BOOK: WORKING WITH MOLECULAR GENETICS (HARDISON)

Ross Hardison
The Pennsylvania State University

LibreTexts™

The Pennsylvania State University

Book: Working with Molecular Genetics (Hardison)

Ross Hardison

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the hundreds of other texts available within this powerful platform, it is freely available for reading, printing and "consuming." Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects.

Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.

The LibreTexts mission is to unite students, faculty and scholars in a cooperative effort to develop an easy-to-use online platform for the construction, customization, and dissemination of OER content to reduce the burdens of unreasonable textbook costs to our students and society. The LibreTexts project is a multi-institutional collaborative venture to develop the next generation of open-access texts to improve postsecondary education at all levels of higher learning by developing an Open Access Resource environment. The project currently consists of 14 independently operating and interconnected libraries that are constantly being optimized by students, faculty, and outside experts to supplant conventional paper-based books. These free textbook alternatives are organized within a central environment that is both vertically (from advance to basic level) and horizontally (across different fields) integrated.

The LibreTexts libraries are Powered by MindTouch® and are supported by the Department of Education Open Textbook Pilot Project, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org. More information on our activities can be found via Facebook (https://facebook.com/Libretexts), Twitter (https://twitter.com/libretexts), or our blog (http://Blog.Libretexts.org).

This text was compiled on 12/01/2022

TABLE OF CONTENTS

Unit II: Replication, Maintenance and Alteration of the Genetic Material

Licensing

*A detailed breakdown of this resource's licensing can be found in **Back Matter/Detailed Licensing**.*

SECTION OVERVIEW

Unit I: Genes, Nucleic Acids, Genomes and Chromosomes

Unit I of this textbook explores the structure and properties of genes and chromosomes, along with a hefty dose of nucleic acid biochemistry.

1: Fundamental Properties of Genes

2: Structures of Nucleic Acids

3: Isolating and Analyzing Genes

4: Genomes and Chromosomes

Thumbnail: DNA Double Helix. (Public Domin; US NIH).

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

1: Fundamental Properties of Genes

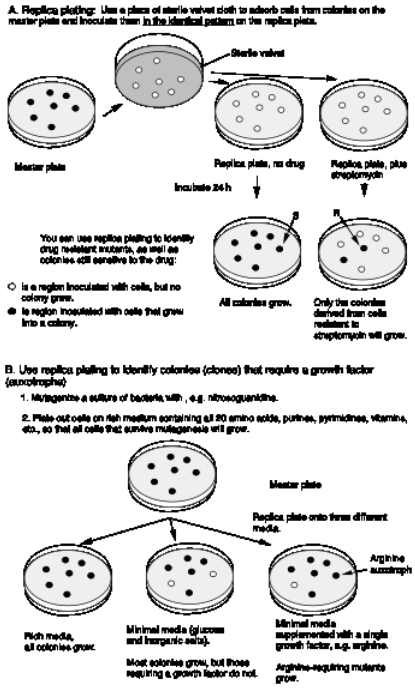


Figure 1.10. F-factor mediated conjugal transfer of DNA in bacteria.

Gene mapping by conjugal transfer

Conjugal transfer can also be used for genetic mapping. By using many different hfr strains, each with the F factor integrated at a different part of the *E.* colichromosome, the positions of many genes were mapped. These studies showed that the genetic map of the *E. coli* chromosome is circular. During conjugal transfer, genes closer to the site of F integration are transferred first. By disrupting the mating at different times, one can determine which genes are closer to the integration site. Thus on the *E. coli*chromosome, genes are mapped in terms of minutes (i.e., the time it takes to transfer to recipient).

For example, for an hfr strain with the F factor integrated at 0 min on the *E. coli*map, conjugal transfer to a female recipient would transfer

- *leuACBD* at 1.7 min
- *pyrH* at 4.6 min
- *proAB* at 5.9 min
- *bioABFCD* at 17.5 min.

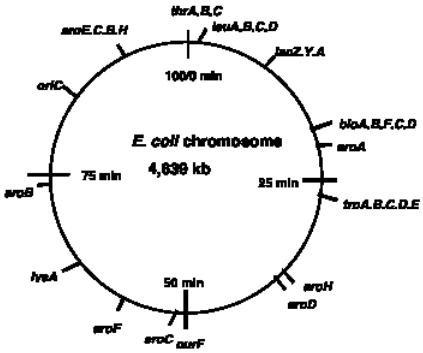


Figure 1.11.Circular genetic map of *E. coli*.

Bacteriophage

Bacteriophageare viruses that infect bacteria. Because of their very large number of progeny and ability to recombine in mixed infections (more than one strain of bacteria in an infection), they have been used extensively in high-resolution definition of genes. Much of what we know about genetic fine structure, prior to the advent of techniques for isolating and sequencing genes, derive form studies in bacteriophage.

Bacteriophage have been a powerful model genetic system, because they have small genomes, have a short life cycle, and produce many progeny from an infected cell. They provide a very efficient means for transfer of DNA into or between cells. The large number of progeny makes it possible to measure very rare recombination events.

Lytic bacteriophage form **plaques**on lawns of bacteria; these are regions of clearing where infected bacteria have lysed. Early work focused on mutants with different **plaque morphology**, e.g. T2 *r*, which shows rapid lysis and generates larger plaques, or on mutants with **different host range**, e.g. T2 *h*, which will kill both host strains B and B/2.

A cis-trans complementation test defines a cistron, which is a gene

Seymour Benzer used the *r*IIlocus of phage T4 to define genes by virtue of their behavior in a complementation test, and also to provide fundamental insight into the structure of genes (in particular, the arrangement of mutable sites - see the next section). The difference in plaque morphology between *r*and *r*+phage is easy to see (large versus small, respectively), and Benzer isolated many *r* mutants of phage T4. The wild type, but not any *rII*mutants, will grow on *E. coli*strain K12(l), whereas both wild type and mutant phage grow equally well on *E. coli*strain B. Thus the wild phenotype is readily detected by its ability to grow in strain K12 (l).

If *E. coli*strain K12 (l) is co-infected with 2 phage carrying mutations at different positions in *rIIA*, you get no multiplication of the phage (except the extremely rare wild type recombinants, which occur at about 1 in 106 progeny). In the diagram below, each line represents the chromosome from one of the parental phage.

rIIA rIIB

phage 1 _|__x_____|_____|_

phage 2 _|_____x_|_____|_

Likewise, if the two phage in the co-infection carry mutations at different positions in *rIIB*, you get no multiplication of the phage (except the extremely rare wild type recombinants, about 1 in 106).

rIIA rIIB

phage 3 _|_____|_x_____|_

phage 4 _|_____|_____x_|_

However, if one of the co-infecting phage carries a mutation in *rIIA*and the other a mutation in *rIIB*, then you see multiplication of the phage, forming a very large number of plaques on *E. coli*strain K12 (l).

rIIA rIIB

phage 1 _|__x_____|_____|_ Provides wt rIIB protein

phage 4 _|_____|_____x_|_ Provides wt rIIA protein

Together these two phage provide all the phage functions - they **complement**each other. This is a positive complementation test. The first two examples show no complementation, and we place them in the same **complementation group**. Mutants that do not complement are placed in the same complementation group; they are different mutant alleles of the same gene. Benzer showed that there were two complementation groups (and therefore two genes) at the *r II* locus, which he called A and B.

In the mixed infection with phage 1 and phage 4, you also obtain the rare wild type recombinants, but there are more recombinants than are seen in the co-infections with different mutant alleles. Why?

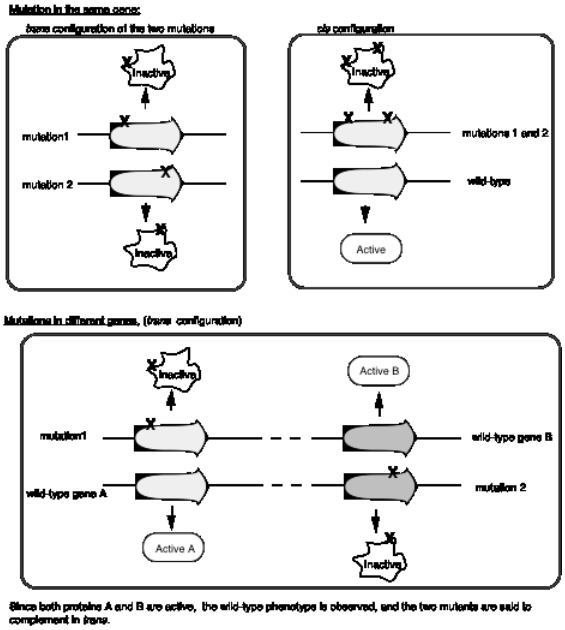


Figure 1.14. Remember that mutations in the *r*gene cause rapid lysis of infected cells, i.e. the length of the lytic cycle is shorter. The difference in plaque morphology between *r*and *r*+phage is easy to see (large versus small, respectively). These two genes are very close together, and many mutations were independently isolated in each. This was summarized in the discussion on complementation above.

Consider the results of infection of a bacterial culture with two mutant alleles of gene *rIIA.*

T4*rIIA6* _|_____x_____|_

and T4*rIIA27* _|_____x_____|_

(x marks the position of the mutation in each allele).

Progeny phage from this infection include those with a parental genotype (in the great majority), and at a much lower frequency, two types of recombinants:

wild type T4 *r*+ _|_____|_

double mutant T4*rIIA6 rIIA27* _|_____x_____x_____|_

The wild type is easily scored because it, and not any *rII*mutants, will grow on E. coli strain K12(l), whereas both wild type and mutant phage grow equally well on E. coli strain B. Thus you can **select**for the wild type (and you will see only the desired recombinant). Finding the double mutants is more laborious, because they are obtained only by screening through the progeny, testing for phage that when backcrossed with the parental phage result in no wild type recombinant progeny.

Equal numbers of wild type and double mutant recombinants were obtained, showing that recombination can occur within a gene, and that this occurs by reciprocal crossing over. If recombination were only between genes, then no wild type phage would result. A large spectrum of recombination values was obtained in crosses for different alleles, just like you obtain for crosses between mutants in separate genes.

Several major conclusions could be made as a result of these experiments on recombination within the *rII*genes.

1. A **large number of mutable sites** occur within a gene, exceeding some 500 for the *rIIA* and *rIIB*genes. We now realize that these correspond to the **individual base pairs** within the gene.
2. The **genetic maps are clearly linear**, indicating that the gene is linear. Now we know a gene is a linear polymer of nucleotides.
3. Most mutations are changes at one mutable site (**point mutations**). Many genes can be restored to wild type by undergoing a reverse mutation at the same site (**reversion**).
4. Other mutations cause the **deletion**of one or more mutable sites, reflecting a physical loss of part of the *rII* gene. Deletions of one or more mutable site (base pair) are extremely unlikely to revert back to the original wild type.

One gene encodes one polypeptide

One of the fundamental insights into how genes function is that **one gene encodes one enzyme** (or more precisely, one **polypeptide**). Beadle and Tatum reached this conclusion based on their complementation analysis of the genes required for arginine biosynthesis in fungi. They showed that a mutation in each gene led to a loss of activity of one enzyme in the multistep pathway of arginine biosynthesis. As discussed above in the section on genetic dissection, a large number of Arg auxotrophs (requiring Arg for growth) were isolated, and then organized into a set of complementation groups, where each complementation groups represents a gene.

The classic work of Beadle and Tatum demonstrated a direct relationship between the genes defined by the auxotrophic mutants and the enzymes required for Arg biosynthesis. They showed that a mutation in one gene resulted in the loss of one particular enzymatic activity, e.g. in the generalized scheme below, a mutation in gene 2 led to a loss of activity of enzyme 2. This led to an accumulation of the substrate for that reaction (intermediate N in the diagram below). If there were 4 complementation groups for the Arg auxotrophs, i.e. 4 genes, then 4 enzymes were found in the pathway for Arg biosynthesis. Each enzyme was affected by mutations in one of the complementation groups.

Intermediates:

M ® N ® O ® P ® Arg

enzyme 1 enzyme 2 enzyme 3 enzyme 4

gene 1 gene 2 gene 3 gene 4

Figure 1.15. A general scheme showing the relationships among metabolic intermediates (M, N, O, P), and end product (Arg), enzymes and the genes that encode them.

In general, each step in a metabolic pathway is catalyzed by an enzyme (identified biochemically) that is the product of a particular gene (identified by mutants unable to synthesize the end product, or unable to break down the starting compound, of a pathway). The number of genes that can generate auxotrophic mutants is (usually) the same as the number of enzymatic steps in the pathway. Auxotrophic mutants in a given gene are missing the corresponding enzyme. Thus Beadle and Tatum concluded that one gene encodes one enzyme. Sometimes more than one gene is required to encode an enzyme because the enzyme has multiple, different polypeptide subunits. **Thus each polypeptide is encoded by a gene**.

The metabolic intermediates that accumulate in each mutant can be used to place the enzymes in their **order of** actionin a pathway. In the diagram in Figure 1.15, mutants in gene 3 accumulated substance O. Feeding substance O to mutants in gene 1 or in gene 2 allows growth in the absence of Arg. We conclude that the defects in enzyme 1 or enzyme 2, respectively, are upstream of enzyme 3. In contrast, feeding substance O to mutants in gene 4 will not allow growth in the absence of Arg. Even though this mutant can convert substance O to substance P, it does not have an active enzyme 4 to convert P to Arg. The inability of mutants in gene 4 to grow on substance O shows that enzyme 4 is downstream of enzyme 3.

Imagine that you are studying serine biosynthesis in a fungus. You isolate serine auxotrophs, do all the pairwise crosses of the mutants and discover that the auxotrophs can be grouped into three complementation groups, called A, B and C. You also discover that a different metabolic intermediate accumulates in members of each complementation group - substance A in auxotrophs in the A complementation group, substance B in the B complementation group and substance C in the C complementation group. Each of the intermediates is fed to auxotrophs from each of the three complementation groups as tabulated below. A + means that the auxotroph was able to grow in media in the absence of serine when fed the indicated substance; a - denotes no growth in the absence of serine.

Fed:	mutant in complementation group A	mutant in complementation group B	mutant in complementation group C
substance A	-	+	+
substance B	-	-	-
substance C	-	+	-

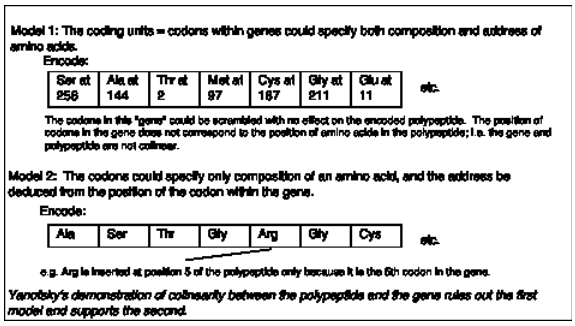


Figure 1.23.Genes and mRNA have untranslated sequences at both the 5’ and 3’ ends.

Eukaryotic mRNAs have covalent attachment of nucleotides at the 5' and 3' ends, and in some cases nucleotides are added internally (a process called *RNA editing*). Recent work shows that additional nucleotides are added post-transcriptionally to some bacterial mRNAs as well.

Regulatory signals can be considered parts of genes

In order to express a gene at the correct time, the DNA also carries signals to start transcription (e.g. promoters), signals for regulating the efficiency of starting transcription (e.g. operators, enhancers or silencers), and signals to stop transcription (e.g. terminators). Minimally, a gene includes the **transcription unit**, which is the segment of DNA that is copied into RNA in the primary transcript. The signals directing RNA polymerase to start at the correct site, and other DNA segments that influence the efficiency of this process are regulatory elements for the gene. One can also consider them to be part of the gene, along with the transcription unit.

A contemporary problem - finding the function of genes

Genes were originally detected by the heritable phenotype generated by their mutant alleles, such as the white eyes in the normally red-eyed *Drosophila*or the sickle cell form of hemoglobin (HbS) in humans. Now that we have the ability to isolate virtually any, and perhaps all, segments of DNA from the genome of an organism, the issue arises as to which of those segments are genes, and what is the function of those genes. (The *genome*is all the DNA in the chromosomes of an organism.) Earlier geneticists knew what the function of the genes were that they were studying (at least in terms of some macroscopic phenotype), even when they had no idea what the nature of the genetic material was. Now molecular biologists are confronted with the opposite problem - we can find and study lots of DNA, but which regions are functions? Many computational approaches are being developed to guide in this analysis, but eventually we come back to that classical definition, i.e. that appropriate mutations in any functional gene should generate a detectable phenotype. The approach of biochemically making mutations in DNA in the laboratory and then testing for the effects in living cells or whole organisms is called "reverse genetics."

Additional Readings

- Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, R. C. and Gelbart, W. M. (1993) An Introduction to Genetic Analysis, Fifth Edition (W. H. Freeman and Company, New York).
- Cairns, J., Stent, G. S. and Watson, J. D., editors (1992) Phage and the Origins of Molecular Biology, Expanded Edition (Cold Spring Harbor Laboratory Press, Plainview, NY).
- Brock, T. D. (1990) The Emergence of Bacterial Genetics (Cold Spring Harbor Laboratory Press, Plainview, NY).
- Benzer, S. (1955) Fine structure of a genetic region in bacteriophage. Proceedings of the National Academy of Sciences, USA 47: 344-354.
- Yanofsky, C. (1963) Amino acid replacements associated with mutation and recombination in the A gene and their relationship to in vitro coding data. Cold Spring Harbor Symposia on Quantitative Biology 18: 133-134.
- Crick, F. (1970) Central dogma of molecular biology. Nature 227:561-563

Questions

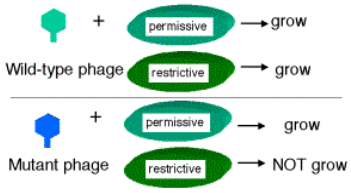
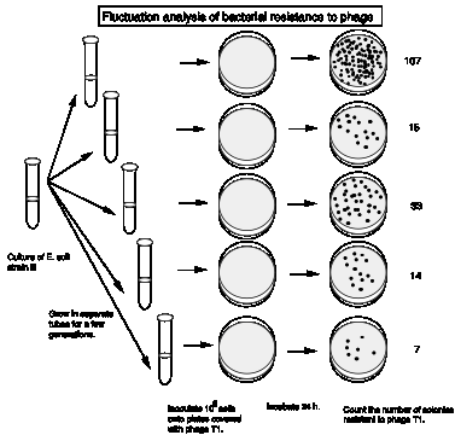


Figure for question 1.10.



The actual results from Luria and Delbrück are summarized in the following table. They examined 87 cultures, each with 0.2 ml of bacteria, for phage resistant colonies.

Number of resistant bacteria	Number of cultures
0	29
1	17
2	4
3	3
4	3
5	2
6-10	5
11-20	6
21-50	7
51-100	5
101-200	2
201-500	4
501-1000	0

Interested students may wish to read about the re-examination of the origin of mutations by Cairns, Overbaugh and Miller (1988, The origin of mutants. Nature 335:142-145). Using a non-lethal selective agent (lactose), they obtained results indicating both pre-adaptive (spontaneous) mutations as well as some apparently induced by the selective agent.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

1.1: Introduction to Genes

Introduction

Species share many traits in common from generation to generation. The bluebird nestlings in the box in my yard will look much like their parents when they are full-grown. The tomato plants that we set out will produce fruits that look, and hopefully taste, like those of their parents. Observable features of organisms, like color, size, and shape, comprise their **phenotype**. Adult male bluebirds share the phenotype of blue wings and a red breast.

A phenotype can be determined by **inherited factors**, by the **environment**, and often by **both**. For example, you are similar to your parents in many aspects of your appearance, your intelligence, and your susceptibility to some diseases, but you are not identical to them in all aspects of these traits. These three traits are clearly the product of both inherited and environmental factors. Considering appearance, I have crooked lower teeth and thinning gray hair, just like my father, but unlike me, neither of my parents has a scar on their knee from a childhood cut. The hair phenotype is inherited, whereas scars are from environmental influences. Quantitative studies show that intellectual capacity is about equally influenced by genetic and environmental factors. Susceptibility to diabetes is partially inherited, but a viral infection may trigger the autoimmune response at its core.

The genetic determinants of the inherited component of a phenotype are called **genes**. The set of genes that make up an organism is its **genotype**. In practice, we will consider only a small subset of the genes in an organism, which comprise a partial genotype. Likewise, an organism's phenotype is all the traits it possesses, but we will only consider partial phenotypes, such as the blue wings of a bluebird or the color of the eyes of a fly.

This chapter will explore some of the basic characteristics of genes, and the experimental evidence for them. Some of the major points include the following.

- Genes are the units of heredity
- They are arranged in a linear fashion along chromosomes.
- Recombination can occur both between and within genes.
- Mutations in different genes required for a phenotype will complement each other in a diploid. This is the basis for genetic dissection of a pathway.
- A gene is composed of a series of mutable sites that are also sites for recombination (now recognized as nucleotides).
- One gene encodes one polypeptide.
- The gene and the polypeptide are colinear.
- Single amino acids are specified by a set of three adjacent mutable sites; this set is called a codon.

In considering experimental evidence for these points, some general genetic techniques as well as genetic techniques for bacteria and phage will be discussed. The first experiments that eventually hypothesized the existence of genes were conducted by Gregor Mendel (see Mendel's Laws).

Genes are mutable

We know that genes are **mutable** because they appear in different forms, called **alleles**. An allele that encodes a normal, functional product (found in nature or a standard laboratory stock) is called the **wild type** allele. Other alleles are altered in a way such that the encoded product differs in function from the wild type. This type of allele is **mutated** or **mutant** (adjective). The alteration in the gene is a **mutation**, and an organism showing the altered phenotype is a **mutant** (noun). Many mutated alleles encode a product that is nonfunctional or less functional than is the wild type, or normal, product; it is easier to break something than to improve it. A **loss-of-function** allele usually shows a **recessive** phenotype, which means that when it is present in the same cell as an allele that produces a different phenotype, the phenotype of the other allele is obtained. If no functional product is made, this loss-of-function allele is a **null** mutation; this can result from no expression or expression of a completely nonfunctional product. Other loss-of-function mutants make less than the normal amount of product, these are called **hypomorphs**. Another class of mutated allele encodes a product that provides an altered or new function. These **gain-of-function** mutations usually show a **dominant** phenotype; e.g. when the gain-of-function allele is in the same cell as a loss-of-function allele, the phenotype of the gain-of-function allele is observed. Another class of gain-of-function mutants makes more than the normal amount of product; these are called **hypermorphs**.

Within a population, the number of alleles at a given locus can vary considerably. Mutant alleles that cause a loss or detrimental change in the function of a gene are selected against, and they are rare in a wild population. In the laboratory, one can utilize

growth conditions that select *for* certain mutants or that maintain mutants, so mutant organisms that would be rare or non-existent in the wild are encountered quite frequently in the laboratory. In many cases, however, alternate forms of genes, i.e. different alleles, have no particular effect on gene function. These variants can be found quite frequently in a population. One common examples of such genetically determined, apparently neutral variation is the ability of some persons to "roll" their tongue. In general, these common alleles are roughly equivalent in function to the wild type allele. Thus they are not providing a strong selective advantage or disadvantage. All the common alleles can be considered the wild type allele. Variant alleles that occur in greater than 5% of population are called **polymorphisms**. The term **variant** includes all alternative forms of a gene, whether they have an effect on function or not. The term **mutant allele** sometimes implies an altered function for the gene.

As will become clearer when we study the fine structure of genes, it is possible to change the structure of the gene (the nucleotide sequence in DNA) without changing the structure of the encoded polypeptide (the amino acid sequence). These **silent substitutions** also generate different alleles, but they can only be detected by examining the structure of the gene; the phenotypes of alleles that differ by silent substitutions are usually identical.

Another possibility is that a mutant allele not only causes a loss of function of the encoded protein, but this altered protein interferes with the activity of other proteins. One way this can happen is by the polypeptide product forming a complex with other polypeptides (e.g. in a heteromultimeric enzyme complex). Sometimes the mutant polypeptide will prevent formation of an active complex with the partner, even in the presence of wild-type polypeptide, thereby leading to a **dominant negative** phenotype. These are of considerable utility now in designing mutant genes and proteins to try to disrupt some cellular function. They are most commonly made in a vector that will drive a high level of expression of the mutant gene, and usually **over-expression** is needed to generate the dominant negative phenotype.

1.2: Genes are the Units of Heredity: Mendel's Laws

Mendel's First Law: Alleles segregate equally

The original experiments by Gregor Mendel involved phenotypic traits (physical, observable characteristics) controlled by single genes. The first one we'll consider is seed color, which can be yellow or green. The dominant allele, denoted *Y*, generates yellow peas in either the homozygous (*YY*) or heterozygous (*Yy*) state, whereas the recessive allele, denoted *y*, generates green peas only in the homozygous state (*yy*). (In plants and flies, the dominant allele is denoted by a capitalized abbreviation and the recessive allele is denoted by a lower case abbreviation.) In a cross between two parents, one homozygous for the dominant allele (*YY*) and the other homozygous for the recessive allele (*yy*), Mendel showed that the F1 progeny were all yellow, i.e. they had had the same phenotype as the parent with the dominant allele. The recessive allele was not contributing to the phenotype.

Had it been lost during the cross? No, when the F1 is crossed with itself, both parental phenotypes were seen in the F2 progeny. The effect of the recessive allele *reappeared* in the second cross, showing that it was still present in the F1 hybrids, but was having no effect. In the F2 progeny, the dominant phenotype (yellow) was observed in 75% of the progeny and the recessive (green) appeared in only 25% of the progeny.

Note that *discrete phenotypes* were obtained (yellow or green), *not a continuum of phenotypes*. The genes are behaving as **units**, not as some continuous function.

The results can be explained by hypothesizing that each parent has two copies of the gene (i.e., two alleles) that **segregate equally**, one per gamete. Since they are homozygous, each parent can form only type of gamete (*Y* or *y*,respectively). When the gametes join in the zygotes of the F1 generation, each individual receives one dominant allele and one recessive allele (*Yy*), and thus all of the F1 generation shows the dominant phenotype (e.g. yellow peas). This is the **uniform phenotype** observed for the F1 generation.

The two alleles did not alter one another when present together in the F1 generation, because when F1 is crossed with F1, the two parental phenotypes are obtained in the F2 generation.

The ratio of 3:1 dominant: recessive observed in the F2 is expected for the equal segregation of the alleles from the F1(*Y* and *y*) and their random rejoining in the zygotes of the F2, producing the genotypes 1 *YY*, 2 *Yy*, and 1 *yy*. Again the genes are behaving as discrete units. These precise mathematical ratios (3:1 for phenotypes in this cross, or 1:2:1 for the genotype) provide the evidence that genes, units of heredity, are determining the phenotypes observed.

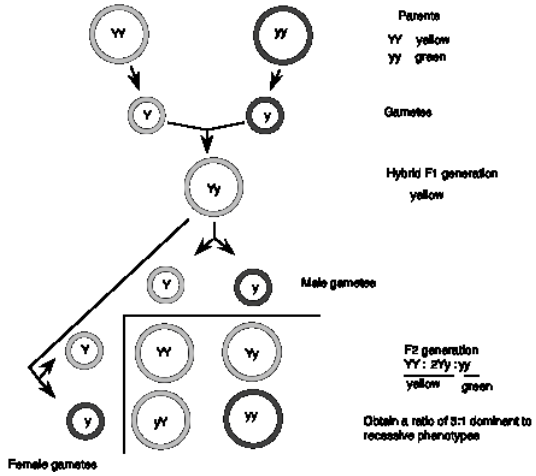


Figure 1.1. Mendel's First Law: Equal segregation of alleles.

Not all loci show the property of **complete dominance** illustrated by the *Y* locus in peas. Sometimes **partial dominance** is observed, in which an intermediate phenotype seen in a heterozygote. An example is the pink color of snapdragons obtained when white and red are crossed. However, the parental phenotypes reappear in the F2 generation, showing that the alleles were not altered in the heterozygote. In this case, **gene dosage** is important in determining the phenotype; two wild-type alleles produce a red flower, but only one wild-type allele produces a pink flower. Sometimes **co-dominance** is observed, in which both alleles contribute equally to the phenotype. An example is the *ABO* blood group locus. Heterozygotes have both the *A* and *B* form of the glycoprotein that is encoded by the different alleles of the gene.

Mendel's Second Law: Different Genes Assort Independently

Now that we have some understanding of the behavior of the different alleles of a single gene, let's consider how two different genes behave during a cross. Do they tend to stay together, or do they assort independently?

Mendel examined two different traits, seed color (as described in the previous section) and seed shape. Two alleles at the locus controlling seed shape were studied, the dominant round (*R*) and recessive wrinkled (*r*) alleles. Mendel crossed one parent that was homozygous for the dominant alleles of these two different genes (round yellow *RRYY*) with another parent that was homozygous for the recessive alleles of those two genes (wrinkled green *rryy*) (see Figure 1.2).

Re-stating the basic question, do the alleles at each locus always stay together (i.e. round with yellow, wrinkled with green) or do they appear in new combinations in the progeny? As expected from the 1st law, the F1 generation shows a uniform round yellow phenotype, since one dominant and one recessive allele was inherited from the parents. When the F2 progeny are obtained by crossing the F1 generation, the parental phenotypes reappear (as expected from the first law), but two **nonparental phenotypes** also appear that differ from the parents: wrinkled yellow and round green!

The results can be explained by the **alleles of each different gene assorting into gametes independently**. For example, in the gametes from the F1 generation, *R* can assort with *Y*or *y*, and *r* can assort with *Y* or *y*, so that four types of gametes form: *RY*, *Ry*, *rY*, and *ry*. These can rejoin randomly with other gametes from the F1 generation, producing the results in the grid shown in Figure 1.2. The alternative, that *R* always assorted with *Y*, etc. was not observed.

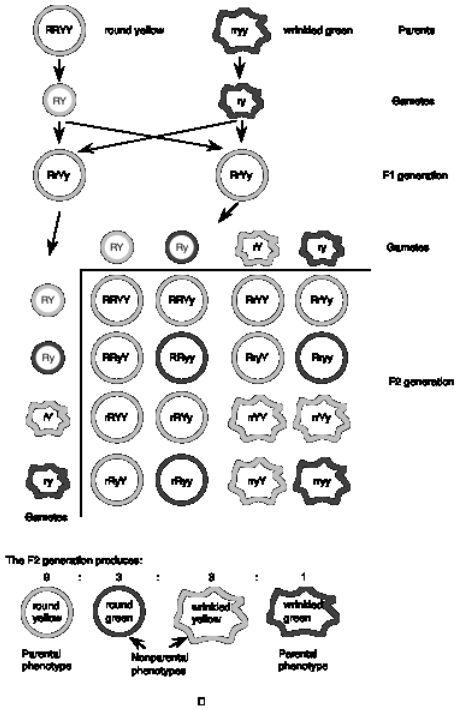


Figure 1.2. Mendel's Second Law: Independent assortment of different genes.

Again, the genes are behaving as units, and the gene for one trait (e.g. color) does not affect a gene for another trait (e.g. shape). Further breeding shows that many nonparental genotypes are present, some of which give a parental phenotype (e.g. *RrYy*). These results are obtained for genes that are **not** linked on chromosomes. Linkage can lead to deviations from these expected ratios in a mating, and this can be used to map the locations of genes on chromosomes, as discussed in the next section.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

1.3: Properties of Genes

Genes are on Chromosomes

In 1902, Sutton and Boveri independently realized that the behavior of genes in Mendelian crosses mimics the movement of chromosomes during meiosis and fertilization. They surmised that the two alleles of each gene correlated with the **homologous pair of chromosomes**. The **equal segregation of alleles** could be explained by the **separation of homologous chromosomes** at anaphase I of meiosis. As diagrammed in Figure 1.3, the chromosome with the *R* allele would go to a different cell than its homolog with the *r* allele at the end of meiosis I, and likewise for the *Y*and *y* alleles. The rejoining of alleles corresponded to the joining of chromosomes, one from each parent, at fertilization. The **independent assortment of different genes** mimics the **independent separation of homologs of different chromosomes** in meiosis. For instance, the paternal copy of chromosome 1 may assort with the maternal copy of chromosome 21 in formation of a gamete. Figure 1.3 shows the dark blue chromosome with the *R* allele assorting with the light red chromosome with the *y*allele, but it is equally likely that it will assort with the dark red chromosome with the *Y*allele. As shown in Figure 1.4, the completion of meiosis results in 4 germ cells for each cell that entered meiosis. All the combinations of alleles of different genes diagrammed in Figure 1.2 can be formed in this process.

This correlation of the behavior of alleles in matings and the movement of chromosomes during meiosis and fertilization produced the **chromosomal theory of inheritance**. One could think of the alleles discerned in genetic crosses as being located at the same locus on the different homologs of a chromosome.

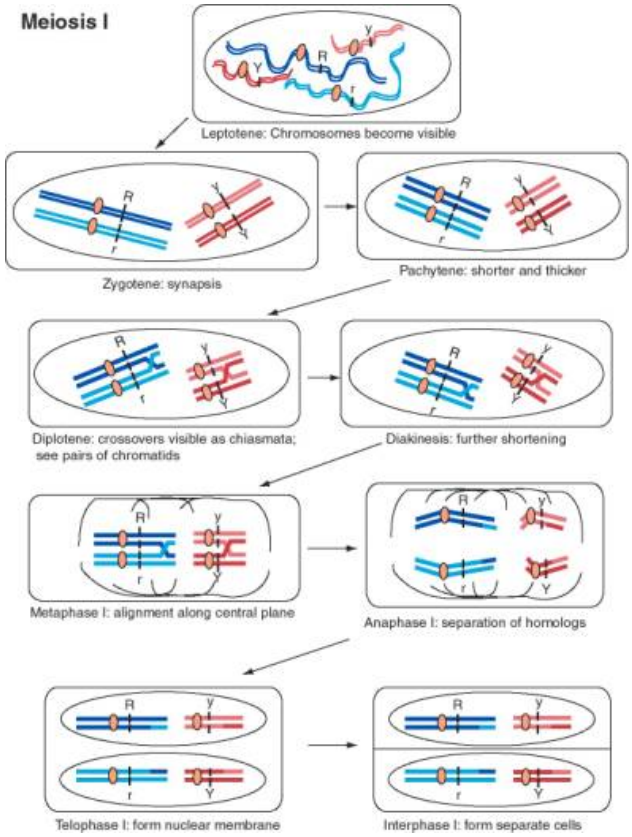


Figure 1.3. Movement of chromosomes during meiosis I, the first divisional process of meiosis. The chromosomes are drawn starting after the synthesis of a copy of each homologous chromosome, so there are two copies of each homolog of a chromosome pair. The two DNA duplexes for each homolog are joined at a single centromere. Meiosis is the process of segregating these four copies of each chromosome (4 alleles for each gene) into four germ cells with one copy of each chromosome. In this diagram, two different chromosome pairs are displayed with each homolog colored a different shade (dark or light red for the shorter chromosome, dark or light blue for the longer chromosome). Each line is a duplex DNA molecule. The R locus is on the longer blue chromosome, with distinctive alleles for each homolog, and the Ylocus is on the shorter red chromosome, again with distinctive alleles for each homolog. Meiosis begins with the leptotene, when the chromosomes become visible as long filaments. The two homologous chromosomes undergo synapsis during zygotene, in which they align along their lengths. The chromosomes become shorter and thicker during pachytene, and crossovers between chromatids of the two different homologs form. The chromosomes start to pull apart in diplotene, at which point the crossovers in chiasmata are visible. The chromosomes shorten further during diakinesis. During metaphase, the chromosomes align along the equatorial plane of the cell, i.e. the plane in which cell division will occur. The nuclear membrane is disassembled at this point. The members of a homologous pair move to opposite poles of the cell during anaphase. This is the cytological even that accounts for the equal segregation of alleles. Note that the centromeres do not separate during anaphase I, and the two sister chromatids stay together. The crossovers are also resolved at this stage. In some organisms, the nuclear membrane reforms during a telophase of meiosis I, followed by cell division and an interphase I.

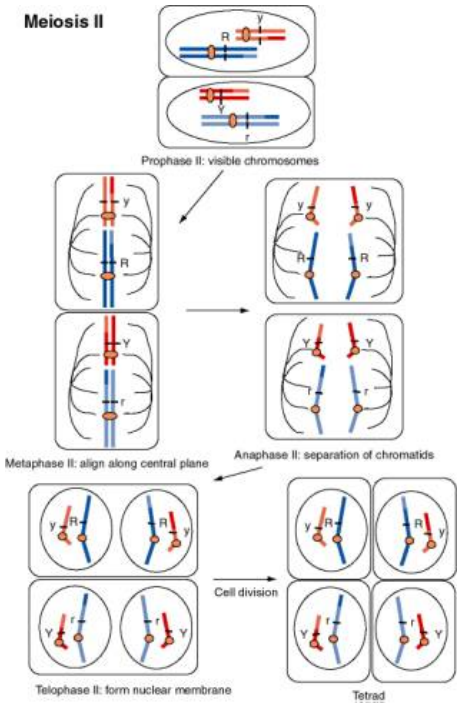


Figure 1.4. Movement of chromosomes during meiosis II, the second divisional process of meiosis. The chromosomes, each with two sister chromatids linked at the centromere, contract and become visible during prophase II. The nuclear membrane disassembles and chromosomes align along the equatorial plane during metaphase II. The centromeres divide and the chromosomes separate during anaphase II. The nuclear membrane reforms during telophase II, and after cell division, a tetrad with one of each chromosome is produced. If the dark blue chromosome had assorted with the dark red chromosome during anaphase I, the resulting spores would be R Y and r y.

Linked genes lie along chromosomes in a linear array

The proponents of the chromosome theory of heredity realized that that the number of genes would probably greatly exceed the number of chromosomes. However, many early genetic studies showed independent assortment between genes with no evidence of linkage. This led to a proposal that a chromosome broke down during meiosis into smaller parts consisting only of individual genes, but such disassembly of chromosomes during meiosis was never observed. Evidence for linkage did eventually come from a demonstration of the absence of independent assortment between different genes. In complementary work, McClintock and Creighton demonstrated an association between different genes and a particular chromosome in 1931.

The behavior of two genes carried on the same chromosome may deviate from the predictions of Mendel's 2nd law. The proportion of parental genotypes in the F2 may be greater than expected because of a reduction in nonparental genotypes. This propensity of some characters to remain associated instead of assorting independently is called **linkage**. When deduced from studies of a population, it is called **linkage disequilibrium**. Figure 1.5. illustrates a cross that shows linkage.

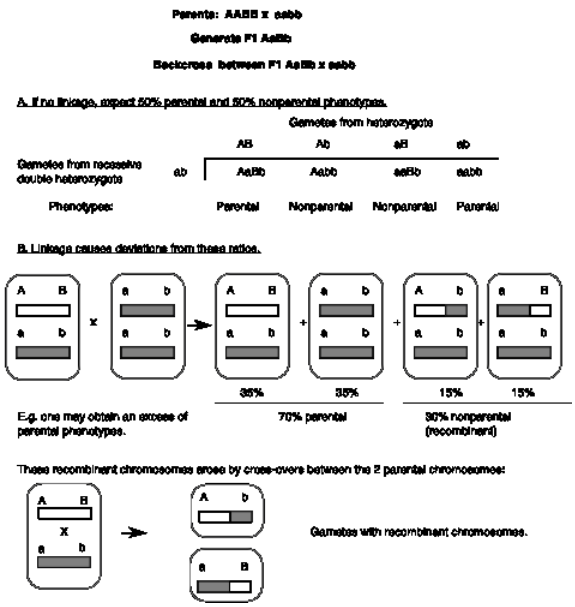


Figure 1.5.Linkage and recombination between genes on the same chromosome.

An F1 heterozygote (*AaBb*) is made by crossing a homozygous dominant parent (*AABB*) with a homozygous recessive parent (*aabb*). A **backcross** is then made between the F1 heterozygote (*AaBb*) and a recessive homozygote (*aabb*), so that the alleles of the recessive parent make no contribution to the phenotype of the progeny. (This is a fairly common cross in genetics, since the genotype of an individual can be ascertained by crossing with such an individual, homozygous recessive at both loci.)

1. As shown in part A of Figure 1.5, if there is *no* linkage, one expects 50% parental phenotypes (from genotypes *AaBb* and *aabb*) and 50% nonparental phenotypes (from genotypes *Aabb*and *aaBb*). This fits with the expectations of Mendel's law of independent assortment of different genes for this backcross. (Sometimes the nonparental phenotypes are called "recombinant" but that confuses this reassortment with events that involve crossovers in the DNA.)
2. If the two genes *are* linked and there is *no* recombination between them, then all progeny will have a parental phenotype. In particular, if genes *A* and *B* are linked, then the backcross *AB/ab* x *ab/ab* yields *AB/ab* progeny 50% of the time and *ab/ab* progeny 50% of the time, *in the absence of recombination*. [In this notation, the alleles to the left of the slash (/) are linked on one chromosome and the alleles to the right of the slash are linked on the homologous chromosome.] Thus only the parental phenotypes are found in the progeny of this cross (i.e. the progeny will show either the dominant characters at each locus or the recessive characters at each locus). Another way of looking at this is that, in the absence of recombination between the homologous chromosomes, all the progeny of this cross will be one of the first two types shown in panel B of Figure 1.5.
3. Note that the dominant alleles can be in the opposite phase, with the dominant *A*allele linked to the recessive *b* allele. For instance, the F1 heterozygote could be formed by a cross between the parents *Ab/Ab*and *aB/aB*to generate *Ab/aB*. In this case, the backcross *Ab/aB*x *ab/ab* will still generate only progeny with parental *pheno*types but a new, nonparental *geno*type (i.e. *Ab/ab*and *aB/ab*; these look like the parents *Ab/Ab* and *aB/aB*). The phase with both dominant alleles on the same chromosome is called the "coupling conformation", whereas the opposite phase is called the "repulsing conformation."
4. But in most cases, recombination can occur between linked genes. In part B of Figure 1.5, there is an *increase* in parental types (from the 50% expected for unlinked genes to the observed 70%) and a *decrease*in nonparental types (30%), showing that allele *A*tends to stay with allele *B*, in contrast to the prediction of the 2nd law. Thus these genes are not assorting independently, and one concludes there is *linkage* between genes *A*and *B*.

The frequency of parental types is not as high as expected for linkage without recombination (which would have been 100%, as discussed above). Indeed, the nonparental types in this experiment result from a physical crossover (breaking and rejoining) between the two homologous chromosomes during meiosis in the *AB/ab* parent. This is a recombination event in the DNA.

(5) We conclude that genes *A* and *B*are linked, and have a recombination frequency of 30%.

map distance = x 100

1 map unit = 1 centiMorgan = 1% recombination

1 centiMorgan = 1 cM = about 1 Mb for human chromosomes

Exercise 1.3.1

In their genetic studies of the fruitfly *Drosophila melanogaster*, Thomas Hunt Morgan and his co-workers found many examples of genes that associated together in groups. One example is the gene for purple eye color (the mutant allele is abbreviated *pr*) that is recessive to the allele for normal red eyes (*pr*+) and the gene for vestigial, or shortened, wings (the mutant allele is abbreviated *vg*) that is recessive to the normal allele for long wings (*vg*+). When a homozygous *purple vestigial*fly is crossed to a homozygous red-eyed long-winged fly, the heterozygous F1 generation shows a normal phenotype. When male heterozygotes are backcrossed to females that are homozygous *purple vestigial* (i.e. homozygous recessive at both loci), only two phenotypes appear in the progeny: the homozygous recessive *purple vestigial*flies and the normal flies.

a. What are the predictions of the backcross if the two genes are not linked?
b. What do the results of the backcross tell you?
c. If the heterozygotes F1 in the backcross are female, then *purple* long-winged and red-eyed *vestigial*flies appear in the progeny. The combined frequency of these recombinant types is 15.2 %. What does this tell you about the arrangement of the genes?

Answer

TBA

Question 1.5 provides some practice in calculating recombination frequencies.

Individual map distances are (roughly) additive

A--10--*B*-5-*C*

---------15----

The recombination distances are not strictly additive if multiple crossovers can occur (see questions 1.6 and 1.7.) Recombination between linked genes occurs by the process of **crossing over** between chromosomes, at **chiasma during meiosis**. The mechanism of recombination is considered in Chapter 8.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

1.4: Complementation and Recombination

What is Complementation?

A general definition of **complementation** is the ability of two mutants in combination to restore a normal phenotype. Dominance observed in heterozygotes reflects the ability of wild-type alleles to complement loss-of-function alleles. You know that a dominant allele will determine the phenotype of a heterozygote composed of a dominant and a recessive allele. Often, recessive alleles are loss-of-function mutations, whereas the dominant allele is the wild type, encoding a functional enzyme. Using the example that led to Mendel's First Law, a cross between *YY* (yellow) peas and *yy* (green) peas yielded yellow peas in the F1 heterozygote (*Yy*). In this case the chromosome carrying the *Y* allele encodes the enzymatic function missing in the product of the recessive *y* allele, and the pathway for pigment biosynthesis continues on to make a yellow product. Thus you could say that the dominant *Y* allele ***complements*** the recessive *y* allele - it provides the missing function.

We can continue the analogy to the classic cross for Mendel's Second Law. Let's look at the same genes, but a different arrangement of alleles. Consider a cross between round green (*RRyy*) and wrinkled yellow (*rrYY*) peas; in this case each parent is providing a dominant allele of one gene and a recessive allele of the other. The F1 heterozygote is round yellow (*RrYy*), i.e., the phenotypes of the dominant alleles are seen. But you could also describe this situation as the chromosomes from *rrYY*peas complementing the deficiency in the *RRyy* chromosomes, and *vice versa*. In particular, the *Y*allele from the *rrYY*parent provides the function missing in the *y* allele from the *RRyy* parent, and the *R* allele from the *RRyy* parent provides the function missing in the *r*allele from the *rrYY* parent. If the phenotype you are looking for is a round yellow pea, you could conclude that mutants in the *R*-gene complement mutants in the *Y*-gene. Since in a heterozygote, the functional allele will provide the activity missing in the mutant allele (if the mutation is a loss-of-function), one could say that dominant alleles complement recessive alleles. Thus dominant alleles determine the phenotype in a heterozygote with both dominant and recessive alleles.

Complementation distinguishes between mutations in the same gene or in different genes

The ability of complementation analysis to determine whether mutations are in the same or different genes is the basis for genetic dissection. In this process, one **finds the genes whose products are required in a pathway**. In the examples from peas, the metabolic pathway to yellow pigments is distinctly different from the pathway to round peas, which is the starch biosynthesis pathway. Complementation analysis is useful in dissecting the steps in a pathway, starting with many mutants that generate the same phenotype. This is a more conventional example of complementation.

Many fungi can propagate as haploids but can also mate to form diploids prior to sporulation. Thus one can screen for mutants in haploids and obtain recessive mutants, and then test their behavior in combination with other mutants in the diploid state. Let's say that a haploid strain of a fungus was mutagenized and screened for arginine **auxotrophs**, i.e. mutants that require arginine to grow. Six of the mutants were mated to form all the possible diploid combinations, and tested for the ability of the diploids to grow in the absence of arginine (**prototrophy**). The results are tabulated below, with a + designating growth in the absence of arginine, and a - designating no growth.

Table 1.1. Growth of the diploids in the absence of arginine

Mutant number	1	2	3	4	5	6
1	-	+	+	-	+	+
2		-	-	+	+	+
3			-	+	+	+
4				-	+	+
5					-	+
6						-

As you would expect, when mutant 1 is mated with itself, the resulting diploid is still an auxotroph; this is the same as being homozygous for the defective allele of a gene. But when mutant 1 is mated with mutant 2 (so their chromosomes are combined), the resulting diploid has prototrophy restored, i.e. it can make its own arginine. This is true for **all**the progeny. *We conclude that mutant 1 will* ***complement*** *mutant 2*. If we say that mutant 1 has a mutation in gene 1 of the pathway for arginine biosynthesis, and mutant 2 has a mutation in gene 2 of this pathway, then the diagram in Figure 1.6 describes the situation in the haploids and the

diploid. (Note that if the organism has more than one chromosome, then genes 1 and 2 need not be on the same chromosome.) Since the enzymes encoded by genes 1 and 2 are needed for arginine biosynthesis, neither mutant in the haploid state can make arginine. But when these chromosomes are combined in the diploid state, the chromosome from mutant 1 will provide a normal product of gene 2, and the chromosome from mutant 2 will provide a normal product of gene 1. Since each provides what is missing in the other, they complement. Mutant 1 will also complement mutant 3, and one concludes that these strains are carrying mutations in different genes required for arginine biosynthesis.

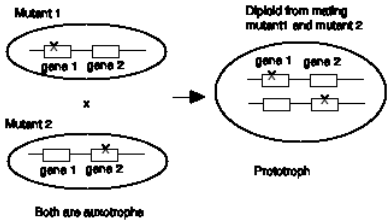


Figure 1.6. Complementation between two haploid mutants when combined in a diploid.

In contrast, the diploid resulting from mating mutant 1 with mutant 4 is still an auxotroph; it will not grow in the absence of arginine. Assuming that both these mutants are recessive (i.e. contain loss-of-function alleles), then we conclude that the mutations are in the same gene (gene 1 in the above diagram). We place these mutants in the same ***complementation group.*** Likewise, mutants 2 and 3 fail to complement, and they are in the same complementation group. Thus mutant 2 and mutant 3 are carrying different mutant alleles of the same gene (gene 2).

Mutant 5 will complement all the other mutants, so it is in a different gene, and the same is true for mutant 6. Thus this mutation and complementation analysis shows that this fungus has at least 4 genes involved in arginine biosynthesis: gene 1 (defined by mutants alleles in strains 1 and 4), gene 2 (defined by mutants alleles in strains 2 and 3), and two other genes, one mutated in strain 5 and the other mutated in strain 6.

Genetic dissection by complementation is very powerful. An investigator can start with a large number of mutants, all of which have the same phenotype, and then group them into sets of mutant alleles of different genes. Groups of mutations that do not complement each other constitute a complementation group, which is equivalent to a gene. Each mutation in a given complementation group is a mutant allele of the gene. The product of each gene, whether a polypeptide or RNA, is needed for the cellular function that, when altered, generates the phenotype that was the basis for the initial screen. The number of different complementation groups, or genes, gives an approximation of the number of polypeptides or RNA molecules utilized in generating the cellular function.

> Question 1.2. Consider the following complementation analysis. Five mutations in a biosynthetic pathway (producing auxotrophs in a haploid state) were placed pairwise in a cell in *trans*(diploid analysis). The diploid cells were then assayed for reconstitution of the biosynthetic pathway; complementing mutations were able to grow in the absence of the end product of the pathway (i.e. they now had a functional biosynthetic pathway). A + indicates a complementing pair of mutations; a - means that the two mutations did not complement.

Mutation number

1 2 3 4 5

1 - + - + -

2 - + + +

3 - + -

4 - +

5 -

> a) Which mutations are in the same complementation group (representing mutant alleles of the same gene)?
>
> b) What is the minimal number of enzymatic steps in the biosynthetic pathway?

Recombination

Note that **all** the diploid progeny fungi from the mating of mutant strains 1 and 2 have the ability to grow on arginine, and this complementation does not require any change in the two chromosomes (Figure 1.6.). The only thing that is happening is that the functional alleles of each gene are providing active enzymes. If genes 1 and 2 are on the same chromosome, at a **low frequency**, recombinations between the two chromosomes in the diploid can lead to crossovers, resulting in one chromosome with wild-type alleles of each gene and another chromosome with the mutant alleles of each gene (Figure 1.7). This can be observed in fungi by inducing sporulation of the diploid. Each spore is haploid, and the vast majority will carry one of the two parental chromosomes, and hence be defective in either gene 1 or gene 2. But **wild type recombinants** can be observed at a low frequency; these will be prototrophs. The double-mutant recombinants will be auxotrophs, of course, but these can be distinguished from the parental single mutants by the inability of the double mutants to complement either mutant strain 1 or strain 2.

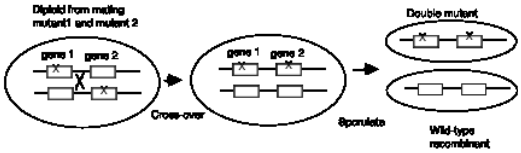


Figure 1.7. Recombination between homologous chromosomes in a diploid

Note that this ***recombination*** is a physical alteration in the chromosomes. The frequency of its occurrence is directly proportional to the distance the genes are apart, which is the basis for mapping genes by their recombination distances. Recombination occurs in a small fraction of the progeny, whereas all the progeny of a complementing diploid have the previously lost function restored.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

1.E: Fundamental Properties of Genes (Exercises)

Question 1.5. Calculating recombination frequencies

Corn kernels can be colored or white, determined by the alleles *C*(colored, which is dominant) or *c*(white, which is recessive) of the *colored*gene. Likewise, alleles of the *shrunken*gene determine whether the kernels are nonshrunken (*Sh*, dominant) or shrunken (*sh*, recessive). The geneticist Hutchison crossed a homozygous colored shrunken strain (*CC shsh*) to a homozygous white nonshrunken strain (*cc ShSh*) and obtained the heterozygous colored nonshrunken F1. The F1 was backcrossed to a homozygous recessive white shrunken strain (*cc shsh*). Four phenotypes were observed in the F2 progeny, in the numbers shown below.

Phenotype	Number of plants
colored shrunken	21,379
white nonshrunken	21,096
colored nonshrunken	638
white shrunken	672

a) What are the predicted frequencies of these phenotypes if the *colored*and *shrunken*genes are not linked?

b) Are these genes linked, and if so, what is the recombination frequency between them?

Question 1.6. Constructing a linkage map:

Consider three genes, A, B and C, that are located on the same chromosome. The arrangement of the three genes can be determined by a series of three crosses, each following two of the genes (referred to as two-factor crosses). In each cross, a parental strain that is homozygous for the dominant alleles of the two genes (e.g. *AB/AB*) is crossed with a strain that is homozygous for the recessive alleles of the two genes (e.g. *ab/ab*), to yield an F1 that is heterozygous for both of the genes (e.g. *AB/ab*). In this notation, the slash (*/*) separates the alleles of genes on one chromosome from those on the homologous chromosome. The F1 (*AB/ab*) contains one chromosome from each parent. It is then backcrossed to a strain that is homozygous for the recessive alleles (*ab/ab*) so that the fates of the parental chromosomes can be easily followed. Let's say the resulting progeny in the F2 (second) generation showed the parental phenotypes (*AB* and *ab*) 70% of the time. That is, 70% of the progeny showed only the dominant characters (*AB*) or only the recessive characters (*ab*), which reflect the haploid genotypes *AB/ab*and *ab/ab*, respectively, in the F2 progeny. The remaining 30% of the progeny showed recombinant phenotypes (*Ab*and *aB*) reflecting the genotypes *Ab/ab*and *aB/ab*in the F2 progeny. Similar crosses using F1's from parental *AC/AC*and *ac/ac*backcrossed to a homozygous recessive strain (*ac/ac*) generated recombinant phenotypes *Ac*and *aC*in 10% of the progeny. And finally, crosses using F1's from parental *BC/BC*and *bc/bc*backcrossed to a homozygous recessive strain (*bc/bc*) generated recombinant phenotypes *Bc*and *bC*in 25% of the progeny.

a. What accounts for the appearance of the recombinant phenotypes in the F2 progeny?

b. Which genes are closer to each other and which ones are further away?

c. What is a linkage map that is consistent with the data given?

Question 1.7

Why are the distances in the previous problem not exactly additive, e.g. why is the distance between the outside markers (*A* and *B*) not 35 map units (or 35% recombination)? There are several possible explanations, and this problem explores the effects of multiple crossovers. The basic idea is that the further apart two genes are, the more likely that recombination can occur multiple times between them. Of course, two (or any even number of) crossover events between two genes will restore the parental arrangement, whereas three (or any odd number of) crossover events will give a recombinant arrangement, thereby effectively decreasing the observed number of recombinants in the progeny of a cross.

For the case examined in the previous problem, with genes in the order *A*___*C*_____*B*, let the term *ab*refer to the frequency of recombination between genes *A*and *B*, and likewise let *ac*refer to the frequency of recombination between genes *A*and *C*, and *cb*refer to the frequency of recombination between genes *C*and *B*.

a) What is the probability that when recombination occurs in the interval between *A*and *C*, an independent recombination event also occurs in the interval between *C*and *B*?

b) What is the probability that when recombination occurs in the interval between *C*and *B*, an independent recombination event also occurs in the interval between *A*and *C*?

c) The two probabilities, or frequencies, in a and b above will effectively lower the actual recombination between the outside markers *A*and *B*to that observed in the experiment. What is an equation that expresses this relationship, and does it fit the data in problem 3?

d. What is the better estimate for the distance between genes *A*and *B*in the previous problem?

Question 1.8 Complementation and recombination in microbes.

The State College Bar Association has commissioned you to study an organism, Alcophila latrobus, which thrives on Rolling Rock beer and is ruining the local shipments. You find three mutants that have lost the ability to grow on Rolling Rock (RR).

a) Recombination between the mutants can restore the ability to grow on RR. From the following recombination frequencies, construct a linkage map for mutations 1, 2, and 3.

Recombination between Frequency

1- and 2- 0.100

1- and 3- 0.099

2- and 3- 0.001

b) The following diploid constructions were tested for their ability to grow on RR. What do these data tell you about mutations 1, 2, and 3?

Grow on RR?

1) 1- 2+ / 1+ 2- yes

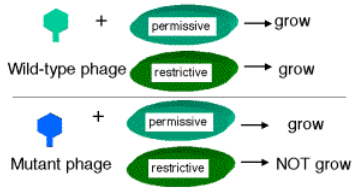
2) 1- 3+ / 1+ 3- yes

3) 2- 3+ / 2+ 3- no

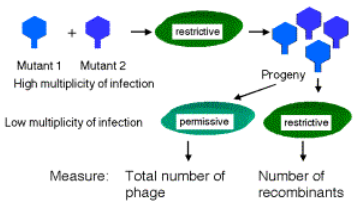
Question 1.9 Using recombination frequencies and complementation to deduce maps and pathways in phage.

A set of four mutant phage that were unable to grow in a particular bacterial host (lets call it restrictive) were isolated; however, both mutant and wild type phage will grow in another, permissive host. To get information about the genes required for growth on the restrictive host, this host was co-infected with pairs of mutant phage, and the number of phage obtained after infection was measured. The top number for each co-infection gives the total number of phage released (grown on the permissive host) and the bottom number gives the number of wild-type recombinant phage (grown on the restrictive host). The wild-type parental phage gives 1010 phage after infecting either host. The limit of detection is 102 phage.

Phenotypes of phage, problem 1.9:



Assays after co-infection with mutant phage:



Results of assays, problem 1.9:

Number of phage

mutant 1 mutant 2 mutant 3 mutant 4

mutant 1 total <102

recombinants <102

mutant 2 total 1010 <102

recombinants 5x106 <102

mutant 3 total 1010 1010 <102

recombinants 107 5x106 <102

mutant 4 total 105 1010 1010 <102

recombinants 105 5x106 107 <102

a) Which mutants are in the same complementation group? What is the minimum number of genes in the pathway for growth on the restrictive host?

b) Which mutations have the shortest distance between them?

c) Which mutations have the greatest distance between them?

d) Draw a map of the genes in the pathway required for growth on the restrictive host. Show the positions of the genes, the positions of the mutations and the relative distances between them.

Question 1.10

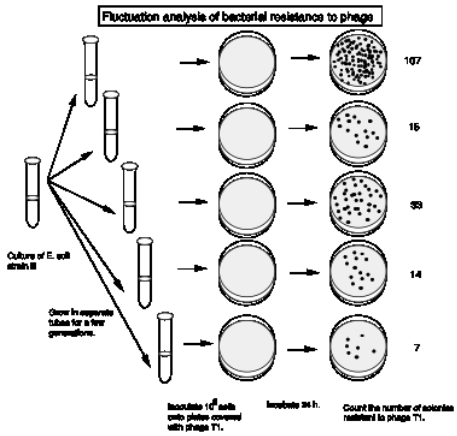
One of the classic experiments in bacterial genetics is the **fluctuation analysis**of Luria and Delbrück (1943, Mutations of bacteria from virus sensitivity to virus resistance, Genetics 28: 491-511). These authors wanted to determine whether **mutations arose spontaneously**while bacteria grew in culture, or if the **mutations were induced**by the conditions used to select for them. They knew that bacteria resistant to phage infection could be isolated from infected cultures. When a bacterial culture is infected with a lytic phage, initially it “clears” because virtually all the cells are lysed, but after several hours phage-resistant bacteria will start to grow.

Luria and Delbrück realized that the two hypothesis for the source of the mutations could be distinguished by a quantitative analysis of the number of the phage-resistant bacteria found in many infected cultures. The experimental approach is outlined in the figure below. Many cultures of bacteria are grown, then infected with a dose of phage T1 that is sufficient to kill all the cells, except those that have acquired resistance. These resistant bacteria grow into colonies on plates and can be counted.

a. What are the predictions for the distribution of the number of resistant bacteria in the two models? Assume that on average, about 1 in 107 bacteria are resistant to infection by phage T1.

b. What do results like those in the figure and table tell you about which model is correct?

Figure for question 1.10.



The actual results from Luria and Delbrück are summarized in the following table. They examined 87 cultures, each with 0.2 ml of bacteria, for phage resistant colonies.

Number of resistant bacteria	Number of cultures
0	29
1	17
2	4
3	3
4	3
5	2
6-10	5
11-20	6
21-50	7
51-100	5
101-200	2
201-500	4
501-1000	0

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

1.S: Additional Readings (Summary)

Additional Readings

- Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, R. C. and Gelbart, W. M. (1993) An Introduction to Genetic Analysis, Fifth Edition (W. H. Freeman and Company, New York).
- Cairns, J., Stent, G. S. and Watson, J. D., editors (1992) Phage and the Origins of Molecular Biology, Expanded Edition (Cold Spring Harbor Laboratory Press, Plainview, NY).
- Brock, T. D. (1990) The Emergence of Bacterial Genetics (Cold Spring Harbor Laboratory Press, Plainview, NY).
- Benzer, S. (1955) Fine structure of a genetic region in bacteriophage. Proceedings of the National Academy of Sciences, USA 47: 344-354.
- Yanofsky, C. (1963) Amino acid replacements associated with mutation and recombination in the A gene and their relationship to in vitro coding data. Cold Spring Harbor Symposia on Quantitative Biology 18: 133-134.
- Crick, F. (1970) Central dogma of molecular biology. Nature 227:561-563

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Central Dogma: DNA to RNA to protein

A few years after he and James Watson had proposed the double helical structure for DNA, Francis Crick (with other collaborators) proposed that a less stable nucleic acid, RNA, served as a messenger RNA that provided a transient copy of the genetic material that could be translated into the protein product encoded by the gene. Such mRNAs were indeed found. These and other studies led Francis Crick to formulate this “central dogma” of molecular biology (Figure 1.21).

This model states that **DNA serves as the repository of genetic information**. It can be **replicated** accurately and indefinitely. The **genetic information is expressed** by the DNA first serving as a template for the **synthesis of (messenger) RNA**; this occurs in a process called **transcription**. The mRNA then serves as a template, which is read by ribosomes and **translated**into **protein**. The protein products can be enzymes that catalyze the many metabolic transformations in the cell, or they can be structural proteins.

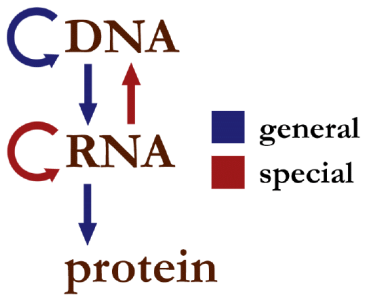


Figure 1.21.The central dogma of molecular biology. (Public Domain; Narayanese).

Although there have been some additional steps added since its formulation, the central dogma has stood the test of time and myriad experiments. It provides a strong unifying theme to molecular genetics and information flow in cell biology and biochemistry.

Although in many cases a gene encodes one polypeptide, other genes encode a **functional RNA**. Some genes encode **tRNAs** and **rRNAs** needed for translation, others encode other structural and catalytic RNAs. Genes encode some product that is used in the cell, i.e. that when altered generates an identifiable phenotype. More generally, genes encode RNAs, some of which are functional as transcribed (or with minor alterations via processing) such as tRNAs and rRNAs, and others are messengers that are then translated into proteins. These proteins can provide structural, catalytic and regulatory roles in the cell.

Note the **static role of DNA** in this process. Implicit in this model is the idea that DNA does not provide an active cellular function, but rather it encodes macromolecules that are functional. However, the expression of virtually all genes is highly regulated. The sites on the DNA where this control is exerted are indeed functional entities, such as promoters and enhancers. In this case, the DNA is directly functional (*cis*-regulatory sites), but the genes being regulated by these sites still encode some functional product (RNA or protein).

Studies of retroviruses lead Dulbecco to argue that the flow of information is not unidirectional, but in fact RNA can be converted into DNA (some viral RNA genomes are converted into DNA proviruses integrated into the genome). Subsequently Temin and Baltimore discovered the enzyme that can make a DNA copy of RNA, i.e. reverse transcriptase.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Transcription and mRNA structure

Several aspects of the structure of genes can be illustrated by examining the general features of a bacterial gene as now understood.

A gene is a string of nucleotides in the duplex DNA that encodes a mRNA, which itself codes for protein. Only one strand of the duplex DNA is copied into mRNA (Figure 1.22). Sometimes genes overlap, and in some of those cases each strand of DNA is copied, but each for a different mRNA. The strand of DNA that reads the same as the sequence of mRNA is the **nontemplate strand**. The strand that reads as the reverse complement of the mRNA is the **template strand**.

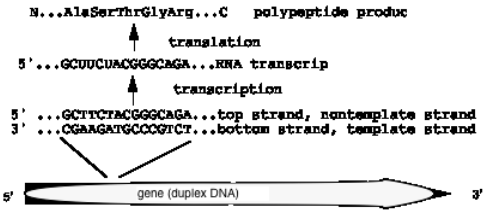


Figure 1.22.Only one strand of duplex DNA codes for a particular product.

Note

The term "sense strand" has two **opposite** uses (unfortunately). Sidney Brenner first used it to designate the strand that served as the template to make RNA (bottom strand above), and this is still used in many genetics texts. However, now many authors use the term to refer to the strand that reads the same as the mRNA (top strand above). The same confusion applies to the term "coding strand" which can refer to the strand encoding mRNA (bottom strand) or the strand "encoding" the protein (top strand). Interestingly, "antisense" is used exclusively to refer to the strand that is the reverse complement of the mRNA (bottom strand).

Figure 1.22 helps illustrate the origin of terms used in gene expression. Copying the information of DNA into RNA stays in the same "language" in that both of these polymers are nucleic acids, hence the process is called transcription. An analogy would be writing exercises where you had to copy, e.g. a poem, from a book onto your paper - you transcribed the poem, but it is still in English. Converting the information from RNA into DNA is equivalent to converting from one "language" to another, in this case from one type of polymer (the nucleic acid RNA) to a different one (a polypeptide or protein). Hence the process is called translation. This is analogous to translating a poem written in French into English.

Figure 1.23 illustrates the point that a gene may be longer than the region coding for the protein because of 5' and/or 3' **untranslated regions**.

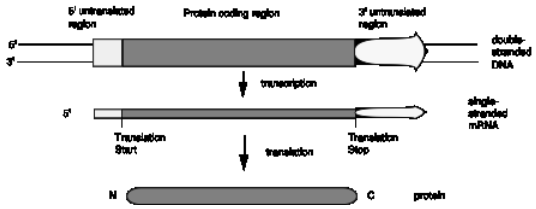


Figure 1.23: Genes and mRNA have untranslated sequences at both the 5’ and 3’ ends.

Eukaryotic mRNAs have covalent attachment of nucleotides at the 5' and 3' ends, and in some cases nucleotides are added internally (a process called *RNA editing*). Recent work shows that additional nucleotides are added post-transcriptionally to some bacterial mRNAs as well.

Regulatory signals can be considered parts of genes

In order to express a gene at the correct time, the DNA also carries signals to start transcription (e.g. promoters), signals for regulating the efficiency of starting transcription (e.g. operators, enhancers or silencers), and signals to stop transcription (e.g. terminators). Minimally, a gene includes the **transcription unit**, which is the segment of DNA that is copied into RNA in the primary transcript. The signals directing RNA polymerase to start at the correct site, and other DNA segments that influence the efficiency of this process are regulatory elements for the gene. One can also consider them to be part of the gene, along with the transcription unit.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Finding the Functions of Genes

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Genetic Methods in Microorganisms

The genetic systems found in bacteria and fungi are particularly powerful. The small size of the **genome** (all the genetic material in an organism), the ability to examine both haploid and diploid forms, and the ease of large-scale screens have made them the method of choice for many investigations.

- Genetic Methods Introduction
- Conjugation in Bacteria
- Gene mapping by conjugal transfer
- Bacteriophage

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Bacteriophage

Bacteriophage are viruses that infect bacteria. Because of their very large number of progeny and ability to recombine in mixed infections (more than one strain of bacteria in an infection), they have been used extensively in high-resolution definition of genes. Much of what we know about genetic fine structure, prior to the advent of techniques for isolating and sequencing genes, derive form studies in bacteriophage.

Bacteriophage have been a powerful model genetic system, because they have small genomes, have a short life cycle, and produce many progeny from an infected cell. They provide a very efficient means for transfer of DNA into or between cells. The large number of progeny makes it possible to measure very rare recombination events.

Lytic bacteriophage form **plaques** on lawns of bacteria; these are regions of clearing where infected bacteria have lysed. Early work focused on mutants with different **plaque morphology**, e.g. T2 *r*, which shows rapid lysis and generates larger plaques, or on mutants with **different host range**, e.g. T2 *h*, which will kill both host strains B and B/2.

A cis-trans complementation test defines a cistron, which is a gene

Seymour Benzer used the *rII*locus of phage T4 to define genes by virtue of their behavior in a complementation test, and also to provide fundamental insight into the structure of genes (in particular, the arrangement of mutable sites - see the next section). The difference in plaque morphology between *r*and *r*+phage is easy to see (large versus small, respectively), and Benzer isolated many *r* mutants of phage T4. The wild type, but not any *rII*mutants, will grow on *E. coli*strain K12(l), whereas both wild type and mutant phage grow equally well on *E. coli*strain B. Thus the wild phenotype is readily detected by its ability to grow in strain K12 (l).

If *E. coli*strain K12 (l) is co-infected with 2 phage carrying mutations at different positions in *rIIA*, you get no multiplication of the phage (except the extremely rare wild type recombinants, which occur at about 1 in 106 progeny). In the diagram below, each line represents the chromosome from one of the parental phage.

rIIA rIIB

phage 1 _|__x_____|_____|_

phage 2 _|_____x_|_____|_

Likewise, if the two phage in the co-infection carry mutations at different positions in *rIIB*, you get no multiplication of the phage (except the extremely rare wild type recombinants, about 1 in 106).

rIIA rIIB

phage 3 _|_____|_x_____|_

phage 4 _|_____|_____x_|_

However, if one of the co-infecting phage carries a mutation in *rIIA*and the other a mutation in *rIIB*, then you see multiplication of the phage, forming a very large number of plaques on *E. coli*strain K12 (l).

rIIA rIIB

phage 1 _|__x_____|_____|_ Provides wt rIIB protein

phage 4 _|_____|_____x_|_ Provides wt rIIA protein

Together these two phage provide all the phage functions - they **complement** each other. This is a positive complementation test. The first two examples show no complementation, and we place them in the same **complementation group**. Mutants that do not complement are placed in the same complementation group; they are different mutant alleles of the same gene. Benzer showed that there were two complementation groups (and therefore two genes) at the *r II* locus, which he called A and B.

Exercise 1.3

In the mixed infection with phage 1 and phage 4, you also obtain the rare wild type recombinants, but there are more recombinants than are seen in the co-infections with different mutant alleles. Why?

Benzer's experiments analyzing the *rII*locus of bacteriophage T4 formalized the idea of a ***cis-trans*complementation test** to define a **cistron**, which is an operational definition of a gene. First, let's define *cis* and *trans*when used to refer to genes. In the ***cis***

configuration, both mutations are on the same chromosome. In the ***trans*** configuration, each mutation is on a different chromosome

Mutations in the same gene will not complement in *trans*, whereas mutations in different genes will complement in *trans* (Figure 1.12). In the *cis* configuration, the other chromosome is wild type, and wild-type will complement any recessive mutation. The **complementation group** corresponds to a genetic entity we call a **cistron**, it is equivalent to a **gene**.

This test requires a diploid situation. This can be a natural diploid (2 copies of each chromosome) or a partial, or merodiploid, e.g. by conjugating with a cell carrying an F' factor. Some bacteriophage carry pieces of the host chromosome; these are called transducing **phage**. Infection of *E. coli* with a transducing phage carrying a mutation in a host gene is another way to create a merodiploid in the laboratory for complementation analysis.

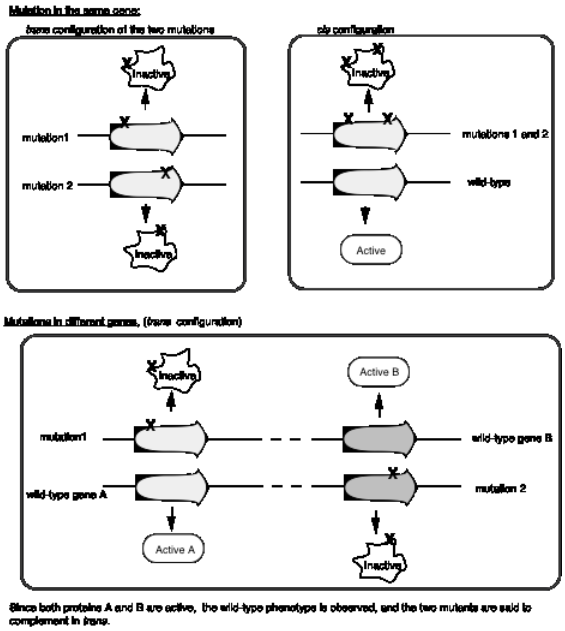


Figure 1.12. The complementation test defines the cistron and distinguishes between two genes.

Recombination within genes allows construction of a linear map of mutable sites that constitute a gene

Once the recombination analysis made it clear that chromosomes were linear arrays of genes, these were thought of as "string of pearls" with the genes, or "pearls," separated by some non-genetic material (Figure 1.13). This putative non-genetic material was thought to be the site of recombination, whereas the genes, the units of inheritance, were thought to be resistant to recombination. However, by examining the large number of progeny of bacteriophage infections, one can demonstrate that **recombination can occur within a gene**. This supports the second model shown in Figure 1.13. Because of the tight packing of coding regions in phage genomes, recombination almost always occurs within genes in bacteriophage, but in genomes with considerable non-coding regions between genes, recombination can occur between genes as well.

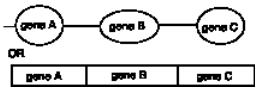


Figure 1.13.Models for genes as either discrete mutable units separate by non-genetic material (top) or as part of a continuous genetic material (bottom).

The tests between these two models required screening for genetic markers (mutations) that are very close to each other. When two markers are very close to each other, the recombination frequency is extremely low, so enough progeny have to be examined to resolve map distances of, say 0.02 centiMorgans = 0.02 map units = 0.02 % recombinants. This means that 2 out of 10,000 progeny will show recombination between two markers that are 0.02 map units apart, and obviously one has to examine at least 10,000 progeny to reliably score this recombination. That's the power of microbial genetics - you actually can select or screen through this many progeny, sometimes quite easily.

An example of recombination in phage is shown in Figure 1.14. Wild type T2 phage forms small plaques and kills only *E. coli*strain B. Thus different alleles of *h*can be distinguished by plating on a mixture of *E. coli*strains B and B/2. The phage carrying

mutant *h*allele will generate clear plaques, since they kill both strains. Phage with the wild type *h*+ give turbid plaques, since the B/2 cells are not lysed but B cells are. When a mixture of *E. coli*strains B and B/2 are co-infected with both T2 *hr*and T2 *h*+*r*+, four types of plaques are obtained. Most have the parental phenotypes, clear and large or turbid and small. These plaques contain progeny phage that retain the parental genotypes T2 *hr*and T2 *h*+*r*+, respectively. The other two phenotypes are nonparental, i.e. clear and small or turbid and large. These are from progeny with recombinant genotypes, i.e. T2 *hr*+and T2 *h*+*r*. In this mixed infection, recombination occurred between two phage genomes in the same cell.

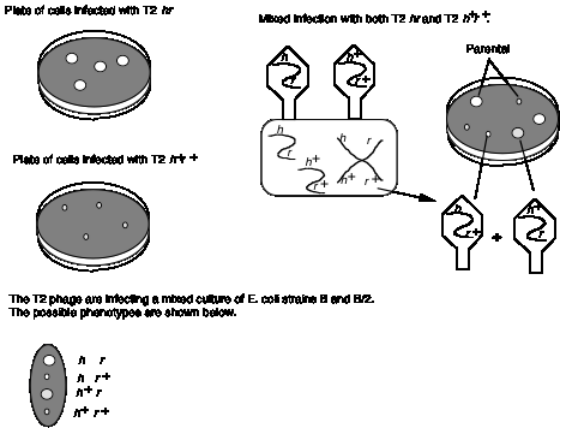


Figure 1.14. Recombination in bacteriophage

The first demonstration of recombination within a gene came from work on the *rIIA*and *rIIB*genes of phage T4. These experiments from Seymour Benzer, published in 1955, used techniques like that diagrammed in Figure 1.14. Remember that mutations in the *r*gene cause rapid lysis of infected cells, i.e. the length of the lytic cycle is shorter. The difference in plaque morphology between *r*and *r*+phage is easy to see (large versus small, respectively). These two genes are very close together, and many mutations were independently isolated in each. This was summarized in the discussion on complementation above.

Consider the results of infection of a bacterial culture with two mutant alleles of gene *rIIA.*

T4*rIIA6* _|_____x_____|_

and T4*rIIA27* _|_____x_____|_

(x marks the position of the mutation in each allele).

Progeny phage from this infection include those with a parental genotype (in the great majority), and at a much lower frequency, two types of recombinants:

wild type T4 *r*+ _|_____|_

double mutant T4*rIIA6 rIIA27* _|_____x_____x_____|_

The wild type is easily scored because it, and not any *rII*mutants, will grow on E. coli strain K12(l), whereas both wild type and mutant phage grow equally well on E. coli strain B. Thus you can **select**for the wild type (and you will see only the desired recombinant). Finding the double mutants is more laborious, because they are obtained only by screening through the progeny, testing for phage that when backcrossed with the parental phage result in no wild type recombinant progeny.

Equal numbers of wild type and double mutant recombinants were obtained, showing that recombination can occur within a gene, and that this occurs by reciprocal crossing over. If recombination were only between genes, then no wild type phage would result. A large spectrum of recombination values was obtained in crosses for different alleles, just like you obtain for crosses between mutants in separate genes.

Several major conclusions could be made as a result of these experiments on recombination within the *rII*genes.

1. A **large number of mutable sites** occur within a gene, exceeding some 500 for the *rIIA* and *rIIB*genes. We now realize that these correspond to the **individual base pairs** within the gene.
2. The **genetic maps are clearly linear**, indicating that the gene is linear. Now we know a gene is a linear polymer of nucleotides.
3. Most mutations are changes at one mutable site (**point mutations**). Many genes can be restored to wild type by undergoing a reverse mutation at the same site (**reversion**).

4. Other mutations cause the **deletion**of one or more mutable sites, reflecting a physical loss of part of the *rII* gene. Deletions of one or more mutable site (base pair) are extremely unlikely to revert back to the original wild type.

One gene encodes one polypeptide

One of the fundamental insights into how genes function is that **one gene encodes one enzyme** (or more precisely, one **polypeptide**). Beadle and Tatum reached this conclusion based on their complementation analysis of the genes required for arginine biosynthesis in fungi. They showed that a mutation in each gene led to a loss of activity of one enzyme in the multistep pathway of arginine biosynthesis. As discussed above in the section on genetic dissection, a large number of Arg auxotrophs (requiring Arg for growth) were isolated, and then organized into a set of complementation groups, where each complementation groups represents a gene.

The classic work of Beadle and Tatum demonstrated a direct relationship between the genes defined by the auxotrophic mutants and the enzymes required for Arg biosynthesis. They showed that a mutation in one gene resulted in the loss of one particular enzymatic activity, e.g. in the generalized scheme below, a mutation in gene 2 led to a loss of activity of enzyme 2. This led to an accumulation of the substrate for that reaction (intermediate N in the diagram below). If there were 4 complementation groups for the Arg auxotrophs, i.e. 4 genes, then 4 enzymes were found in the pathway for Arg biosynthesis. Each enzyme was affected by mutations in one of the complementation groups.

Intermediates:

M ® N ® O ® P ® Arg

enzyme 1 enzyme 2 enzyme 3 enzyme 4

gene 1 gene 2 gene 3 gene 4

Figure 1.15. A general scheme showing the relationships among metabolic intermediates (M, N, O, P), and end product (Arg), enzymes and the genes that encode them.

In general, each step in a metabolic pathway is catalyzed by an enzyme (identified biochemically) that is the product of a particular gene (identified by mutants unable to synthesize the end product, or unable to break down the starting compound, of a pathway). The number of genes that can generate auxotrophic mutants is (usually) the same as the number of enzymatic steps in the pathway. Auxotrophic mutants in a given gene are missing the corresponding enzyme. Thus Beadle and Tatum concluded that one gene encodes one enzyme. Sometimes more than one gene is required to encode an enzyme because the enzyme has multiple, different polypeptide subunits. **Thus each polypeptide is encoded by a gene**.

The metabolic intermediates that accumulate in each mutant can be used to place the enzymes in their **order of action**in a pathway. In the diagram in Figure 1.15, mutants in gene 3 accumulated substance O. Feeding substance O to mutants in gene 1 or in gene 2 allows growth in the absence of Arg. We conclude that the defects in enzyme 1 or enzyme 2, respectively, are upstream of enzyme 3. In contrast, feeding substance O to mutants in gene 4 will not allow growth in the absence of Arg. Even though this mutant can convert substance O to substance P, it does not have an active enzyme 4 to convert P to Arg. The inability of mutants in gene 4 to grow on substance O shows that enzyme 4 is downstream of enzyme 3.

Exercise 1.4

Imagine that you are studying serine biosynthesis in a fungus. You isolate serine auxotrophs, do all the pairwise crosses of the mutants and discover that the auxotrophs can be grouped into three complementation groups, called A, B and C. You also discover that a different metabolic intermediate accumulates in members of each complementation group - substance A in auxotrophs in the A complementation group, substance B in the B complementation group and substance C in the C complementation group. Each of the intermediates is fed to auxotrophs from each of the three complementation groups as tabulated below. A + means that the auxotroph was able to grow in media in the absence of serine when fed the indicated substance; a - denotes no growth in the absence of serine.

Fed:	mutant in complementation group A	mutant in complementation group B	mutant in complementation group C
substance A	-	+	+
substance B	-	-	-

Fed:	mutant in complementation group A	mutant in complementation group B	mutant in complementation group C
substance C	-	+	-

In the biosynthetic pathway to serine in this fungus, what is the order of the enzymes encoded in the three complementation groups? Enzyme A is encoded by the gene that when altered generates mutants that fall into complementation group A, etc.

The gene and its polypeptide product are colinear

Once it was determined that a gene was a linear array of mutable sites, that genes are composed of a string of nucleotides called DNA (see Chapter 2), and that each gene encoded a polypeptide, the issue remained to be determined how exactly that string of nucleotides coded for a particular amino acid sequence. This problem was studied along several avenues, culminating in a major achievement of the last half of the 20th century – the deciphering of the genetic code. The detailed assignment of particular codons (triplets of adjacent nucleotides) will be discussed in Chapter 13. In the next few sections of this chapter, we will examine how some of the basic features of the genetic code were deciphered.

A priori, the coding units within a gene *could*encode both the composition and the address for each amino acid, as illustrated in Model 1 of Figure 1.17. In this model, the coding units could be scrambled and still specify the same protein. In such a situation, the polypeptide would not be colinear with the gene.

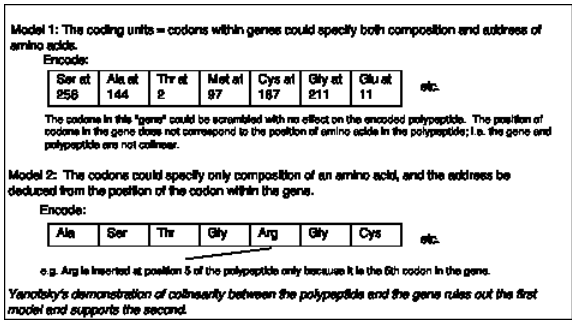


Figure 1.16.Alternative models for gene and codon structure.

In an alternative model (Model 2 in Figure 1.16), the coding units only specify the composition, but not the position, of an amino acid. The "address" of the amino acid is derived from the position of the coding unit within the gene. This model would predict that the gene and its polypeptide product would be colinear - e.g. mutation in the 5th coding unit would affect the 5th amino acid of the protein, etc.

Charles Yanofsky and his co-workers (1964) tested these two models and determined that the **gene and the polypeptide product are indeed colinear**. They used recombination frequencies to map the positions of different mutant alleles in the gene that encodes a particular subunit of the enzyme tryptophan synthase. They then determined the amino acid sequence of the wild type and mutant polypeptides. As illustrated in Figure 1.17, the position of a mutant allele on the recombination map of the gene corresponds with the position of the amino acid altered in the mutant polypeptide product. For instance, allele *A101* maps to one end of the gene, and the corresponding Glu ® Val replacement is close to the N terminus of the polypeptide. Allele *A64*maps close to the other end of the gene, and the corresponding Ser ® Leu replacement is close to the C terminus of the polypeptide. This correspondence between the positions of the mutations in each allele and the positions of the consequent changes in the polypeptide show that Model 1 can be eliminated and Model 2 is supported.

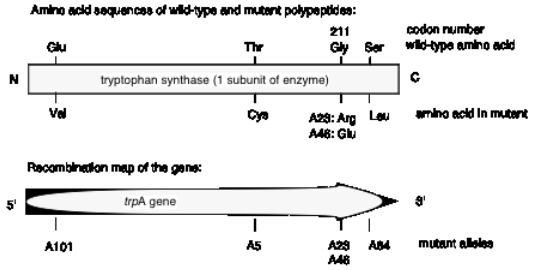


Figure 1.17.The polypeptide is colinear with the gene.

Mutable sites are base pairs along the double helix

The large number of mutable sites found in each gene, and between which recombination can occur, leads one to conclude that the mutable sites are base pairs along the DNA. Sequence determination of the wild type and mutant genes confirms this conclusion.

Single amino acids are specified by three adjacent nucleotides, which are a codons

This conclusion requires three pieces of information.

First of all, **adjacent mutable sites specify amino acids**. Reaching this conclusion required investigation of the fine structure of a gene, including rare recombination between very closely linked mutations within a gene. Yanofsky and his colleagues, working with mutations the *trpA* gene of *E. coli*, encoding tryptophan synthase, showed that different alleles mutated in the same codon could recombine (albeit at very low frequency). (This is the same laboratory and same system that was used to show that a gene and its polypeptide product are colinear.) Thus recombination between two different alleles can occur within a codon, which means that a codon must have more than one mutable site. We now recognize that a mutable site is a nucleotide in the DNA. Thus adjacent mutable sites (nucleotides) encode a single amino acid.

Let's look at this in more detail (Figure 1.18). Yanofsky and colleagues examined two different mutant alleles of *trpA*, each of which caused alteration in amino acid 211 of tryptophan synthase. In the mutant allele *A23*, wild type Gly is converted to mutant Arg. In the mutant allele *A46*, wild type Gly is converted to mutant Glu.

GGA (Gly 211) --> AGA (Arg 211) mutant allele A23

GGA (Gly 211) --> GAA (Glu 211) mutant allele A46

A23 ´ A46 AGA ´ GAA ® GGA (wild type Gly 211 in 2 out of 100,000 progeny)

Figure 1.18. Recombination can occur between two mutant alleles affecting the same codon.

Alleles *A23*and *A46*are not alternative forms of the same mutable site, because recombination to yield wild type occurs, albeit at a very low frequency (0.002%; the sites are very close together, in fact in the same codon!). If they involved the same mutable site, one would never see the wild-type recombinant.

The second observation is that the **genetic code is non-overlapping**. This was shown by demonstrating that a mutation at a single site alters only one amino acid. This conflicts with the predictions of an overlapping code (see Figure 1.19), and thus the code must be non-overlapping.

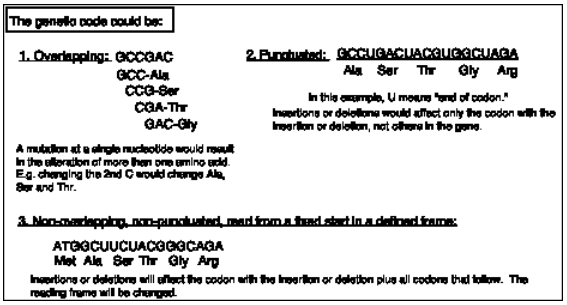


Figure 1.19. Predictions of the effects of nucleotide substitutions, insertions or deletions on polypeptides encoded by an overlapping, a punctuated, or a nonoverlapping, nonpunctuated code.

The third observation is that the **genetic code is read in triplets** from a fixed starting point. This was shown by examining the effect of **frameshift mutations**. As shown in Figure 1.19, a code lacking punctuation has a certain reading frame. Insertions or deletions of nucleotides are predicted to have a drastic effect on the encoded protein because they will change that reading frame. The fact that this was observed was one of the major reasons to conclude that the mRNA molecules encoded by genes are read in successive blocks of three nucleotides in a particular reading frame.

For the sequence shown in Figure 1.20, insertion of an A shifts the reading frame, so all amino acids after the insertion differ from the wild type sequence. (The 4th amino acid is still a Gly because of degeneracy in the code: both GGC and GGG code for Gly.) Similarly, deletion of a U alters the entire sequence after the deletion.

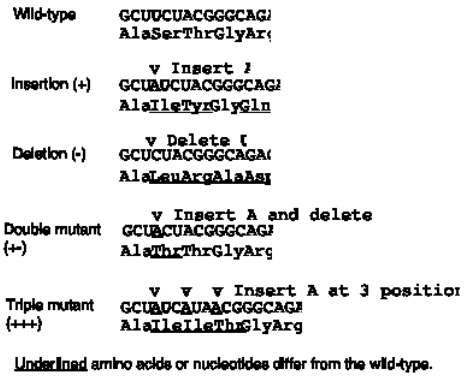


Figure 1.20. Frameshift mutations show that the genetic code is read in triplets.

These observations show that the nucleotide sequence is read, or translated, from a fixed starting point without punctuation. An alternative model is that the group of nucleotides encoding an amino acid (the codon) could also include a signal for the end of the codon (Model 2 in Figure 1.19). This could be considered a "comma" at the end of each codon. If that were the case, insertions or deletions would only affect the codon in which they occur. However, the data show that all codons, including and after the one containing the insertion or deletion, are altered. Thus the genetic code is not punctuated, but is read in a particular frame that is defined by a fixed starting point (Model 3 in Figure 1.19). That starting point is a particular AUG, encoding methionine. (More about this will be covered in Chapter 13).

The results of frame-shift mutations are so drastic that the proteins are usually not functional. Hence a screen or selection for loss-of-function mutants frequently reveals these frameshift mutants. Simple nucleotide substitutions that lead to amino acid replacements often have very little effect on the protein, and hence have little, or subtle, phenotypes.

A double mutant generated by crossing over between the insertion (+) and deletion (-) results in an (almost) normal phenotype, i.e. reversion of insertion or deletion.

A gene containing **three closely spaced insertions** (or deletions) of single nucleotides will produce a **functional product**. However, four or five insertions or deletions do not give a functional product (Crick, Barnett, Brenner and Watts-Tobin, 1961). This provided the best evidence that the **genetic code is read in groups of three nucleotides**(not two or four). Over the next 5 years the code was worked out (by 1966) and this inference was confirmed definitively.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Conjugation in Bacteria

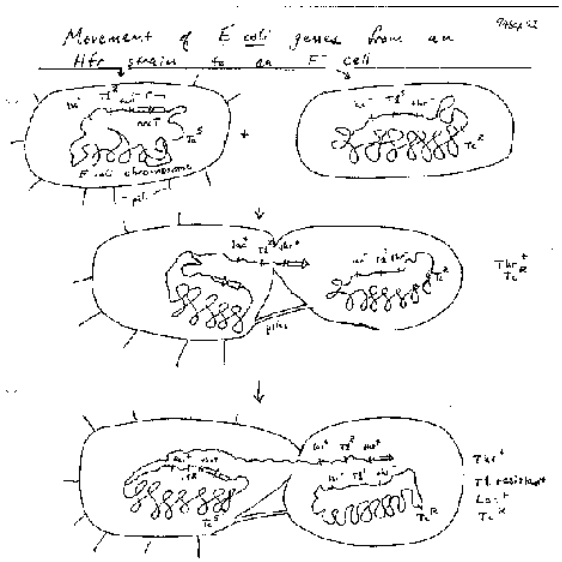


Figure 1.10. F-factor mediated conjugal transfer of DNA in bacteria.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Gene mapping by conjugal transfer

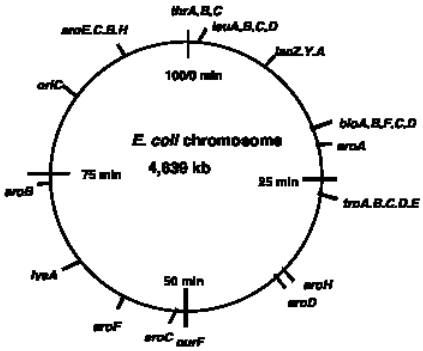


Figure 1.11.Circular genetic map of *E. coli*.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Genetic Methods Introduction

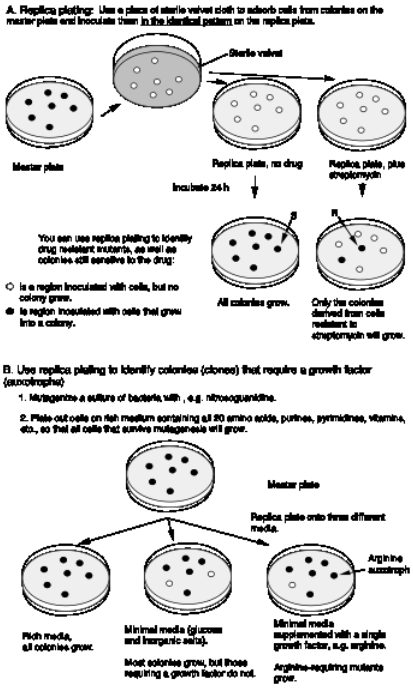


Figure 1.9. Coat color in Siamese cats is determined by a temperature sensitive mutation in an enzyme needed for pigment formation. Siamese are homozygous $c^h c^h$, which encodes an enzyme that is active at low temperature (in the extremities of the cat) but inactive elsewhere.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

CHAPTER OVERVIEW

2: Structures of Nucleic Acids

DNA and RNA are both **nucleic acids**, which are the polymeric acids isolated from the nucleus of cells. DNA and RNA can be represented as simple strings of letters, where each letter corresponds to a particular **nucleotide**, the monomeric component of the nucleic acid polymers. Although this conveys almost all the information content of the nucleic acids, it does not tell you anything about the underlying chemical structures. This chapter will be review the evidence that nucleic acids are the genetic material, and then exploring the chemical structure of nucleic acids.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Welcome to the Biology Library. This Living Library is a principal hub of the LibreTexts project, which is a multi-institutional collaborative venture to develop the next generation of open-access texts to improve postsecondary education at all levels of higher learning. The LibreTexts approach is highly collaborative where an Open Access textbook environment is under constant revision by students, faculty, and outside experts to supplant conventional paper-based books.

2.5: B-Form, A-Form, and Z-Form of DNA

Three major forms of DNA are double stranded and connected by interactions between complementary base pairs. These are terms A-form, B-form,and Z-form DNA.

B-form DNA

The information from the base composition of DNA, the knowledge of dinucleotide structure, and the insight that the X-ray crystallography suggested a helical periodicity were combined by Watson and Crick in 1953 in their proposed model for a double helical structure for DNA. They proposed two strands of DNA -- each in a right-hand helix -- wound around the same axis. The two strands are held together by H-bonding between the bases (in anti conformation) as shown in Figure 2.5.1.

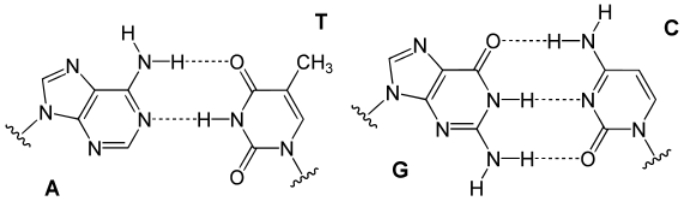


Figure* 2.5.1*: *(left) An A:T base pair and (right) a G:C base pair*

Bases fit in the double helical model if pyrimidine on one strand is always paired with purine on the other. From **Chargaff's rules**, the two strands will pair A with T and G with C. This pairs a keto base with an amino base, a purine with a pyrimidine. Two H-bonds can form between A and T, and three can form between G and C. This third H-bond in the G:C base pair is between the additional exocyclic amino group on G and the C2 keto group on C. The pyrimidine C2 keto group is not involved in hydrogen bonding in the A:T base pair.

These are the complementary base pairs. The base-pairing scheme immediately suggests a way to replicate and copy the the genetic information.

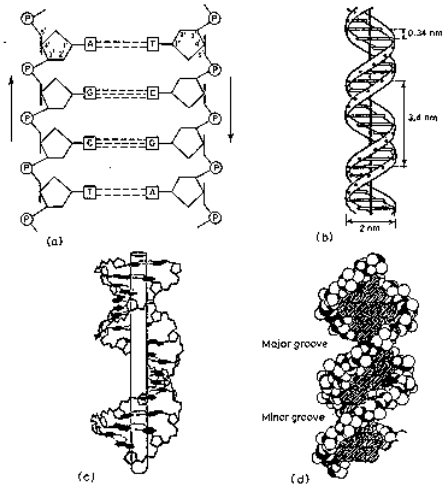


Figure 2.5.2: Antiparallel (a), plectonemically coiled (b, c, d) DNA strands. The arrows in a are pointed 3’ to 5’, but they illustrate the antiparallel nature of the duplex. The two strands of the duplex are antiparallel and plectonemically coiled. The nucleotides arrayed in a 5' to 3' orientation on one strand align with complementary nucleotides in the the 3' to 5' orientation of the opposite strand.

The two strands are not in a simple side-by-side arrangement, which would be called a *paranemic joint* (Figure 2.5.3). (This will be encountered during recombination in Chapter 8.) Rather the two strands are coiled around the same helical axis and are intertwined with themselves (which is referred to as a **plectonemic coil**). One consequence of this intertwining is that the two strands cannot be separated without the DNA rotating, one turn of the DNA for every "untwisting" of the two strands.

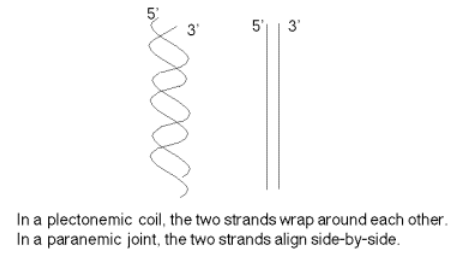


Figure 2.5.3: *Duplex DNA has the two strands wrapped around each other in a plectonemic coil (left), not a paranemic duplex (right).*

Dimensions of B-form (the most common) of DNA

- 0.34 nm between bp, 3.4 nm per turn, about 10 bp per turn
- 1.9 nm (about 2.0 nm or 20 Angstroms) in diameter

Major and minor groove

The major groove is wider than the minor groove in DNA (Figure 2.5.2*d*), and many sequence specific proteins interact in the major groove. The N7 and C6 groups of purines and the C4 and C5 groups of pyrimidines face into the major groove, thus they can make specific contacts with amino acids in DNA-binding proteins. Thus specific amino acids serve as H-bond donors and acceptors to form H-bonds with specific nucleotides in the DNA. H-bond donors and acceptors are also in the minor groove, and indeed some proteins bind specifically in the minor groove. Base pairs stack, with some rotation between them.

A-form nucleic acids and Z-DNA

Three different forms of duplex nucleic acid have been described. The most common form, present in most DNA at neutral pH and physiological salt concentrations, is B-form. That is the classic, right-handed double helical structure we have been discussing. A thicker right-handed duplex with a shorter distance between the base pairs has been described for RNA-DNA duplexes and RNA-RNA duplexes. This is called A-form nucleic acid.

A third form of duplex DNA has a strikingly different, left-handed helical structure. This Z DNA is formed by stretches of alternating purines and pyrimidines, e.g. GCGCGC, especially in negatively supercoiled DNA. A small amount of the DNA in a cell exists in the Z form. It has been tantalizing to propose that this different structure is involved in some way in regulation of some cellular function, such as transcription or regulation, but conclusive evidence for or against this proposal is not available yet.

Differences between A-form and B-form nucleic acid

The major difference between A-form and B-form nucleic acid is in the conformation of the deoxyribose sugar ring. It is in the C2' endoconformation for B-form, whereas it is in the C3' endoconformation in A-form. As shown in Figure 2.5.4, if you consider the plane defined by the C4'-O-C1' atoms of the deoxyribose, in the C2' endoconformation, the C2' atom is above the plane, whereas the C3' atom is above the plane in the C3' endoconformation. The latter conformation brings the 5' and 3' hydroxyls (both esterified to the phosphates linking to the next nucleotides) closer together than is seen in the C2' endoconfromation (Figure 2.16). Thus the distance between adjacent nucleotides is reduced by about 1 Angstrom in A-form relative to B-form nucleic acid (Figure 2.5.4).

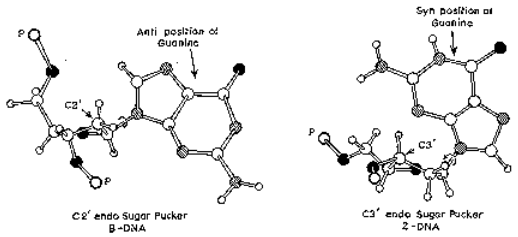


Figure 2.5.4: Syn and anti conformations of the base relative to the sugar in nucleotides.

A second major difference between A-form and B-form nucleic acid is the placement of base-pairs within the duplex. In B-form, the base-pairs are almost centered over the helical axis (Figure 2.5.4), but in A-form, they are displaced away from the central axis and closer to the major groove. The result is a ribbon-like helix with a more open cylindrical core in A-form.

Z-form DNA

Z-DNA is a radically different duplex structure, with the two strands coiling in left-handed helices and a pronounced zig-zag (hence the name) pattern in the phosphodiester backbone. As previously mentioned, Z-DNA can form when the DNA is in an alternating purine-pyrimidine sequence such as GCGCGC, and indeed the G and C nucleotides are in different conformations, leading to the zig-zag pattern. The big difference is at the G nucleotide. It has the sugar in the C3' endoconformation (like A-form nucleic acid, and in contrast to B-form DNA) and the guanine base is in the synconformation. This places the guanine back over the sugar ring, in contrast to the usual anticonformation seen in A- and B-form nucleic acid. Note that having the base in the anticonformation places it in the position where it can readily form H-bonds with the complementary base on the opposite strand. The duplex in Z-DNA has to accomodate the distortion of this G nucleotide in the synconformation. The cytosine in the adjacent nucleotide of Z-DNA is in the "normal" C2' endo, anticonformation.

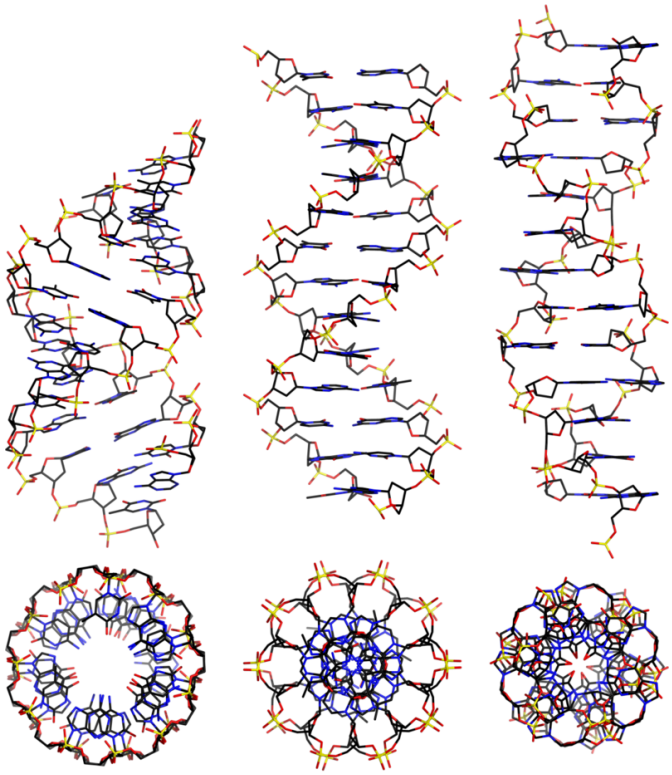


Figure 2.5.5: B-form (left), A-form (middle) and Z-DNA (right). (CC BY-SA 4.0; Mauroesguerroto)

Even classic B-DNA is not completely uniform in its structure. X-ray diffraction analysis of crystals of duplex oligonucleotides shows that a given sequence will adopt a distinctive structure. These variations in B-DNA may differ in the propeller twist (between bases within a pair) to optimize base stacking, or in the 3 ways that 2 successive base pairs can move relative to each other: twist, roll, or slide.

Table 2.5.1 *Comparisons of B-form, A-form and Z-DNA*

	B-Form	A-Form	Z-Form
helix sense	Right Handed	Right Handed	Left Handed
base pairs per turn	10	11	12
vertical rise per bp	3.4 Å	2.56 Å	19 Å
rotation per bp	+36°	+33°	-30°
helical diameter	19 Å	19 Å	19 Å

2.8: Intro

DNA and RNA are both **nucleic acids**, which are the polymeric acids isolated from the nucleus of cells. DNA and RNA can be represented as simple strings of letters, where each letter corresponds to a particular **nucleotide**, the monomeric component of the nucleic acid polymers. Although this conveys almost all the information content of the nucleic acids, it does not tell you anything about the underlying chemical structures. This chapter will be review the evidence that nucleic acids are the genetic material, and then exploring the chemical structure of nucleic acids.

Southern blot-hybridizations

After separation by electrophoresis, DNA fragments are transferred to a membrane (nylon or nitrocellulose) and immobilized; this replica of the DNA pattern in the gel is called a "blot." A specific labeled probe is hybridized to the blot to detect related sequences. After nonspecifically bound probe is washed away, the specific hybrids are detected by autoradiography of the blot.

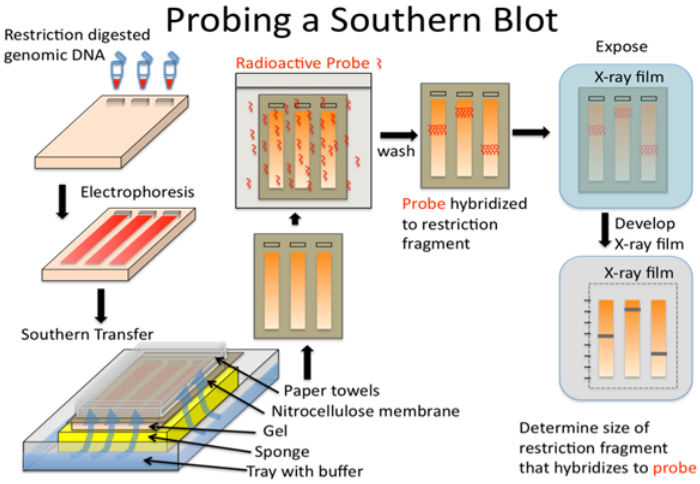


Figure 2.28: A diagram of Southern blotting. Genomic DNA that has been digested with a restriction enzyme is separated on an agarose gel, then the DNA is transferred from the gel to a nylon membrane (grey sheet) by blotting. The DNA is immobilized on the membrane, then probed with a radioactively labeled DNA fragment that is complementary to a target sequence. After stringent washing, the blot is exposed to X-ray film to detect what size frament the probe is bound. In this case, the probe bound to different-sized fragments in lanes 1, 2, and 3. In the last image the orange represent the position of the digested DNA, but it is not actually present on the X-ray film. (Original-J. Locke-CC:AN)

Restriction sites can be used as genetic markers. One can identify restriction fragment length polymorphisms (RFLPs) that are linked to a particular locus. This can be be used to

1. Develop a diagnostic test for a disease locus (e.g. sickle cell disease)
2. Help isolate the gene.
3. DNA fingerprinting for highly variable loci.

Sizes of DNAs and chromosomes, and methods to resolve them

The next figure presents views of chromosomes and DNA segments on four different, expanding scales. The top level compares the sizes of intact chromosomes from four of the organisms we will be discussing in this course. The scale on yeast chromosome III is then expanded so that it can be compared to some of the viral and plasmid genomes that are in common use. Next, a higher resolution view of the plasmid pBR322 is given, and finally the highest resolution that we are usually concerned with, i.e. the nucleotide sequence.

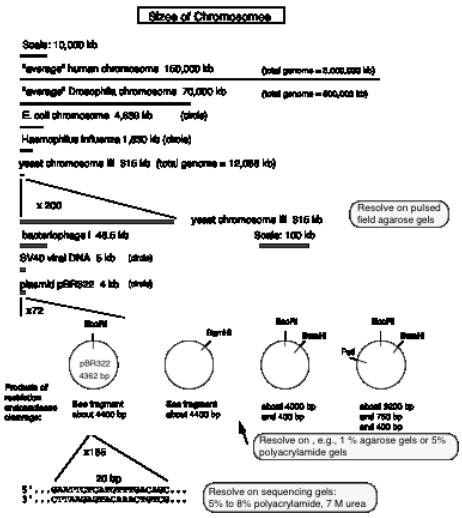


Figure 2.29.

Determining the sequence of DNA and RNA

The basic approach is to **generate a nested set of DNA fragments**that start a common site and **end in either A, G ,C or T**. These sets of (labeled) DNA fragments are separated on a denaturing polyacrylamide gel that has a resolution of 1 bp. The resulting pattern allows the sequence to be read. Base-specific chemical modification and degradation, developed by Maxam and Gilbert, was a widely used approach. Nucleotide-specific cleavage of RNA by a set of Rnases can be used to sequence RNA. We will focus on the most common method of sequencing DNA, that of nucleotide-specific chain termination.

The **dideoxynucleotide chain termination method**was developed in the laboratory of Fred Sanger at Cambridge. A 2', 3' dideoxynucleotide can be incorporated into DNA, as directed by the template strand. However, the missing 3'-OH precludes further polymerization. Hence the newly synthesized chain of nucleotides ends at **base-specific**, **chain terminating** dideoxynucleotide. Reactions are run such that all the products end in a G, a C, an A, or aT, but they all begin at the same place. This generates a nested set of products whose length is a measure of the position of all G's in a target sequence, or all C's, etc. Thus one can deduce that the target sequence is complementary to, e.g. G at position 1, T at positon 2, C at positions 3 and 4, etc. for hundreds of nucleotides per run.

In more detail, a specific primer is annealed to the template, upstream from the region to be sequenced. DNA polymerase will catalyze the synthesis of new DNA from the 3' end of that primer (elongation). The primer therefore generates a common end to all the product fragments. (This is the basis for the nested set in this approach).

The synthesized DNA is labeled with either a radioactive nucleotide, such as [a35S]deoxy-thio-ATP, or a fluorescent dye, often attached to the primer.

A base-specific chain-terminator is included in each of four reactions:

- 2',3' dideoxyGTP in the "G" reaction.
- 2',3' dideoxyATP in the "A" reaction.
- 2',3' dideoxyTTP in the "T" reaction.
- 2',3' dideoxyCTP in the "C" reaction.

The DNA polymerase will elongate from each annealed primer until it incorporates a 2', 3' dideoxynucleotide. No additional nucleotides can be added to this product, since it has no 3' OH, thus it is a chain-terminator. This termination occurs only at G residues (complementary to C's in the template) in the "G" reaction, only at A residues in the "A" reaction, etc. Thus the products of each reaction comprise a nested set of fragments, with the specific primer at the 5' end and the base-specific chain terminator at the 3' end. The products are resolved on a sequencing gel, exposed to X-ray film and the sequence read, as in Figure 2.30.

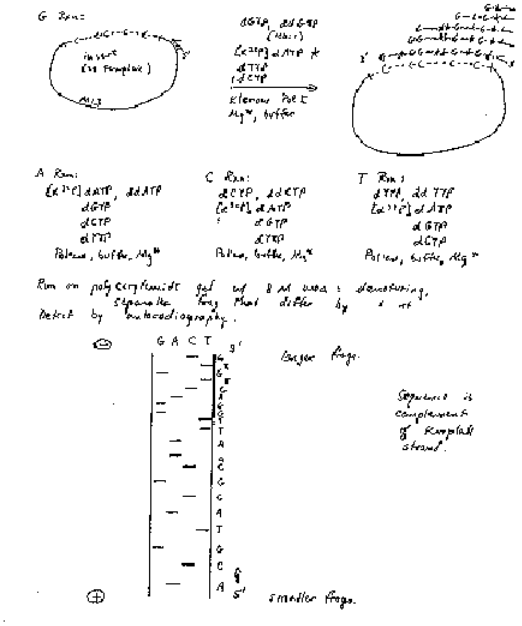


Figure 2.30. Sanger dideoxynucleotide chain termination sequencing.

The dideoxynucleotide chain-termination approach is the method used in **automated sequenators**. Different color fluorescent dyes (usually attached to the primer) are included in each base-specific reaction. Therefore the products of all four can be run in 1 lane of the resolving gel, allowing >20 sequencing sets to be analyzed at one time. A laser scans continuously along one zone of the gel, and records when a (e.g.) red, green, blue or yellow fluoresence is detected in each lane, meaning that the primer extended to a (e.g.) A, G, C or T is passing through the detection zone. These data are automatically processed, and a readout is generated with the peaks for each fluorescent dye as function of time of the gel running and the deduced sequence. An example of the output is shown below in black-and-white; the original output is in color (a different color for each nucleotide). Manual editing of the deduced sequence can be done based on the raw data, but in large scale sequencing projects, each region is determined about 8 different times and other software is used to determine the most frequently ocurring nucleotide at each position.

The capacity of automated sequencing machines is extraordinary. New machines using capillary gel electrophoresis are used to generated millions of nucleotides per day in the major sequencing centers. This technology allows large, complex genomes to be sequenced rapidly, as discussed in Chapter 4.

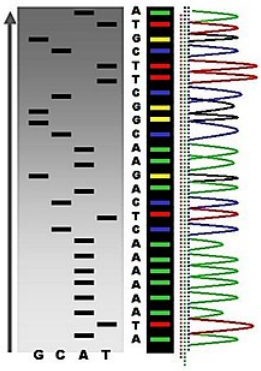


Figure 2.31. Example of output from automated dideoxynucleotide sequencing. An example of the results of automated chain-termination DNA sequencing. (CC BY-SA 3.0; Abizar Lakdawalla).

Supercoiling of Topologically Constrained DNA

Topologically closed DNA can be circular (covalently closed circles) or loops that are constrained at the base. The coiling (or wrapping) of duplex DNA around its own axis is called **supercoiling** (Figure 2.32 middle).

- **Negative** supercoils twist the DNA about its axis in the opposite direction from the clockwise turns of the right-handed (R-H) double helix.

- Negatively supercoiled DNA is underwound (and thus favors unwinding of duplex).
- Negatively supercoiled DNA has R-H supercoil turns (Figure 2.32).
- **Positive** supercoils twist the DNA in the same direction as the turns of the R-H double helix.
- Positively supercoiled DNA is overwound (helix is wound more tightly).
- Positively supercoiled DNA has L-H supercoil turns.

The clockwise turns of R-H double helix (A or B form) generate a positive Twist (T); see Figure 2.32 left. The couterclockwise (ccw) turns of L-H helix (Z) generate a negative T.

T= Twisting number

- For B form DNA, it is + (# bp/10 bp per twist)
- For A form DNA, it is + (# bp/11 bp per twist)
- For Z DNA, it is - (# bp/12 bp per twist)

W= Writhing Numberis the turning of the axis of the DNA *duplex* in space

- Relaxed molecule W=0
- Negative supercoils, W is negative
- Positive supercoils, W is positive

L= Linking number= total number of times one strand of the double helix (of a closed molecule) encircles (or links) the other.

$$L = W + T \tag{2.8.1}$$

- L cannot change unless one or both strands are broken and reformed.
- A change in the linking number, DL, is partitioned between T and W (Figure 2.32 right). Thus:

$$DL = DW + DT \tag{2.8.2}$$

if $DL = 0$, $DW = -DT$

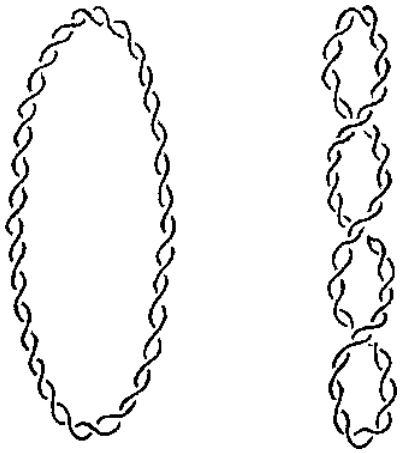


Figure 2.32. Relationship between supercoiling and twisting. Drawing showing the difference between a circular DNA chromosome (a plasmid) with a secondary helical twist only, and one containing an additional tertiary superhelical twist superimposed on the secondary helical winding. (CC BY-SA 3.0; JoKalliauer).

Ethidium Bromide intercalates in DNA, and untwists (or unwinds) the duplex by -27° per molecule of ethidium bromide intercalated. Thus intercalation of 14 molecules of ethidium bromide will untwist the duplex by 378o, i.e. slightly more than one full twist (which would be 360°). For this process of intercalation, DL=0, since no covalent bonds in the DNA are broken or reformed. The change in twist, DT, is negative, and thus DW is positive. Thus intercalation of ethidium bromide can relax a negatively supercoiled circle, and further intercalation will make the DNA positively supercoiled (Figure 2.33).

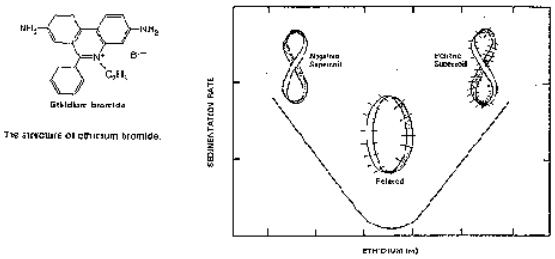


Figure 2.33.

It is useful to have an expression for supercoiling that is independent of length. The *superhelical density* is simply the number of superhelical (S.H.) turns per turn (or twist) of double helix.

$$\text{Superhelical density} = s = \frac{W}{T} \tag{2.8.3}$$

This is -0.05 for natural bacterial DNA. i.e., in bacterial DNA, there is 1 negative S.H. turn per 200 bp (calculated from 1 negative S.H. turn per 20 twists = 1 negative S.H. turn per 200 bp)

Negative supercoiled DNA has energy stored that favors unwinding, or a transition from B-form to Z DNA.

For **s** = -0.05, $\Delta G = -9 Kcal/mole$, which favoring unwinding

Thus negative supercoiling could favor initiation of transcription and initiation of replication.

Topoisomerases

Topoisomerases catalyze a change in the linking number of DNA.

- Topo I = nicking-closing enzyme, can relax positive or negative supercoiled DNA, makes a transient break in 1 strand. *E. coli* Topo I specifically relaxes negatively supercoiled DNA. Calf thymus Topo I works on both negatively and positively supercoiled DNA.
- Topo II = gyrase: uses the energy of ATP hydrolysis to introduce negative supercoils. Its mechanism of action is to make a transient double strand break, pass a duplex DNA through the break, and then re-seal the break.

Measuring a change in linking number

One can measure a change in linking number (DL) by sedimentation, electrophoresis, or electron microscopy, as illustrated in Figure 2.34.

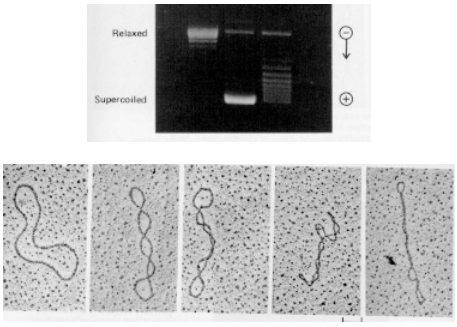


Figure 2.34.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

CHAPTER OVERVIEW

3: Isolating and Analyzing Genes

The first two chapters covered many important aspects of genes, such as how they function in inheritance, how they code for protein (in general terms) and their chemical nature. All this was learned without having a single gene purified. A full understanding of a gene, or the entire set of genes in a genome, requires that they be isolated and then studied intensively. Once a gene is “in hand”, in principal one can determine both its biochemical structures and its function(s) in an organism. One of the goals of biochemistry and molecular genetics is to assign particular functions to individual or composite structures. This chapter covers some of the techniques commonly used to isolate genes and illustrates some of the analyses that can be done on isolated genes.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

3.1: Recombinant DNA, Polymerase Chain Reaction and Applications to Eukaryotic Gene Structure and Function

Methods to purify some abundant proteins were developed early in the 20th century, and some of the experiments on the fine structure of the gene (colinearity of gene and protein for *trp*A and tryptophan synthase) used microbial genetics and proteins sequencing. However, methods to isolate genes were not developed until the 1960's, and the were applicable to only a few genes.

All this changed in the late 1970's with the development of recombinant DNA technology, or molecular cloning. This technique enabled researchers to isolate any gene from any organism from which one could isolate intact DNA (or RNA). The full potential to provide access to all genes of organisms is now being realized as full genomes are sequenced. One of the by-products of the intense investigation of individual DNA molecules after the advent of recombinant DNA was a procedure to isolate any DNA for which one knows the sequence. This technique, called the polymerase chain reaction (PCR), is far easier than traditional molecular cloning methods, and it has become a staple of many laboratories in the life sciences. After covering the basic techniques in recombinant DNA technology and PCR, their application to studies of eukaryotic gene structure and function will be discussed.

Like many advances in molecular genetics, recombinant DNA technology has its roots in bacterial genetics.

Transducing Phage

The first genes isolated were bacterial genes that could be picked up by bacteriophage. By isolating these hybrid bacteriophage, the DNA for the bacterial gene could be recovered in a highly enriched form. This is the basic principal behind recombinant DNA technology.

Some bacteriophage will integrate into a bacterial chromosome and reside in a dormant state (Figure 3.1). The integrated phage DNA is called a **prophage**, and the bacterium is now a **lysogen**. Phage that do this are **lysogenic**. Induction of the lysogen will result in excision of the prophage and multiplication to produce many progeny, i.e. it enters a **lytic phase**in which the bacteria are broken open and destroyed. The nomenclature is descriptive. The bacteria carrying the prophage show no obvious signs of the phage (except immunity to superinfection with the same phage, covered later in Part Four), but when induced (e.g. by stress or UV radiation) they will generate a lytic state, hence they are called lysogens. Induced lysogens make phage from the prophage that was integrated. Phage that always multiply when they infect a cell are called **lytic**.

Excision of a prophage from a lysogen is **not**always precise. Usually only the phage DNA is cut out of the bacterial chromosome, but occassionally some adjacent host DNA is included with the excised phage DNA and encapsidated in the progeny. These **transducing phage** are usually biologically inactive because the piece of the bacterial chromosome replaces part of the phage chromosome; these can be propagated in the presence of helper phage that provide the missing genes when co-infected into the same bacteria. When DNA from the transducing phage is inserted into the newly infected cell, the bacterial genes can **recombine** into the host chromosome, thereby bringing in new alleles or even new genes and genetically altering the infected cell. This process is called **transduction**.

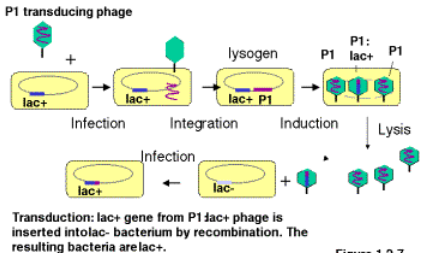


Figure 3.1. Transfer of bacterial genes by transduction: A lac+ transducing phage can convert a lac- strain to lac+ by infection (and subsequent crossing over).

Note that the transducing phage are carrying one or a small number of bacterial genes. This is a way of **isolating the genes**. The bacterial gene in the transducing phage has been separated from the other 4000 bacterial genes (in *E. coli*). By isolating large numbers of the transducing phage, the phage DNA, including the bacterial genes, can be obtained **in large quantities**for biochemical investigation. One can isolate mg or mg quantities of a single DNA molecule, which allows for precise structural determination and detailed investigation.

A **generalized transducing phage**can integrate at many different locations on the bacterial chromosome. Imprecise excision from any of those locations generates a particular transducing phage, carrying a short sections of the bacterial genome adjacent to the

integration site. Thus a generalized transducing phage such as P1 can pick up many different parts of the *E. coli* genome.

A **specialized transducing phage**integrates into only one or very few sites in the host genome. Hence it can carryonly a few specific bacterial genes, e.g., l *lac*(Figure 3.2).

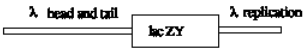


Figure 3.2. An example of a l transducing phage carrying part of the lacoperon.

This process of isolating a particular bacterial gene on a transducing phage is mimicked in **recombinant DNA technology**, in which a gene or genome fragment from any organism is isolated on a recombinant phage or plasmid.

3.2: Overview of Recombinant DNA Technology

Recombinant DNA technology utilizes the power of microbiological selection and screening procedures to allow investigators to isolate a gene that represents as little as 1 part in a million of the genetic material in an organism. The DNA from the organism of interest is divided into small pieces that are then placed into individual cells (usually bacterial). These can then be separated as individual colonies on plates, and they can be screened through rapidly to find the gene of interest. This process is called **molecular cloning**.

Joining DNA in vitro to form recombinant molecules

Restriction endonucleasescut at defined sequences of (usually) 4 or 6 bp. This allows the DNA of interest to be cut at specific locations. The physiological function of restriction endonucleases is to serve as part of system to protect bacteria from invasion by viruses or other organisms. (See Chapter 7)

Table 3.1. List of restriction endonucleases and their cleavage sites. A ' means that the nuclease cuts between these 2 nucleotides to generate a 3' hydroxyl and a 5' phosphate.

Enzyme	Site		Enzyme	Site
*Alu*I	AG'CT		*Not*I	GC'GGCCGC
*Bam*HI	G'GATCC		*Pst*I	CTGCA'G
*Bgl*II	A'GATCT		*Pvu*II	CAG'CTG
*Eco*RI	G'AATTC		*Sal*I	G'TCGAC
*Hae*III	GG'CC		*Sau*3AI	'GATC
*Hha*I	GCG'C		*Sma*I	CCC'GGG
*Hinc*II	GTY'RAC		*Spe*I	A'CTAGT
*Hind*III	A'AGCTT		*Taq*I	T'CGA
*Hin*fI	G'ANTC		*Xba*I	T'CTAGA
*Hpa*II	C'CGG		*Xho*I	C'TCGAG
*Kpn*I	GGTAC'C		*Xma*I	C'CCGGG
*Mbo*I	'GATC			

N = A,G,C or T

R = A or G

Y = C or T

S = G or C

W = A or T

a. Sticky ends

(1) Since the recognition sequences for restriction endonucleases are pseudopalindromes, an off-center cleavage in the recognition site will generate either a 5' overhang or a 3' overhang with self-complementary (or "sticky") ends.

e.g. 5' overhang EcoRI G'AATTC

BamHI G'GATCC

3' overhangPstI CTGCA'G

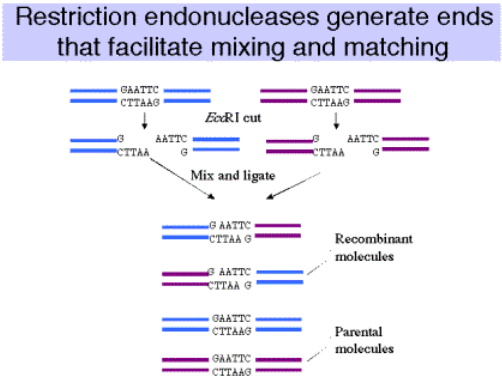
(2) When the ends of the restriction fragments are complementary,

e.g. for EcoRI 5'---G AATTC---3'

3'---CTTAA G---5'

the ends can anneal to each other. **Any two fragments, regardless of their origin (animal, plant, fungal, bacterial) can be joined in vitro to form recombinant molecules** (Figure 3.3)**.**

Figure 3.3.



b. Blunt ends

(1) The restriction endonuclease cleaves in the center of the pseudopalindromic recognition site to generate blunt (or flush) ends.

(2) E.g. HaeIII GG'CC

HincII GTY'RAC

T4 DNA ligase is used to tie together fragments of DNA (Figure 3.4). Note that the annealed "sticky" ends of restriction fragments have **nicks**(usually 4 bp apart). Nicks are breaks in the phosphodiester backbone, but all nucleotides are present. **Gaps**in one strand are missing a string of nucleotides.

T4 DNA ligase uses ATP as source of adenylyl group attached to 5' end of the nick, which is a good leaving group after attack by the 3' OH. (See Chapter 5 on Replication).

At high concentration of DNA ends and of ligase, the enzyme can also ligate together blunt-ended DNA fragments. Thus any two blunt-ended fragments can be ligated together. Note: Any fragment with a 5' overhang can be readily converted to a blunt-ended molecule by fill-in synthesis catalyzed by a DNA polymerase (often the Klenow fragment of DNA polymerase I). Then it can be ligated to another blunt-ended fragment.

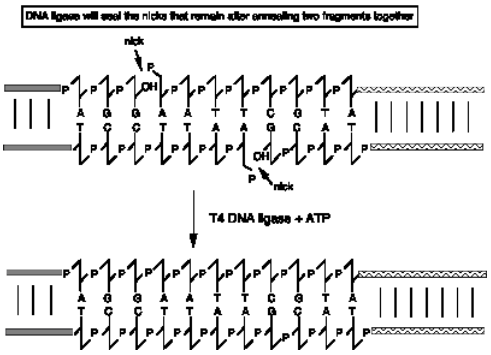


Figure 3.4

Linkers are short duplex oligonucleotides that contain a restriction endonuclease cleavage site. They can be ligated onto any blunt-ended molecule, thereby generating a new restriction cleavage site on the ends of the molecule. Ligation of a linker on a restriction fragment followed by cleavage with the restriction endonuclease is one of several ways to generate an end that is easy to ligate to another DNA fragment.

Annealing of **homopolymer tails** are another way to joint two different DNA molecules.

The enzyme **terminal deoxynucleotidyl transferase**will catalyze the addition of a string of nucleotides to the 3' end of a DNA fragment. Thus by incubating each DNA fragment with the appropriate dNTP and terminal deoxynucleotidyl transferase, one can add complementary homopolymers to the ends of the DNAs that one wants to combine. E.g., one can add a string of G's to the 3'

ends of one fragment and a string of C's to the 3' ends of the other fragment. Now the two fragments will join together via the homopolymer tails.

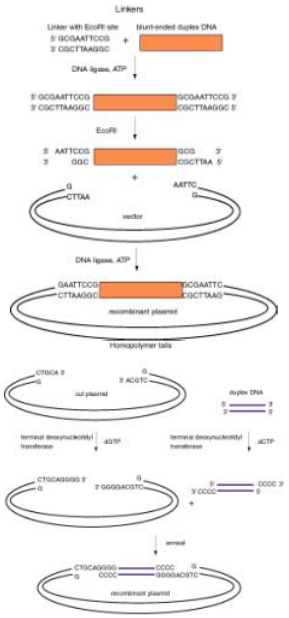


Figure 3.5. Use of linkers (left) and homopolymer tails (right) to make recombinant DNA molecules.

3.3: Introduction of recombinant DNA into cell and replication: Vectors

Vectors used to move DNA between species, or from the lab bench into a living cell, must meet three requirements (Figure 3.3.1):

1. They must be **autonomously replicating** DNA molecules in the host cell. The most common vectors are designed for replicating in bacteria or yeast, but there are vectors for plants, animals and other species.
2. They must contain a **selectable marker** so cells containing the recombinant DNA can be distinguished from those that do not. An example is drug resistance in bacteria.
3. They must have an **insertion site** to accommodate foreign DNA. Usually a unique restriction cleavage site in a nonessential region of the vector DNA. Later generation vectors have a set of about 15 or more unique restriction cleavage sites.

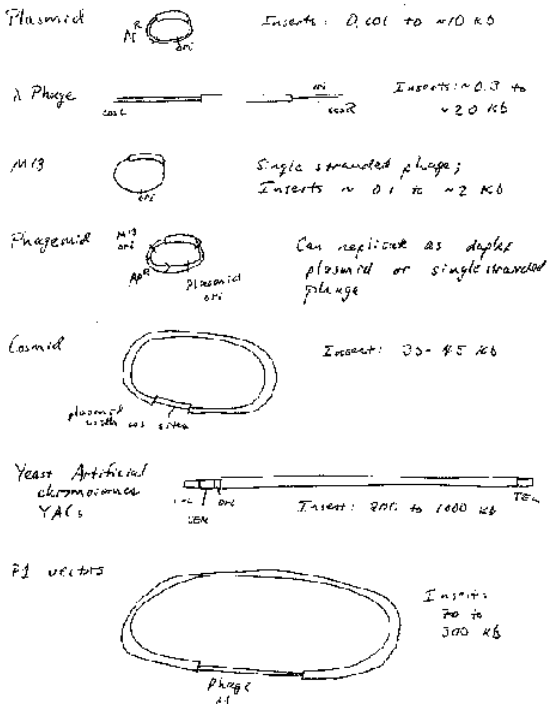


Figure 3.3.1: Summary of vectors for molecular cloning

Plasmid Vectors

Plasmids are autonomously replicating circular DNA molecules found in bacteria. They have their own origin of replication, and they replicate independently of the origins on the "host" chromosome. Replication is usually dependent on host functions, such as DNA polymerases, but regulation of plasmid replication is distinct from that of the host chromosome. Plamsids, such as the *sex-factor F*, can be very large (94 kb), but others can be small (2-4 kb). Plasmids do not encode an essential function to the bacterium, which distinguishes them from chromosomes. Plasmids can be present in a single copy, such as F, or in multiple copies, like those used as most cloning vectors, such as pBR322, pUC, and pBluescript.

In nature, plasmids provide carry some useful function, such as transfer (F), or antibiotic resistance. This is what keeps the plasmids in a population. In the absence of selection, plasmids are lost from bacteria. The antibiotic resistance genes on plasmids are often carried within, or are derived from, transposons, a types of transposable element. These are DNA segments that are capable of "jumping" or moving to new locations (Chapter 9).

A plasmid that was widely used in many recombinant DNA projects is pBR322 (Figure 3.3.2). It replicates from an origin derived from a colicin-resistance plasmid (ColE1). This origin allows a fairly high copy number, about 100 copies of the plasmid per cell. Plasmid pBR322 carries two antibiotic resistance genes, each derived from different transposons. These transposons were initially found in R-factors, which are larger plasmids that confer antibiotic resistance.

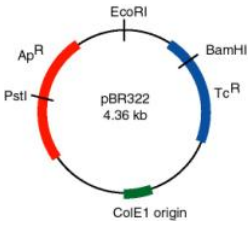


Figure 3.3.2: Features of plasmid pBR322. The gene conferring resistance to ampicillin (ApR) can be interrupted by insertion of a DNA fragment into the PstI site, and the gene conferring resistance to tetracycline (TcR) can be interrupted by insertion of a DNA fragment into the BamHI site. Replication is controlled by the ColE1 origin.

Use of the TcR and ApR genes allows for easy screening for recombinants carrying inserts of foreign DNA. For instance, insertion of a restriction fragment in the *Bam*HI site of the TcR gene inactivates that gene. One can still select for ApR colonies, and then screen to see which ones have lost TcR .

Exercise 3.3.1

What effects on drug resistance are seen when you use the *Eco*RI or *Pst*I sites in pBR322 for inserting foreign DNA?

A generation of vectors developed after pBR322 are designed for even more efficient **screening for recombinant plasmids**, i.e. those that have foreign DNA inserted. The **pUC** plasmids (named for plasmid universal cloning) and plasmids derived from them use a rapid screen for inactivation of the b-galactosidase gene to identify recombinants (Figure 3.3.3).

One can screen for production of functional **b-galactosidase** in a cell by using the chromogenic substrate **X-gal** (a halogenated indoyl b-galactoside). When cleaved by b-galactosidase, the halogenated indoyl compound is liberated and forms a blue precipitate. The pUC vector has the b-galactosidase gene {actually only part of it, but enough to form a functional enzyme with the rest of the gene that is encoded either on the E. coli chromosome or an F' factor}. When introduced into *E. coli*, the colonies are **blue** on plates containing X-gal.

The **multiple cloning sites** (unique restriction sites) are in the b-galactosidase gene (lacZ). When a restriction fragment is introduced into one or more of these sites, the b-galactosidase activity is lost by this insertional mutation. Thus cells containing recombinant plasmids form **white**(not blue) colonies on plates containing X-gal.

The replication origin is a modified ColE1 origin of replication that has been mutated to eliminate a negative control region. Hence the **copy number is very high**(several hundred or more plasmid molecules per cell), and one obtains an very high yield of plasmid DNA from cultures of transformed bacteria. The plasmid has ApR as a selectable marker.

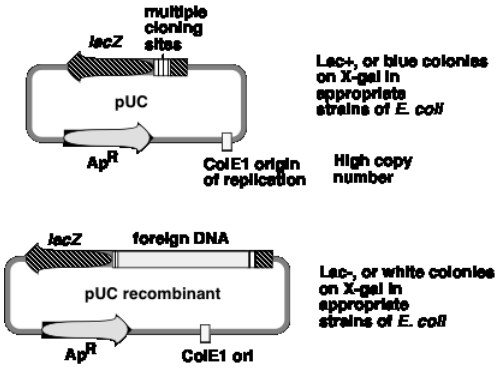


Figure 3.3.3: pUC-type vectors

3.4: Introducting Recombinant DNA into Host Cells

Transformation in *E. coli*

E. coli does not have a natural system for taking up DNA, but when treated with $CaCl_2$, the cells will take up the added DNA (Figure 3.4.1). The recombinant vectors will give a new phenotype to the cells (usually drug resistance), so this process can be considered **DNA-mediated transformation**. An average efficiency is about 10^6 transformants per mg of DNA, although some more elaborate transformation cocktails procedures can give up to about 10^8 transformants per mg of DNA.

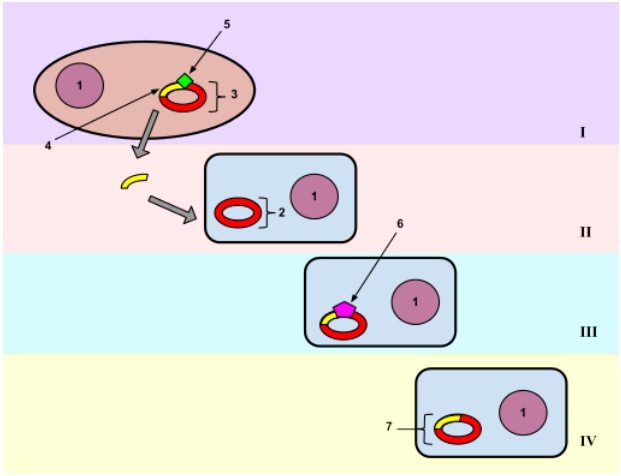


Figure 3.4.1: DNA-mediated transformation of E. coli. A gene from bacterial cell 1 is moved from bacterial cell 1 to bacterial cell 2. This process of bacterial cell 2 taking up new genetic material is called transformation. Step I: The DNA of a bacterial cell is located in the cytoplasm (1), but also in the plasmid, an independent, circular loop of DNA. The gene to be transferred (4) is located on the plasmid of cell 1 (3), but not on the plasmid of bacterial cell 2 (2). In order to remove the gene from the plasmid of bacterial cell 1, a restriction enzyme (5) is used. The restriction enzyme binds to a specific site on the DNA and "cuts" it, releasing the satisfactory gene. Genes are naturally removed and released into the environment usually after a cell dies and disintegrates. Step II: Bacterial cell 2 takes up the gene. This integration of genetic material from the environment is an evolutionary tool and is common in bacterial cells. Step III: The enzyme DNA ligase (6) adds the gene to the plasmid of bacterial cell 2 by forming chemical bonds between the two segments which join them together. Step IV: The plasmid of bacterial cell 2 now contains the gene from bacterial cell 1 (7). The gene has been transferred from one bacterial cell to another, and transformation is complete. (CC BY_SA 3.0; Sprovenzano15)

Usually one will transform with a mixture of recombinant vector molecules, most of which carry a different restriction fragment. Each transformed E. coli cell will pick up only **one** plasmid molecule, so the complex mixture of plasmids in the ligation mix has been separated into a population of transformed bacteria (Figre 3.4.1). The bacterial cells are then plated at a sufficiently low density that individual colonies can be identified. Each colony (or transformant) carries a single plasmid, so as one screens the colonies, one is actually screening through individual DNA molecules. A colony is a visible group of bacterial cells on a plate, all of which are derived from a single bacterial cell. A group of identical cells derived from a single cell is called a **clone**. Since each clone carries a single type of recombinant DNA molecule, the process is called **molecular cloning**.

Phage Vectors

Phase vectors are a more efficient introduction of DNA into bacteria. Phage vectors such as those derived from bacteriophage l can carry **larger inserts** and can be **introduced into bacteria more efficiently**. l phage has a duplex DNA genome of about 50 kb. The internal 20 kb can be replaced with foreign DNA and still retain the lytic functions. Hence restriction fragments up to 20 kb can replace the l sequences, allowing larger genomic DNA fragments to be cloned (Figure 3.4.2).

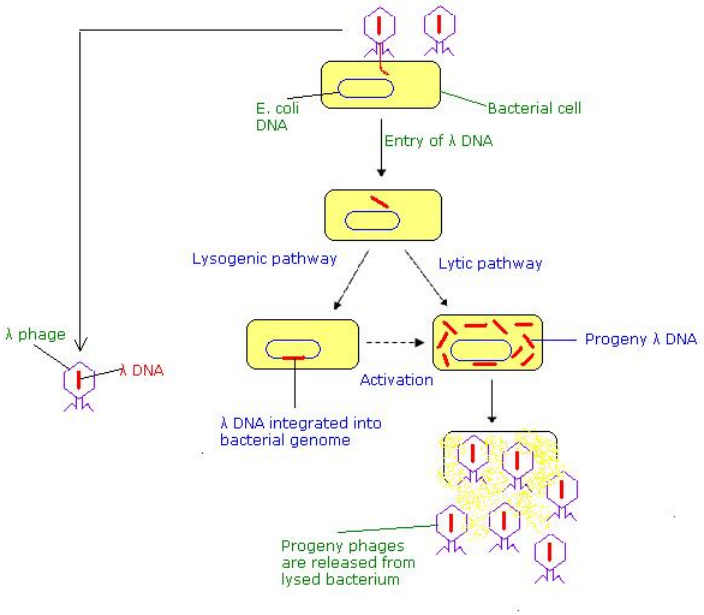


Figure 3.4.2: Lambda vectors for cloning. (public Domain; Tinastella).

Recombinant bacteriophage can be introduction into *E. coli* by **infection**. DNA that has the cohesive ends of l can be packaged in vitro into infective phage particles. Being in a viral particle brings the efficiency of infection reliably over 108 plaque forming units per mg of recombinant DNA.

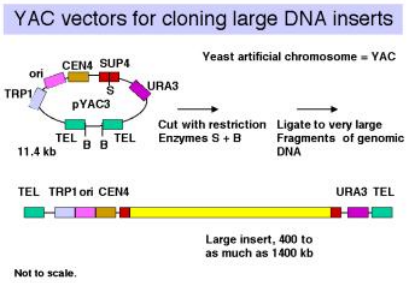
Some other bacteriphage vectors for cloning are derived from the virus M13. One can obtain **single stranded DNA** from M13 vectors and recombinants. M13 is a virus with a genome of single stranded DNA. It has a nonessential region into which foreign genes can be inserted. It has been modified to carry a gene for b-galactosidase as a way to screen for recombinants. Introduction of recombinant M13 DNA into E. coli will lead to an infection of the host, and the progeny viral particles will contain single-stranded DNA. The replicative form is duplex, allowing one to cleave with restriction enzymes and insert foreign DNA.

Some vectors are hybrids between plasmids and single-strand phage; these are called **phagemids**. One example is pBluescript. Phagemids are plasmids (with the modified, high-copy number ColE1 origin) that also have an M13 origin of replication. Infection of transformed bacteria (containing the phagemid) with a helper virus (e.g. derived from M13) will cause the M13 origin to be activated, and progeny viruses carrying single-stranded copies of the phagemid can be obtained. Hence one can easily obtain either double- or single-stranded forms of thes plasmids. {The "blue" comes from the blue-white screening for recombinants that can be done when the multiple cloning sites are in the b-galactosidase gene. The "script" refers to the ability to make RNA copies of either strand in vitro with phage RNA polymerases.}

Vectors Designed to Carry Larger Inserts

Fragments even larger than those carried in l vectors are useful for studies of longer segments of chromosomes or whole genomes. Several vectors have been designed for cloning these very large fragments, 50 to 400 kb.

- **Cosmids** are plasmids that have the cohesive ends of l phage. They can be packaged in vitro into infective phage particles to give a more efficient delivery of the DNA into the cells. They can carry about 35 to 45 kb inserts.
- **Yeast artificial chromosomes** (**YACs**) are yeast vectors with centromeres and telomeres. They can carry about 200 kb or larger fragments (in principle up to 1000 kb = 1 Mb). Thus very large fragments of DNA can be cloned in yeast (Figure 3.4.3). In practice, chimeric clones with fragments from different regions of the genome are obtained fairly often, and some of the inserts are unstable.



,

Figure 3.4.3

Vectors derived from bacteriophage **P1** can carry fragments of about 100 kb. Fragments in a similar size range are also cloned into **bacterial artificial chromosomes** (**BACs**), which are derived from the F-factor (Figure 3.4.4). These have a lower copy number (like F) but they are stable and relatively easy to work with in the laboratory. BACs have become one of the most frequently used vectors for large inserts in genome projects.

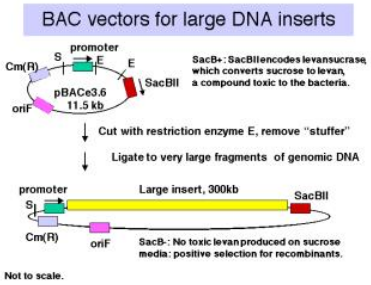


Figure 3.4.4

Shuttle vectors for testing functions of isolated genes

Shuttle vectors can replicate in two different organisms, e.g. bacteria and yeast, or mammalian cells and bacteria. They have the appropriate origins of replication. Hence one can, e.g. clone a gene in bacteria, maybe modify it or mutate it in bacteria, and test its function by introducing it into yeast or animal cells.

3.5: Polymerase Chain Reaction (PCR)

The **polymerase chain reaction (PCR)** is now one of the most commonly used assays for obtaining a particular segment of DNA or RNA. It is rapid and extremely sensitive. By amplifying a designated segment of DNA, it provides a means to isolate that particular DNA segment or gene. This method requires knowledge of the nucleotide sequence at the ends of the region that you wish to amplify. Once that is known, one can make large quantities of that region starting with miniscule amounts of material, such as the DNA within a single human hair. With the availability of almost complete or complete sequences of genomes from many species, the range of genes to which it can be applied is enormous. The applications of PCR are numerous, from diagnostics to forensics to isolation of genes to studies of their expression.

The power of PCR lies in the exponential increase in amount of DNA that results from repeated cycles of DNA synthesis from primers that flank a given region, one primer designed to direct synthesis complementary to the top strand, the other designed to direct synthesis complementary to the bottom strand (Figure 3.5.1). When this is done repeatedly, there is roughly a 2-fold increase in the amount of synthesized DNA in each cycle. Thus it is possible to generate a million-fold increase in the amount of DNA from the amplified region with a sufficient number of cycles. This exponential increase in abundance is similar to a chemical chain reaction, hence it is called the polymerase chain reaction.

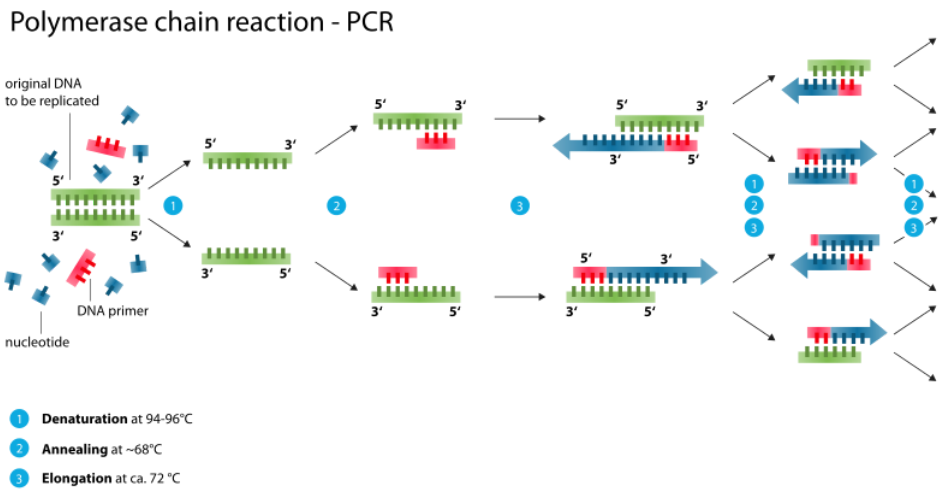


Figure 3.5.1: Polymerase Chain Reaction (PCR). (CC BY-SA 3.0; Enzoklop)

The events in the polymerase chain reaction are examined in more detail in Figure 3.5.2. The several panels show what happens in each cycle. Each cycle consists of a denaturation step at a temperature higher than the melting temperature of the duplex DNA (e.g. 95 ^{o}C), then an annealing step at a temperature below the melting temperature for the primer-template (e.g. 55 ^{o}C), followed by extension of the primer by DNA polymerase using dNTPs provided in the reaction. This is done at the temperature optimum for the DNA polymerase (e.g. 70 ^{o}C for a thermostable polymerase). **Thermocylers** are commercially available for carrying out many cycles quickly and reliably (Figure 3.5.3).

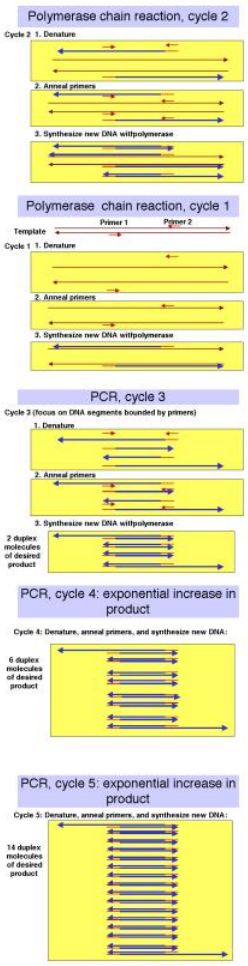


Figure 3.5.2: Steps in the polymerase chain reaction.

The template supplied for the reaction is the only one available in the first cycle, and it is still a major template in the second cycle. At the end of the second cycle, a product is made whose ends are defined by primers. This is the desired product, and it serves as the major template for the remaining cycles. The initial template is still present and can be used, but it does not undergo the exponential expansion observed for the desire product.

If *n*is the number of cycles, the amount of desired product is approximately 2^{n-1} –2 times the amount of input DNA (between the primers). Thus in 21 cycles, one can achieve a million-fold increase in the amount of that DNA (assuming all cycles are completely efficient). A sample with 0.1 pg of the segment of DNA between the primers can be amplified to 0.1 mg in 21 cycles, in theory. In practice, roughly 25-35 cycles are done in many PCR assays.

Figure 3.5.2: A thermal cycler for PCR. (CC BY-SA 3.0; Rror).

The ease if doing PCR was greatly increased by the discovery of DNA polymerases that were stable at high temperatures. These have been isolated from bacteria that grow in hot springs, such as those found in Yellowstone National Park, such as *Thermus*

aquaticus. The **Taq polymerase** from this bacterium will retain activity even at the high temperatures needed for melting the templates, and it is active at a temperature between the melting and annealing temperature. This particular polymerase is rather error-prone, and other thermostable polymerases have been discovered that are more accurate.

3.6: cDNA

cDNA clones are copies of mRNAs

Construction of cDNA clones involves the synthesis of complementary DNA from mRNA and then inserting a duplex copy of that into a cloning vector, followed by transformation of bacteria (Figure 3.6.1).

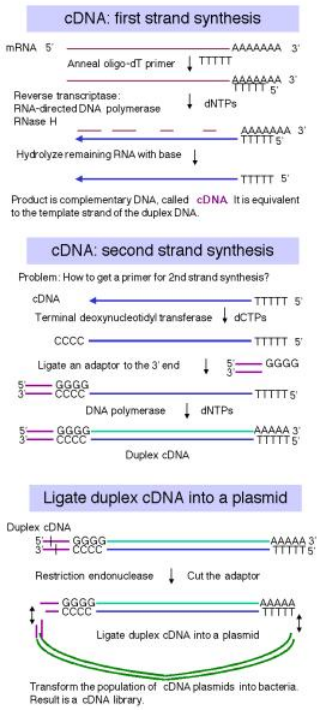


Figure 3.6.1: Making cDNA clones

a. **First strand synthesis**: First, one anneals an oligo dT primer onto the 3' polyA tail of a population of mRNAs. Then reverse transcriptase will begin DNA synthesis at the primer, using dNTPs supplied in the reaction, and copy the mRNA into **complementary DNA**, abbreviated **cDNA**. The mRNA is degraded by the RNase H activity associated with reverse transcriptase and by subsequent treatment with alkali.

b. **Second strand synthesis**: For the primer to make the second strand of DNA (equivalent in sequence to the original mRNA), one can utilize a transient hairpin at the end of the cDNA. (The basis for its formation is not certain.) In other schemes, one generates a primer binding site and uses a primer directed to that site; one way to do this is by homopolymer tailing of the cDNA followed by use of a complementary primer. Random primers can also be used for second strand synthesis; although this precludes the generation of a full-length cDNA (i.e. a copy of the entire mRNA). However, it is rare to generate duplex copies of the entire mRNA by any means.

DNA polymerase (e.g. Klenow polymerase) is used to synthesize the second strand, complementary to the cDNA. The product is **duplex cDNA.**

If the hairpin was used to prime second strand synthesis, it must be opened by a single-strand specific nuclease such as S1.

c. Insertion of the duplex cDNA into a cloning vector:

One method is to use terminal deoxynucleotidyl transferase to add a homopolymer such as poly-dC to the ends of the duplex cDNA and a complementary homopolymer such as poly-dG to the vector.

An alternative approach is to use linkers; these can be employed such that a linker carrying a cleavage site for one restriction endonuclease is on the 5' end of the duplex cDNA and a linker carrying a cleavage site for a different restriction endonuclease is on the 3' end. (In this context, 5’ and 3’ refer to the nontemplate, or "top" strand.) This allows "forced" cloning into the vector, and one has initial information about orientation, based on proximity to one cleavage site or the other.

The cDNA and vector are joined at the ends, using DNA ligase, to form recombinant cDNA plasmids (or phage).

d. The ligated cDNA plasmids are then transformed into E. coli. The resulting set of transformants is a library of cDNA clones.

Screening methods for cDNA clones

a. Brute force examination of individual cDNA plasmids.

If the mRNA is highly abundant in a given tissue, then many of the cDNA clones will be copies of that mRNA. One can examine DNA from individual clones and test for characteristic restriction cleavage patterns or a particular sequence. This was a common approach for screening cDNAs in the early days of recombinant DNA technology.

Starting in the mid-1990's, cooperative efforts from corporations (such as Merck) and publicly funded genome centers (such as at Washington University) have generated the sequence of individual clones from large cDNA libraries from many tissues from human, mouse, and rat. Other consortia have sequenced cDNA libraries from other species. Each sequence is called an "expressed sequence tag" or **EST**. These are now a major source of partially or fully characterized cDNA clones. Hundreds of thousands of ESTs are available, and contain at part of the DNA sequence from many, if not most, human genes. The web site for NCBI (http://www.ncbi.nlm.nih.gov) is an excellent resource for examining the ESTs.

b. Hybridization with a gene-specific probe.

If the sequence of the desired cDNA is known, or if the sequence from homologs from related species is known, one can use synthetic oligonucleotides (or other source of the diagnostic sequence) as a radiolabeled hybridization probe to identify the cDNA of interest.

If the amino acid sequence has been determined for all or even just parts of the protein product of the gene of interest, then one can chemically synthesize oligonucleotides based on the genetic code for those amino acids. The oligonucleotides need to be at least 18 nucleotides or longer (so that they will anneal to specific sites in the genome), and because the genetic code is degenerate (more than one codon per amino acid; discussed in Part Two), they have to be degenerate as well. The oligonucleotides can be used directly as hybridization probes, although it is becoming more common to amplify the region between two oligonucleotides using the polymerase chain reaction, and to use that amplification product as a labeled probe.

The process of hybridization screening is illustrated schematically in Figure 3.16. The colonies of bacteria, each with a single cDNA plasmid, are transferred to a solid substrate (such as a nylon or nitrocellulose membrane), lysed. and the released DNA immobilized onto the membrane. Hybridization of this membrane (with the DNA attached) to a specfic probe allows one to screen through thousands of colonies in a single experiment.

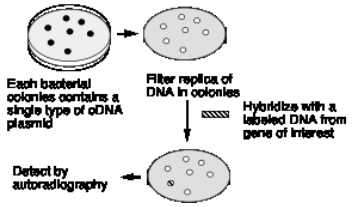


Figure 3.6.2: Hybridization Screening

c. Express the cDNA, i.e. make the protein product encoded by the mRNA, and screen for that protein product (Figure 3.6.3). This is often in bacteria by constructing the clones in a vector that has an active E. coli promoter (for transcription) and efficient translation signals upstream from the site at which the cDNAs were inserted. The transformed bacterial cells will express the encoded protein, and one tries to identify it. One can also screen for expression in yeast, plant or mammalian cells. The expression vector has to contain gene-regulatory signals (such as promoters and enhancers, see Part Three) that allow expression of the desired gene in the appropriate cell.

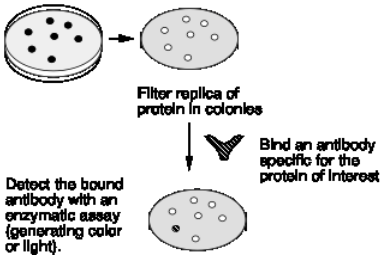


Figure 3.6.3: Screening for an Expressed Gene Product

1. One can use specific *antisera* to detect the desired colony expressing the gene of interest.
2. One can use a labeled ligand that will bind to the expressed cDNA on the cell surface. For example, cDNAs for receptors can be expressed in an appropriate cell (usualy mammalian cells in culture) and identified by newly-acquired ability to bind a labeled hormone (such as growth hormone or erythropoietin)
3. by *complementation* of a known mutation in the host. E.g. a cDNA for the human homolog to yeast p34*cdc2* was isolated by its ability to complement a yeast mutant that had lost the function of this key regulator of progress through the cell cycle.
4. Expression cloning can be done in mammalian cells, as long as one can screen or select for a new function generated by the expression. Use of this method to isolate the receptor for the glycoprotein hormone erythropoietin is illustrated in Figure 3.6.4.

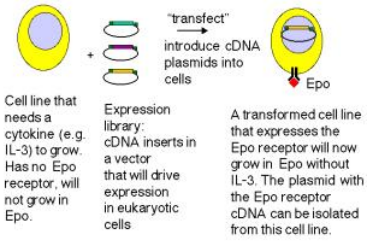


Figure 3.6.4: Expression screening in eukaryotic cells.

d. Differential analysis:

Often one is interested in finding all the genes (or their mRNAs) that are expressed uniquely in some differentiated or induced state of cells. Two classic examples are (i) identifying the genes whose products regulate the determination process that causes a multipotential mouse cell line (like 10T1/2 cells) to differentiate into muscle cells, and (ii) ,using the fact that the T-cell receptor is expressed only in T-lymphocytes, but not in their sister lineage B-lymphocytes, to help isolate cDNA clones for that mRNA. Both of these projects used subtractive hybridization to highly enrich for the cDNA clones of interest.

In this technique, the cDNA from the differentiating or induced cell of interest is hybridized to mRNA from a related cell line, but which has not undergone the key differentiation step. This allows one to remove mRNA-cDNA duplexes that contain the cDNAs for all the genes expressed in common between the two types of cells. The resulting single-stranded are enriched for the cDNAs that are involved in the process under study.

The subtractive hybridization scheme used in isolation of the muscle determination gene *MyoD* is illustrated in Figure 3.6.5.

A conceptually equivalent strategy, using PCR (see next section) rather than cDNA cloning, is differential display of PCR products from cells that differ by some process (e.g. differentiation, induction, growth arrest versus stimulation, etc.). In this technique, one uses several sets of PCR primers annealed to cDNA to mRNA from the two types of cells that are being compared. The sets of primers are empirically designed to allow many regions of cDNA to be amplified. The amplification products are resolved (or displayed) on polyacrylamide gels, and the products specific to the cell type of interest are isolated and used to screen through cDNA libraries. This technique is also called representational difference analysis.

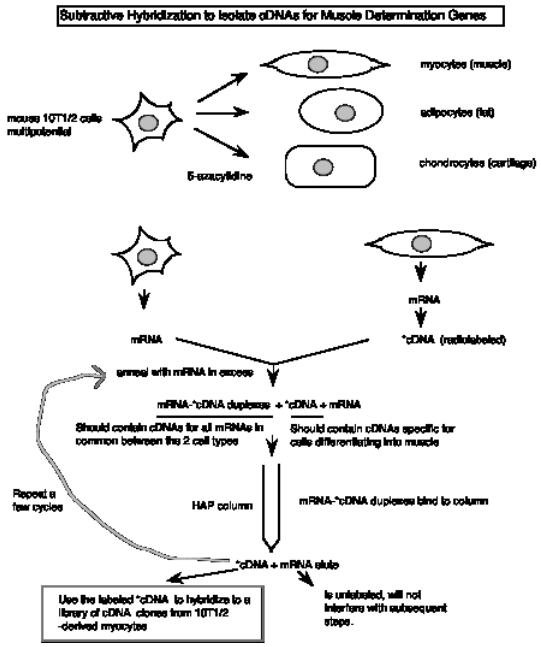


Figure 3.6.5: Differential screening to find cDNAs of mRNAs expressed only in certain cell-types.

The advent of sequencing all or a very large number of genes from various organisms (e.g. *E. coli*, yeast, *Drosophila*, humans) has allowed the development of high-density microchip arrays of DNA from each gene. One can hybridize RNA from cells or tissues of interest, isolated under various metabolic conditions, to identify all (known) genes expressed. Even more useful are assays for genes whose expression *changes* during a shift in cell metabolism (cell cycle, heat shock, hormonal induction, etc.) or as a result of mutation of some other gene (e.g. a gene encoding a transcription factor of interest). This powerful new technology is being used more and more to examine global effects on gene expression.

For a description (and movie) of the Affymetrix GeneChip, go to **http://www.affymetrix.com/technology/index.html**

3.7: Genomic DNA clones

Clones of *genomic DNA*, containing individual fragments of chromosomal DNA, are needed for many purposes. Some examples include:

- to obtain detailed structures of genes,
- to identify regulatory regions, i.e. DNA sequences needed for correct expression of the gene,
- to map and analyze alterations to the genome, e.g. the isolate genes that when mutated cause a hereditary disease,
- to direct alterations in the genome, e.g. by homologous recombination to replace a wild-type allele with a mutant one (to test function of the gene in mouse) or vice versa(to cure a hereditary disease, perhaps eventually in humans).

Construction of libraries of genomic DNA fragments in cloning vectors

Genomic DNA is digested with restriction enzymes (Figure 3.7.1). The more frequently an enzyme cuts (the shorter the recognition sequence), the smaller the average size of DNA fragments. Some enzymes cut very infrequently, such as NotI (8 bp recognition sequence) and can be used to generate very large fragments. Alternatively, one can do a partial digest (not all sites are cleaved) with a particular enzyme and isolate the products that are in the desired size range (e.g. 20 kb). A particularly clever way to do this is to digest partially with Sau3AI or MboI (both cut at 'GATC) and ligate these fragments into vector cut with BamHI (cuts at G'GATCC) - i.e. they have the same sequence in the overhang (or sticky end). In this process one uses vectors that can accomodate large DNA fragments, such as l phage vectors, cosmids, YACs or P1 vectors.

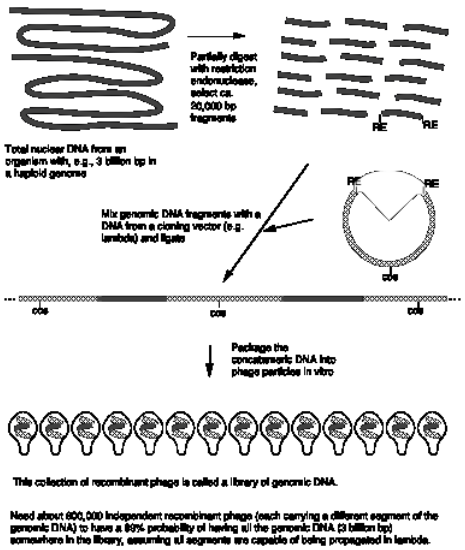


Figure 3.7.1: Construction of a library of genomic DNA

Screening methods for genomic DNA clones

One method is to use **complementation** of a mutation in the host to select or screen for the desired gene. This works just like the situation for cDNA clones described above, and it requires that the cloned fragments be expressed in the host cell. Far more common is to screen by **hybridization** with gene-specific probes (Figure 3.7.2). Frequently the cDNA clone is found first, and the genomic clone then isolated by hybridization screening (using the cDNA clone as a probe) against a library of genomic DNA fragments.

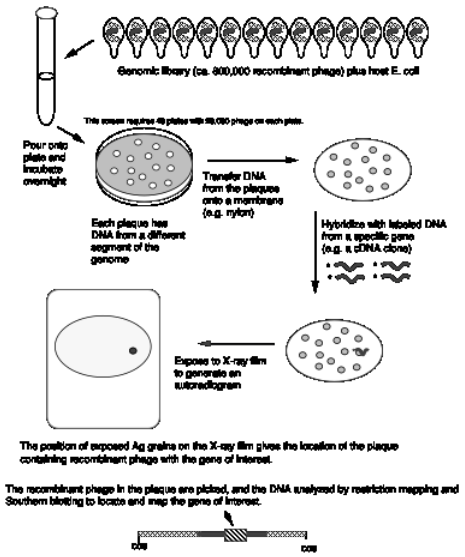


Figure 3.7.2: Screening a library of genomic DNA

3.8: Eukaryotic Gene Structure

Much can be learned about any gene after it has been isolated by recombinant DNA techniques. The structure of coding and noncoding regions, the DNA sequence, and more can be deduced. This is true for bacterial and viral genes, as well as eukaryotic cellular genes. The next sections of this chapter will focus on analysis of eukaryotic genes, showing the power of examining purified copies of genes.

Split genes and introns

Precursors to mRNA longer than mRNA

Initial indications of a complex structure to eukaryotic genes came from analysis of nuclear RNAs during the 1970's. The precursors to messenger RNA, or pre-mRNAs, were found to be surprisingly **long**, considerably larger than the average mRNA size (Figure 3.8.1).

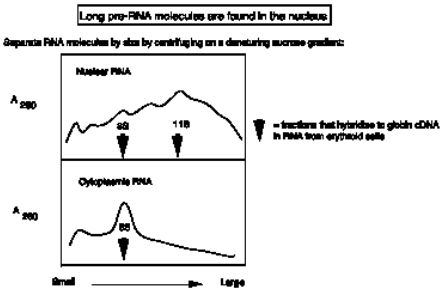


Figure 3.8.1

Denaturing sucrose gradients (with high concentration of formamide, e.g. >50%) separate RNAs on the basis of size. Analysis of nuclear RNA showed that the average size was much larger than the average size of cytoplasmic RNA. Labeled RNA could be "chased" from the nucleus to the cytoplasm - i.e. nuclear RNA was a precursor to mRNA and other cytoplasmic RNAs. Was the extra RNA at the ends? or in the middle of the pre-mRNA? More precisely, one could examine specific RNAs by hybridizing fractions from the denaturing sucrose gradients to labeled copies of, e.g. globin mRNA. The hybridizing RNA from the nucleus was about 11S (as well as mature 8S message), whereas cytoplasmic RNA of about 8S hybridized. Thus the nuclear RNA encoding globin is larger than the cytoplasmic mRNA.

Visualization of mRNA-DNA heteroduplexes revealed extra sequences internal to the mRNA-coding segments

R-loops are hybrids between RNA and DNA that can be visualized in the EM, under conditions where DNA-RNA duplexes are favored over DNA-DNA duplexes (Figure 3.8.2). For a simple gene structure, one sees a continuous RNA-DNA duplex (smooth, slowly curving) and a displaced single strand of DNA (thinner, many more turns and curves – single stranded DNA is not a rigid as double stranded nucleic acid, either duplex DNA or RNA-DNA).

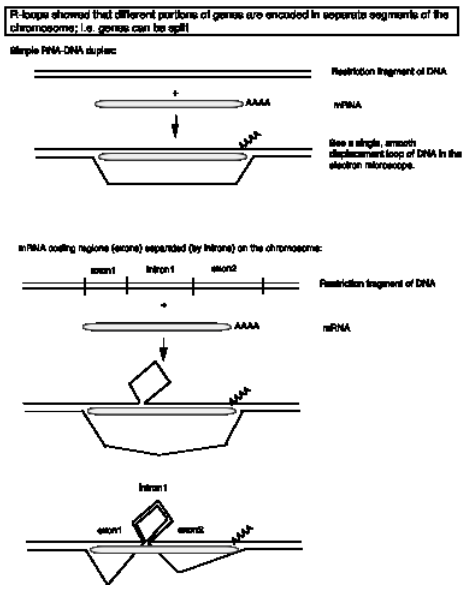


Figure 3.8.2

EM pictures of duplexes between purified adenovirus mRNAs and the genomic DNA showed extensions at both the 3' (poly A) and 5' ends, which are encoded elsewhere on the genome. All late mRNAs have the same sequence at the 5' end; this is dervied from from the tripartite leader. R-loops between late mRNAs and adenovirus DNA fragments including the major late promoter showed duplexes with the leader segments, separated by loops of duplex DNA (Figure 3.23, bottom panel). The RNA-DNA hybrids identify regions of DNA that encode RNA. The surprising result is that RNA-coding portions of a gene are separated by loops of duplex DNA in the R-loop analysis. Examples of R-loops in genes with introns are shown in Figure 3.8.3.

These data showed that the adenovirus **RNAs are encoded in different segments of the viral genome; i.e. the genes are split**. The portion of a gene that encodes mRNA was termed an **exon**. The part of gene does not code for sequences in the mature mRNA is called an **intron**. These observations led to the Nobel Prize for Phil Sharp and Rich Roberts. Louise Chow and Sue Berget were also key players in the discovery of introns.

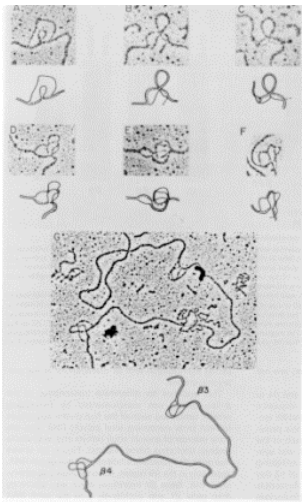


Figure 3.8.3: R-loops between clones of rabbit beta-like globin genes (now called HBEand HBG) and mRNA from rabbit embryonic erythroid cells. A photograph from the electron microscope is shown at the the top of each panel, and an interpretive drawing is included below it. The displaced nontemplate strand of DNA forms partial or complete duplexes with the template strand in the large intron. A small intron is also visible in panel C. Panel G shows the two genes together on one large clone.

Interruptions in cellular genes were discovered subsequently, in the late 1970's, in globin genes, immunoglobulin genes and others. We now realize that mostgenes in complex eukaryotes are split by multiple introns.

Exons are more conserved than introns (in most cases), since alterations in protein-coding regions that alter or decrease function are selected against, whereas many sequences in introns can be altered without affecting the function of the gene product. Important

sequences in introns (such as splice junctions, the branch point, and occassionally enhancers) are covered in some detail in Part Three.

Differences in restiction maps between cDNA and genomic clones reveal introns

Restriction maps based on copies of the mRNA (cDNA) were different from those in genomic DNA - the genes were cleaved by some restriction endonucleases that the cDNAs were not, and some restriction sites were further apart in the genomic DNA. These observations were explained by the presence of intervening sequences or introns (Figure 3.8.4).

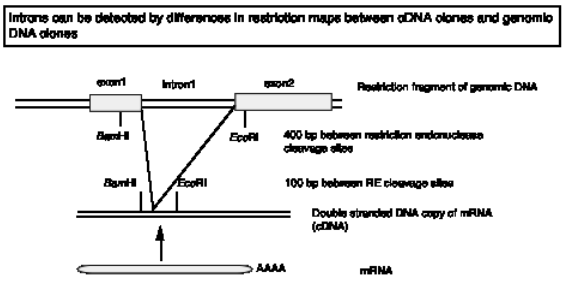


Figure 3.8.4

The experimental procedures to do this involve making a **restriction map** of the clones of genomic DNA, and then **identifying the regions that encode mRNA by hybridization of labeled cDNA probes** to the restriction digests. Cloned genomic DNA digested with appropriate restriction endonucleases, separated by size on an agarose gel, and then transferred onto a nylon or nitrocellulose solid support. This **Southern blot** is then hybridized with a labeled probe specific to the cDNA (composed only of exons). The pattern of labeled fragments on the resulting autoradiogram shows the fragments that contain exons. Alignment of these with the restriction map of the gene gives an approximation of the position of the exons.

The blot-hybridization approach can be combined with a PCR (polymerase chain reaction) analysis for higher resolution. Primers are synthesized that will anneal to adjacent exons. The difference in size of the PCR amplification product between genomic DNA and cDNA is the size of the intron. The PCR product can be cloned and sequenced for more detailed information, e.g. to precisely define the exon/intron junctions.

Subsequently, the nucleotide sequence of exonic regions and preferably the entire gene is determined. The presence of introns were confirmed and their locations defined precisely in DNA sequences of isolated clones of the genes.

Types of Exons

Eukaryotic genes are a combination of introns and exons. However, not all exons do the same thing (Figure 3.8.5). In particular, the protein-coding regions or genes are a subset of the sequences in exons. Exons include both the untranslated regions and the protein-coding, translated regions. Introns are the segments of genes that are present in the primary transcript (or precursor RNA) but are removed by splicing in the production of mature RNA. Methods used to detect coding regions will not find all exons.

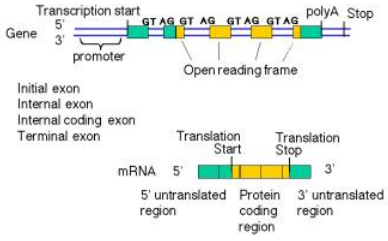


Figure 3.8.5: Types of exons

3.9: Introns and Exons

Multiple, large introns can make some eukaryotic genes very large. Eukaryotic genes can be split into many (>60), sometimes very small exons (e.g. <60 bp, coding for <20 amino acids), separated by very large introns (as large as >100kb), resulting in some enormous genes (>500 kb). E.g. the *DMD*gene (which when mutated can cause Duchenne's muscular dystrophy) is almost 1 Mb, about 1/4 the size of the *E. coli* chromosome! The average size of genes from more complex organisms is considerably larger than those of simpler ones, but the avg. size of mRNA is about the same, reflecting the presence of more and larger introns in the more complex organisms. tRNA and rRNA genes also contain introns

Finding exons in long genomic sequences using computer programs

Far more exons and introns have been discovered (or more accurately, predicted) throught the analysis of genomic DNA sequences than could ever be discovered by direct experimentation. The different types of exons, the enormous length of introns, and other factors have complicated the task of finding reliable diagnostic signatures for exons in genomic sequences. However, considerable progress has been made and continues in current research. Some of the commonly used approaches are summarized in Figure 3.27.

Figure 3.27. Introns in the b-globin gene can be reliably identified computat

Introns are removed by splicing RNA precursors

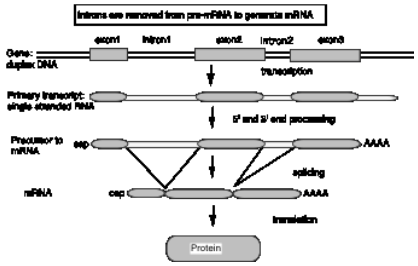


Figure 3.28. Introns are removed from pre-mRNA to generate mRNA.

Alternative splicing generates more than one polypeptide from the same gene

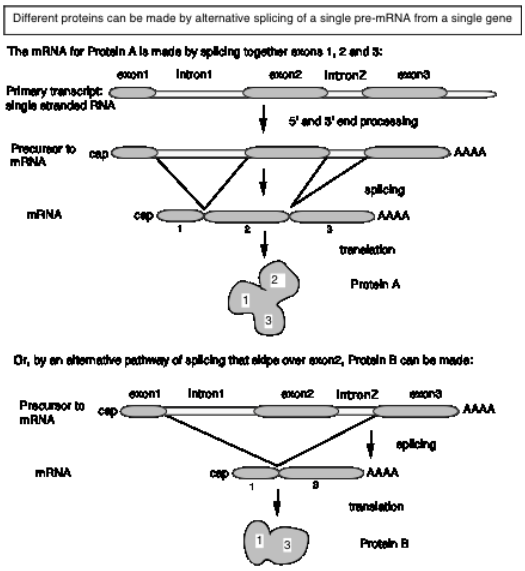


Figure 3.29.

Some segments of RNA may be included in the mature mRNA (exons) but not included on other spliced products. The alternative products may be made in different tissues or at different developmental stages - i.e. alternative splicing can be regulated.

Split genes may enhance the rate of evolution

Many exons encode a unit very close to a protein domain, e.g. the exons of leghemoglobin, or the variable and constant regions of immunoglobulins, or domains (e.g. "kringle") in EGF precursor that are also found in part of the LDL receptor. The exon organization tends to be well conserved in highly divergent species. Introns tend to occur between those portions of genes that encode structural domains of proteins.

Duplication of the exons encoding structural domains and subsequent recombination can lead to more rapid evolution of a new protein, essentially using the parts from earlier evolved genes. Analogous to building a house from prefabricated parts, as opposed to one nail and one board at a time - start with preassembled walls, roof joists etc.

However, the relationship between exons and structural domains of proteins is not exact, and some exon-intron boundaries vary (a little) in genes for different species. A different model holds that the introns are transposable elements (some certainly are - see later). They can insert anywhere in a gene, but they are least disruptive at domain boundaries, and these latter insertions are more likely to be fixed in a population than insertions into the middle of a region encoding a domain. So the results after long years of evolution is that the introns tend to be between region coding domains, but the gene was originally intact, not assembled from discrete exons.

Multigene families and gene clusters

Many eukaryotic genes are found in multiple copies. Some of them are developmentally regulated, such as *HOX*gene clusters and globin gene clusters .

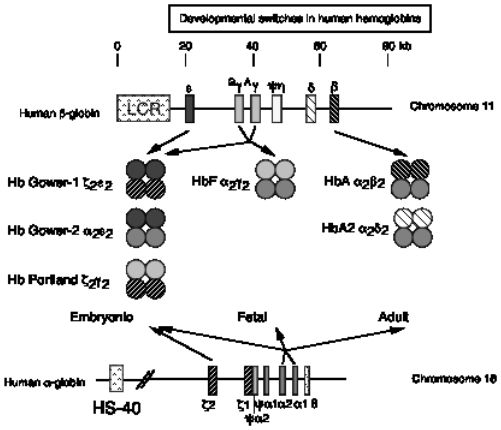


Figure 3.30.

A **multigene family** contains multiple genes of similar sequence encoding similar proteins; e.g. globin genes (Figure 3.30). Globin genes are expressed at different times of development. The order of developmental expression is the same as their order along the chromosome, e.g. the e-globin gene is expressed in early embryonic red cells, the g-globin gene is expressed at a high level in fetal red cells, and the b-globin gene is expressed in red cells after birth. As we will see later, this correlates with their distance from a dominant control element at the 5' end of the cluster, the Locus Control Region.

The order of *HOX*genes is also aligned with their spatial expression in the embryo. This is another example of alignment between chromosomal position and regulation of expression.

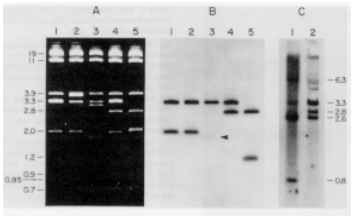
Other multi-gene families include those encoding histones, immunoglobulins, actins, cyclins, cyclin-dependent protein kinases, and rRNAs. Some of these families are linked in gene clusters, but others are dispersed around the genome. Having multiple copies of genes may be more the rule than the exception in eukaryotic genomes.

Experimental techniques that reveal multigene families include the following.

Purification and analysis of a particular kind of protein, e.g. hemoglobins, immunoglobulins, and many enzymes, *may reveal heterogeneity*. Further purification (via chromatography and electrophoresis) and sequencing can show that the observed heterogeneity is a result of related but not identical proteins, and one deduces that these similar proteins are encoded by multiple genes with similar sequences, i.e. a multigene family.

Analysis of the clones obtained by screening a library of cloned genomic DNA may reveal multiple related sequences, each with a distinctive restriction map. In many cases these are clones of different, related genes that comprise a multigene family (Figure 3.31).

Southern blot-hybridization of restriction-cleaved genomic DNA can reveal multiple copies of genes, simply as multiple bands on the hybridized blot. Although the number of fragments generated from total genomic DNA is too many to resolve on a gel, after transfer to a membrane, particular fragments can be visualized by hybridization with a specific probe. The number of hybridizing fragments is roughly correlated with the number of copies of related genes. Some genes are cleaved by the restriction enzyme, producing multiple bands, but some fragments can have multiple genes. A true measure of the number of related genes comes from more detailed restriction mapping or sequencing.



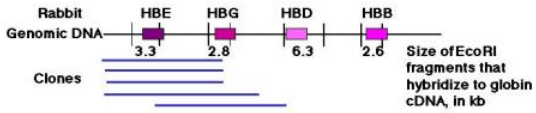


Figure 3.31. Blot-hybridization analysis of clones of genomic DNA and genomic DNA showing that mutliple copies of genes are present. A set of overlapping clones containing rabbit genomic DNA were digested and run on an agarose gel (panel A), blotted onto a membrane and hybridized with a radiolabeled probe that detected embryonic hemoglobin genes, and exposed to X-ray film. The resulting autoradiogram is shownin panel B. Panel C shows the results of a blot-hybridization analysis of rabbit total genomic DNA, using the same probe. Many of the same bands are seen as in the cloned DNA, confirming the existence of multiple hybridizing fragments. Mapping the fragments showed that they represented separate genes.

Keeping multigene families homogeneous

Sometimes multiple copies of genes are maintained as virtually identical over the course of evolution: e.g. rRNA genes, histone genes, a-globin genes (in primates). In these cases, the multiple copies are **coevolving**(**concerted evolution**).

sequence differences

Human: A | A | A | among human genes: 1%

between human & chimp5%

Chimp: A | A | A | among chimp genes: 1%

between chimp & monkey 10%

Monkey: A | A | A | among monkey genes: 1%

Since all three primates have 3 A genes, we infer that the common ancestor had 3 genes (the duplications preceded the speciation events). If in the time since human and chimp diverged, the A genes have diverged 5%, why haven't the A genes in human (e.g.) also diverged 5% from each other? They have been apart even longer than the human and chimp chromosomes carrying them! The A genes within a species are "talking to each other", or co-evolving or evolving in concert.

Sequence homogeneity in a multigene family can arise because of recent gene amplification (Figure 3.32 part1). In this case the genes have not been separate from each other long enough to accumulate variation in their sequences. Other multigene families have existed for a long time, but maintain sequence homogeneity despite ample opportunity for divergence. Two mechanisms have been seen that maintain similarity. The first is multiple rounds of unequal crossing over. As illustrated in Figure 3.32, part 2, the expansions and contractions of repeated genes can result in a new variant predominanting in the gene cluster. The other method for maintaining homogeneity is **gene conversion** between homologs. When a new mutation arises, it can be removed by conversion with the unmutated allele, or the mutation can be passed on the the other allele. Either way, the sequences of the two alleles becomes the same.

Sometimes the products of the gene duplications, or duplicative transpositions, accumulate mutations so they are no longer functional. These remnants of once-active genes are called **pseudogenes**.

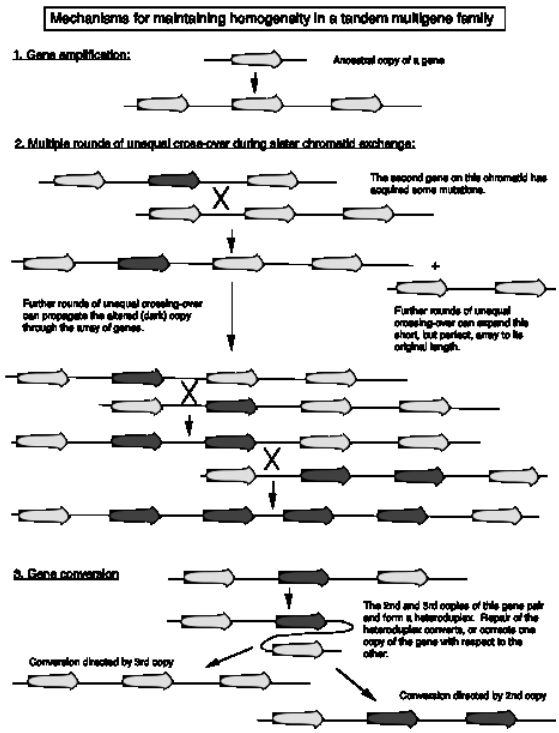


Figure 3.32.

3.10: Functional analysis of isolated genes

Gene expression

"Northern blots" or RNA blot-hybridization

In the reverse of Southern blot-hybridizations, one can separate RNAs by size on a denaturing agarose gel, and transfer them to nylon or other appropriate solid support. Labeled DNA can then be used to visualize the corresponding mRNA (Figure 3.33). Ed Southern initially used labeled rRNA to find the complementary regions in immobilized, digested DNA, so this "reverse" of Southern blot-hybridizations, i.e. using a labeled DNA probe to hybridize to immobilized RNA, is often referred to as **"Northern" blot-hybridizations**.

One can hybridize a labeled DNA clone to a panel of RNA samples from a wide variety of tissues to determine in what tissues a particular cloned gene is expressed (top panel of Figure 3.33. More precisely, this technique reveals the tissues in which the genes is transcribed into stable RNA. The results allow one to determine the **tissue specificity** of expression, e.g. a gene may only be expressed in liver, or only in erythroid cells (e.g. the b-globin gene). This helps give some general idea of the possible function of the gene, since it should reflect the function of that tissue. Other genes are expressed in almost all cells or tissue types (such as *GAPDH*); these are referred to as **housekeeping genes**. They are involved in functions common to all cells, such as basic energy metabolism, cell structure, etc. The relative amounts of RNA in the different lanes can be directly compared to see, e.g., which tissues express the gene most **abundantly**.

One can hybridize a labeled DNA clone to a panel of RNA samples from a progressive stages of development to determine the **developmental stage**when during development a particular cloned gene is expressed as RNA (bottom panel of Figure 3.33). For instance, a gene product may be required for determination decisions early in development, and only be expressed in early embryos.

Once the DNA sequence of the gene of interest is known, and its intron-exon structure determined, highly sensitive **RT-PCR assays**can be designed (Figure 3.34). The RNA from the cell or tissue of interest is copied into cDNA using reverse transcriptase and dNTPs, and then primers are annealed for PCR. Ideally, the primers are in different exons so that the product of amplifying the cDNA will be smaller than the product of amplifying the genomic DNA.

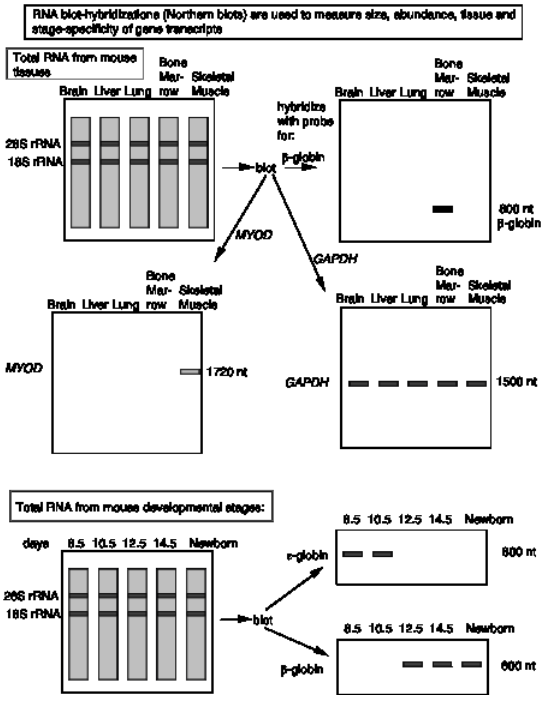


Figure 3.33.

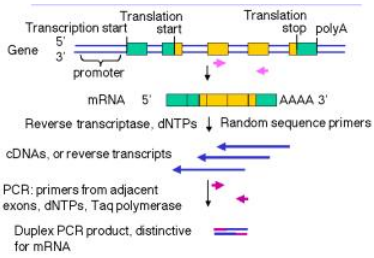


Figure 3.34.Reverse transcription-PCR (RT-PCR) assay for mRNA.

In situ hybridizations / immunochemistry

In complementary approaches, the labeled DNA can be hybridized *in situ*to thin sections of a tissue or embryo or other specimen, and the resulting pattern of grains visualized along the specimen in the microscope (Figure 3.35). Also, antibody probes against the protein product can be used to localize it in the specimen. This gives a more detailed picture of the **pattern of expression**, with resolution to the particular cells that are expressing the gene. The RNA blot-hybridization techniques described in a. above look at the RNA in all the cells from a tissue, and do not provide the level of resolution to single cells.

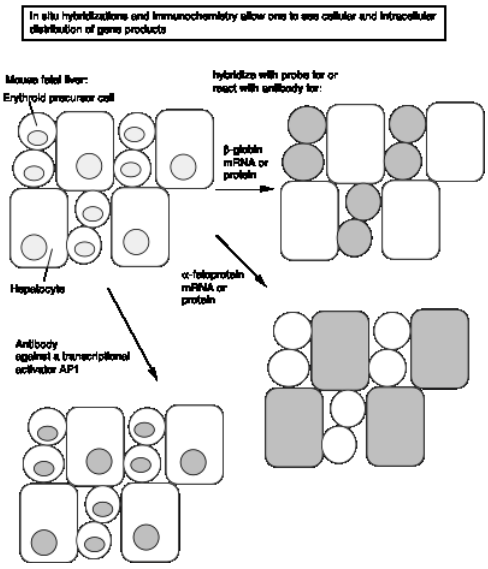


Figure 3.35.

Microarrays

As large numbers of sequenced mRNAs and genes become available, technology has been developed to look at expression of very large numbers of genes simulatneously. DNA sequences specific for each gene in a bacterium or yeast can be spotted in a high-density array with 400 r more spots. Some technologies use many more spots, with mutliple sequences per gene. Microarrays, or “gene chips” are available for many species, some with tens of thousands of different sequences or “probes.” RNA from different tissues can be converted to cDNA with a distinctive fluorescent label, and then hybridized to the gene chip. Differences in level of expression can be measured. Thus global changes in gene expression can now be measured.

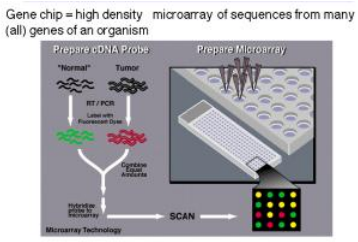


Figure 3.36.Hybridization of RNA to high density microarrays of gene sequences, or “gene chips”.

Database searches

An increasingly powerful approach is to determine candidates for the the function of your gene by **searching the databases**with the sequence, looking for matches to known proteins and genes. These matches provide clues as to protein function.

The power of this approach increases as the amount of sequences deposited in databases expand. Sequences of many genes are already known. The sequenced genes from more complex organisms, such as plants and animals, tend to be the ones more easily isolated using the techniques discussed in recombinant DNA technology. However, the sequences of genes expressed at a low level are starting to accumulate in the databases.

One remarkable advance in the past few years is the increasing number of organisms whose entire genome has been sequenced. About 10 bacterial genomes have been sequenced, and the number increases every few months. Genomics sequences for two eukaryotes are now available. That of the yeast *Saccharomyces cerevisiae*has been known for a few years, and the genome of the nematode *Caenorhabditis elegans*was completed in 1998. These sequences are being analyzed intensively, and a very high fraction of all the genes in each genome can be reliably detected using computational tools (one part of *bioinformatics*). It has become clear that many of the enzymes used in basic metabolism, regulation of the cell cycle, cellular signaling cascades, etc. are highly conserved across a broad phylogenetic spectrum. Thus it is common to find significant sequence matches in the genomes of model organisms when they are queried by the sequence of a previously unknown gene, e.g. from humans or mouse. The function already established for that gene in worms or yeast is a highly reliable guide to the function of the homologous gene in humans. The worm *C. elegans*is multicellular, and fate of each of its cells during development has been mapped. Thus it is possible that many functions involved in cellular interactions and cell-cell signaling will be conserved in this species, thus expanding the list of potential targets for a search in the databases.

This potential is being realized as working draft sequences of the human and mouse genomes are being analyzed. Within these data is a good approximation of sequences from virtually all human and mouse genes. Random clones have been partially sequenced from libraries of cDNAs from various human tissues, normalized to remove much of the products of abundant mRNAs and thus increasing the frequency of products of rare mRNAs. These sequences from the ends of the cDNA clones are called expressed sequence tags, or ESTs. The name is derived from the fact that since they are in cDNA libraries, they are obviously expressed at the level of mRNA, and some are used as tags in generating high-resolution maps of human chromosome. Hundreds of thousands of these have now been sequenced in collaborative efforts between pharmaceutical companies, other companies and universities. The database dbEST records all those in the public domain, and it is a strong complement to the databases recording all known sequences of genes. Many different parts of the same, or highly related, cDNAs, are recorded as separate entries in dbEST. Projects are underway to group all the sequences from the same (or highly related) gene into a a unified sequence. One example is the Unigene project at NCBI. The number of entries grows continually, but in the summer of 1998 there are about 50,000 entries, each representing about one gene. The number is higher now. Current estimates of the number of human genes are around 30,000, so it is possible that some UniGene clusters represent only parts of genes, and some genes match more than one cluster.

Very efficient search engines have been designed for handling queries to these databases, and several are freely available over the World Wide Web. One of the most popular and useful sites for this and related activities is maintained by the National Center for Biotechnology Information (**http://www.ncbi.nlm.nih.gov/**). Their Entrez browser provides integrated access to sequence, mapping and some functional information, PubMed provides access to abstracts of papers in journals in the National Library of Medicine, and the BLAST server allows rapid searches through various sequence databases. dbEST and the Unigene collection are maintained here, many genome maps are available, and three-dimensional structures of proteins and nucleic acids are available.

Make the protein product and analyze it

It is often possible to **express the gene** and make the encoded protein in large amounts. The protein can be purified and assayed for various enzymatic or other activities. Hypotheses for such activities may come from database searches.

Directed mutation

The previously describe approaches give some idea about gene function, but they do not firmly establish those functions. Indeed, this is a modern problem of trying to assign a function to an isolated gene. Several “reverse genetic” approaches can now be taken to tackle this problem. The most powerful approach to determining the physiological role(s) of a gene product is to **mutate** the gene in an appropriate organism and search for an **altered phenotype**.

The easiest experiment to do, but sometimes most difficult to interpret, is a **gain of function** assay. In this case, one forces expression of the gene in a transgenic organism, which often already has a wild type copy of the gene. One can look for a

phenotype resulting from **over-expression** in tissues where it is normally expressed, or **ectopic expression** in tissues where it is normally silent.

In some organisms, it is possible to engineer a **loss of function**of the gene. The most effective method is to use homologous recombination to replace the wild type gene with one engineered to have no function. This **knock-outmutation** will prevent expression of the endogenous gene and one can see the effects on the whole organism. Unfortunately, the efficiency of homologous recombination is low in many organisms and cell lines, so this is not always feasible. Other methods for knocking out expression are being developed, although the mechanism for their effect (when successful) is still being studied. In some cases, one can block expression of the endogenous gene by forcing production of **antisense**RNA. Another method that is effective in some, but currently not all organisms, is the use of **double-stranded, interfering RNA (RNAi)**. Duplex RNAs less than 30 nucleotide pairs long from the gene of interest can prevent expression of genes in worms, flies, and plants. Some success in mammals was recently reported.

Another way to generate a loss-of-function phenotype is to express **dominant negative alleles** of the gene. These mutant alleles encode stable proteins that form an aberrant structure that prevents functioning of the endogenous protein. This usually requires some protein-protein interaction (e.g. homodimers or heterodimers).

Localization on a genetic map

Sometimes the gene you have isolated maps to a region on a chromosome with a known function. Of course, many genes are probably located in that region, so it is critical to show that a candidate gene really is the one that when mutated causes an altered phenotype. This can be done by showing that a wild type copy of the candidate gene will restore a normal phenotype to the mutant. If a marker is known to be very tightly linked to the candidate gene, one can test whether this marker is always in linkage disequilibrium with the determinant of the mutant phenotype, i.e. in a large number of crosses, the marker for the candidate gene and the mutant phenotype never separated by recombination.

The mapping is often done with gene-specific probes for **in situ hybridizations** to mitotic chromosomes. One then aligns the hybridization pattern with the chromosome banding patterns to map the isolated gene. Another method is to hybridize to a panel of DNAs from hybrid cells that contain only part of the chromosomal complement of the genome of interest. This is particularly powerful with radiation hybrid panels.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

3.E: Isolating and Analyzing Genes (Exercises)

3.2 Altering the ends of DNA fragments for ligation into vectors.

(Adapted from POB)

a) Draw the structure of the end of a linear DNA fragment that was generated by digesting with the restriction endonuclease *Eco*RI. Include those sequences remaining from the *Eco*RI recognition sequence.

b) Draw the structure resulting from the reaction of this end sequence with DNA polymerase I and the four deoxynucleoside triphosphates.

c) Draw the sequence produced at the junction if two ends with the structure derived in (b) are ligated.

d) Design two different short synthetic DNA fragments that would permit ligation of structure (a) with a DNA fragment produced by a *Pst*I restriction digest. In one of these synthetic fragments, design the sequence so that the final junction contains the recognition sequences for both *Eco*RI and *Pst*I. Design the sequence of the other fragment so that neither the *Eco*RI nor the *Pst*I sequence appears in the junction.

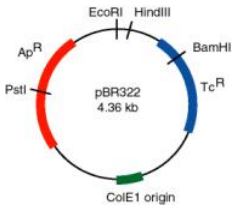
3.3. What properties are required of vectors used in molecular cloning of DNA?

3.4. A student ligated a *Bam*HI fragment containing a gene of interest to a pUC vector digested with BamHI, transformed *E. coli*with the mixture of ligation products and plated the cells on plates containing the antibiotic ampicillin and the chromogenic substrate X-gal. Which colonies should the student pick to find the ones containing the recombinant plasmid (with the gene of interest in pUC)?

3.5. Starting with an isolated mRNA, one wishes to make a double stranded copy of the mRNA and insert it at the *Pst*I site of pBR322 via G-C homopolymer tailing. One then transforms *E. coli*with this recombinant plasmid, selecting for tetracycline resistance. What are the four enzymatic steps used in preparing the cDNA insert? Name the enzymes and describe the intermediates.

3.6 A researcher needs to isolate a cDNA clone of giraffe actin mRNA, and she knows the size (Mr = 42,000) and partial amino acid sequence of giraffe actin protein and has specific antibodies against giraffe actin. After constructing a bank of cDNA plasmids from total mRNA of giraffe fibroblasts (dG-dC tailed into the *Pst*I site of pBR322), what methods of screening the bank could be used to identify the actin cDNA clone?

3.7 The restriction map of pBR322 is



The distance in base pairs between restriction sites is as follows:

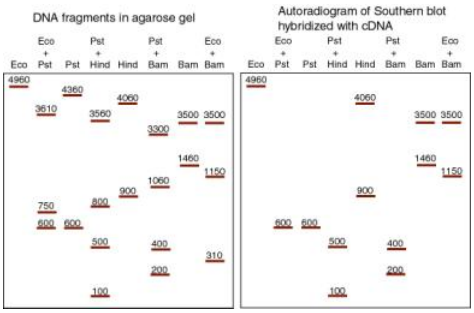
PstI to EcoRI 750 bp

EcoRI to HindIII 50 bp

HindIII to BamHI 260 bp

BamHI to PstI 3300 bp

A recombinant cDNA plasmid, pAlc-1, has double-stranded cDNA inserted at the *Pst*I site of pBR322, using a technique that retains this cleavage site at both ends of the insert. Digestion of pBR322 and pAlc-1 with restriction endonucleases gives the following pattern after gel electrophoresis (left). The sizes of the fragments are given in base pairs. The DNA fragments were transferred out of the gel onto nitrocellulose and hybridized with radiolabeled cDNA from wild-type *A. latrobus*(a Southern blot-hybridizaton). Hybridizing fragments are shown in the autoradiogam diagram on the right.



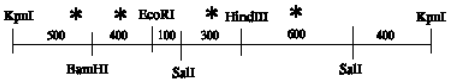
a) What is the size of the cDNA insert?

b) What two restriction endonucleases cleave within the cDNA insert?

c) For those two restriction endonucleases, each DNA fragment in the single digest is cut by *Pst*I into two DNA fragments in the double digest (i.e. the restriction endonuclease plus *Pst*I). Determine which fragments each single digest fragment is cut into, and use this information to construct a map.

d) Draw a restriction map for pAlc-1, showing sites for *Pst*I, *Eco*RI, *Bam*HI and *Hind*III. Indicate the distance between sites and show the cDNA insert clearly.

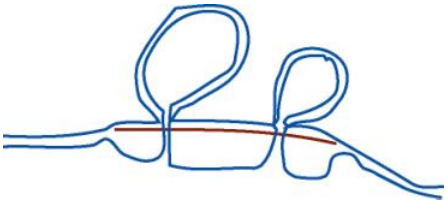
3.8. You isolate and clone a *Kpn*I fragment from *A. latrobus*genomic DNA that encodes the mRNA cloned in pAlc-1 (as analyzed in question 3.7). The restriction map of the genomic fragment is



Each fragment that hybridizes to pAlc-1 is indicated by an asterisk. What does this map, especially when compared to that in problem 3.7, tell you about the structure of the gene? Be as quantitative as possible.

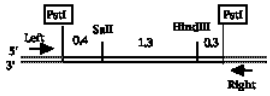
3.9. Some particular enzyme is composed of a polypeptide chain of 192 amino acids. The gene that encodes it has 1,440 nucleotide pairs. Explain the relationship between the number of amino acids in this polypeptide and the number of nucleotide pairs in its gene.

3.10. When viewed in the electron microscope, a hybrid between a cloned giraffe actin gene (genomic DNA) and mature actin mRNA looks like this:



What can you conclude about actin gene structure in the giraffe?

3.11. DNA complementary to pepper mRNA was synthesized using oligo (dT) as a primer for first strand synthesis. The second strand (synonymous with the mRNA) was then synthesized, and the population of double stranded cDNAs were ligated into a plasmid vector using a procedure that leaves PstI sites flanking the cDNA insert (i.e. the terminal PstI sites for each clone are not part of the cDNA). This cDNA library was screened for clones made from the mRNA from the pepper *yellow* gene. One clone was isolated, and subsequent analysis of the pattern of restriction endonuclease cleavage patterns showed it had the following structure:



The map shows the positions of restriction endonuclease cleavage sites and the distance between them in kilobases (kb). The map of the cDNA insert is shown with solid lines, and plasmid vector DNA flanking the cDNA is shown as dotted lines. The top strand is oriented 5' to 3' from left to right, and the bottom strand is oriented 5' to 3' from right to left. The positions and orientations of

two oligonucleotides to prime synthesis for sequence determination are shown, and are placed adjacent to the strand that will be synthesized in the sequencing reaction.

a) Oligonucleotides that anneal to the plasmid vector sequences that flank the duplex cDNA insert were used to prime synthesis of DNA for sequencing by the Sanger dideoxynucleotide procedure. A primer that annealed to the vector sequences to the left of the map shown above generated the sequencing gel pattern shown below on the left. A primer that annealed to the vector sequences to the right of the map shown above generated the sequencing gel pattern shown below on the right. The gels were run from the negative electrode at the top to the positive electrode at the bottom, and the segment presented is past the PstI site (i.e. do not look for a PstI recognition site).

Left primer Right primer

G	A	T	C			G	A	T	C
		___					___		
	___							___	
___								___	
	___							___	
		___							___
			___			___			
		___							___
			___			___			
			___			___			
___							___		
___								___	
	___							___	
___								___	
___								___	
___								___	
___								___	
___								___	
___								___	
___								___	

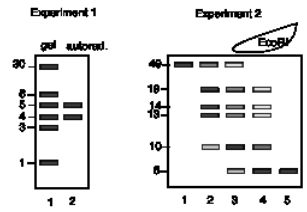
a) What is the DNA sequence of the left and right ends of the insert in the cDNA clone? Be sure to specify the 5' to 3' orientation, and the strand (top or bottom) whose sequence is reported. The terms left, right, top and bottom all refer to the map shown above for the cDNA clone.

b) Which end of the cDNA clone (left or right in the map above) is most likely to include the sequence synonymous with the 3' end of the mRNA?

c) What restriction endonuclease cleavage sites do you see in the sequencing data given?

3.12. Genomic DNA from the pepper plant was ligated into EcoRI sites in a l phage vector to construct a genomic DNA library. This library was screened by hybridization to the *yellow*cDNA clone. The pattern of EcoRI cleavage sites for one clone that hybridized to the *yellow*cDNA clone was analyzed in two experiments.

In the first experiment, the genomic DNA clone was digested to completion with EcoRI, the fragments separated on an agarose gel, transferred to a nylon filter, and hybridized with the radioactive *yellow*cDNA clone. The digest pattern (observed on the agarose gel) is shown in lane 1, and the pattern of hybridizing fragments (observed on an autoradiogram after hybridization) is shown in lane 2. Sizes of the EcoRI fragments are indicated in kb. The right arm of this l vector is 6 kb long, and the left arm is 30 kb.



In the second experiment, the genomic DNA clone was digested with a range of concentrations of EcoRI, so that the products ranged from a partial digest to a complete digest. The cleavage products were annealed to a radioactive oligonucleotide that hybridized only to the right cohesive end (*cos*site) of the l vector DNA. This simply places a radioactive tag at the right end of all the products of the reaction that extend to the right end of the l clone (partial or complete); digestion products that do not include the right end of the l clone will not be seen. The results of the digestion are shown above, on the right. Lane 1 is the clone of genomic DNA in l that has not been digested, lane 5 is the complete digest with EcoRI, and lanes 2, 3 and 4 are partial digests using increasing amounts of EcoRI. The sizes of the radioactive DNA fragments (in kb) are given, and the density of the fill in the boxes is proportional to the intensity of the signal on the autoradiogram.

a) What is the map of the EcoRI fragments in the genomic DNA clone, and which fragments encode mRNA for the *yellow*gene? You may wish to fill in the figure below; the left and right arms of the l vector are given. Show positions of the EcoRI cleavage sites, distances between them (in kb) and indicate the fragments that hybridize to the cDNA clone.

EcoRI EcoRI

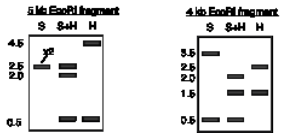
Left arm ___| |_ Right arm

(30 kb) (6 kb)

In a third experiment, the pepper DNA from the genomic DNA clone was excised, hybridized with *yellow*mRNA under conditions that favor RNA-DNA duplexes and examined in the electron microscope to visualize R-loops. A pattern like the following was observed. The lines in the figure can be duplex DNA, RNA-DNA duplexes and single-stranded DNA.

b) What do the R-loop data indicate? Please draw an interpretation of the R-loops, showing clearly the two DNA strands and the mRNA and distinguishing between the template (bottom, or message complementary) and nontemplate (top, or message synonymous) strands.

The EcoRI fragments that hybridize to the *yellow*cDNA clone were isolated and digested with SalI (S in the figure below), HindIII (H), and the combination of SalI plus HindIII (S+H). The resulting patterns of DNA fragments are shown below; all will hybridize to the *yellow*cDNA clone. Cleavage of the 5 kb EcoRI fragment with SalI generates two fragments of 2.5 kb.



c) What are the maps of the SalI and HindIII site(s) in each of the EcoRI fragments? Show positions of the cleavage sites and distances between them on the diagram below.

5 kb EcoRI fragment: 4 kb EcoRI fragment:

EcoRI EcoRI EcoRI EcoRI

|_____| |_____|

d) Compare these restriction maps with that of the cDNA clone (problem 1.38) and the R-loops shown above. Assuming that the SalI and HindIII sites in the genomic DNA correspond to those in the cDNA clone, what can you deduce about the intron/exon structure of the *yellow*gene(s) contained within the 5 kb and 4 kb EcoRI fragments? Please diagram the exon-intron structure in as much detail as the data permit (i.e. show the size of the intron(s) and positions of intron/exon junctions as precisely as possible).

5 kb EcoRI fragment: 4 kb EcoRI fragment:

EcoRI EcoRI EcoRI EcoRI

|_____| |_____|

e) Considering all the data (maps of cDNA and genomic clones and R-loop analysis), what can you conclude about the number and location(s) of *yellow*gene(s) in this genomic clone?

3.13 You have isolated an 1100 base pair (bp) cDNA clone for a gene called *azure* that when mutated causes blue eyes in frogs. You also isolate a 3000 bp *Sal*I genomic DNA fragment that hybridizes to the *azure*cDNA. The map of the *azure* cDNA is as follows, with sizes of fragments given in bp.

Digestion of the 3000 bp *Sal*I fragment of genomic DNA with the indicated restriction endonucleases yields the following pattern of fragments, all of which hybridize to the *azure*cDNA. Remember that the starting fragment has *Sal*I sites at each end. Sizes of fragments are in bp.

Restriction enzymes

BamHI Bam+Pst PstI Pst+Eco EcoRI Bam+Eco

2700

2300

2000

1900 1900

1200

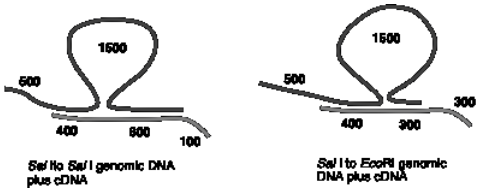
1100 1100

800

700 700 700

300 300 300

The *Sal*I to *Sal*I (3000 bp) genomic fragment was hybridized to the 1100 bp cDNA fragment, and the heteroduplexes were examined in the electron microscope. Measurements on a large number of molecules resulted in the determination of the sizes indicated in the structure on the left, i.e. duplex regions of 400 and 600 bp are interrupted by a single stranded loop of 1500 nucleotides and are flanked by single stranded regions of 500 and 100 nucleotides. When the same experiment is carried out with the 2700 bp *Sal*I to *Eco*RI genomic DNA fragment hybridized to the cDNA fragment, the structure on the right is observed.



a) What is the restriction map of the 3000 bp *Sal*I to *Sal*I genomic DNA fragment from the *azure*gene? Specify distances between sites in base pairs.

b) How many introns are present in the *azure*genomic DNA fragment?

c) Where are the exons in the *azure*genomic DNA fragment? Draw the exons as boxes on the restriction map of the 3000 bp *Sal*I to *Sal*I genomic DNA fragment? Specify (in base pairs) the distances between restriction sites and the intron/exon boundaries.

3.14 The T-cell receptor is present only on T-lymphocytes, not on B-lymphocytes or other cells. Describe a strategy to isolate the T-cell receptor by subtractive hybridization, using RNA from T-lymphocytes and from B-lymphocytes.

3.15. How many exons are in the human insulin (*INS*) gene, how big are they, and how large are the introns that separate them? Use three different bioinformatic approaches to answer this.

a. Align the available genomic sequence containing *INS*(encoding insulin) with the sequence of the mRNA to find exons and introns in the *INS*gene. The sequence files are:

*INS*mRNA: accession number NM_000207

INS gene (includes part of *TH*and *IGF2*in addition to *INS*): accession number L15440

Files can be obtained from NCBI (http://www.ncbi.nlm.nih.gov), or from the course web site (www.bmb.psu.edu/Courses/bmb400/default.htm)

Align the mRNA (cDNA) and genomic sequence using the *BLAST2*sequences server at

http://www.ncbi.nlm.nih.gov/blast/

and the *sim4*server at

pbil.univ-lyon1.fr/sim4.html

*Sim4*is designed to take into account terminal redundancy at the exon/intron junctions, whereas *BLAST2*does not. Do you see this effect in the output?

b. Use the *ab initio*exon finding program *Genscan*, available at

genes.mit.edu/GENSCAN.html

to predict exons in the *INS*genomic sequence (L15440).

How does this compare with the results of analyzing with the program *genscan*?

c. What do you see for *INS*at the Human Genome Browser and Ensembl? They are accessed at:

http://genome.ucsc.edu/goldenPath/hgTracks.html

http://www.ensembl.org/

CHAPTER OVERVIEW

4: Genomes and Chromosomes

Chapter 4 has two parts: genomes and chromosomes. Initial studies on genome structure used the kinetics of hybridization of nucleic acids to determine the bulk features of genomes, e.g. how big is a particular genome, how much is single-copy and how much is repeated, and how much of that genome is transcribed into nuclear or mRNA in a particular tissue. More detailed whole-genome mapping and sequencing projects are now revolutionizing biology. Some of the information on whole-genome sequences of bacteria, the yeast *Saccharomyces cerevisiae*, worms, flies and mammals (humans and mice) will be reviewed. All this genomic DNA is packaged into chromosomes, and Chapter 4 will also review some of their cytological features, and discuss their packaging into nucleosomes and higher order structure. Transitions between types of chromatin structure are fundamental to issues of gene regulation in eukaryotes; this will be explored in more biochemical detail in Part Four of the text.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

4.1: Reassociation kinetics measure sequence complexity

Low complexity DNA Sequences Reanneal Faster than Do High Complexity Sequences

The components of complex genomes differ not only in repetition frequency (highly repetitive, moderately repetitive, single copy) but also in sequence complexity. **Complexity** (denoted by N) **is the number of base pairs of unique or nonrepeating DNA** in a given segment of DNA, or component of the genome. This is different from the length (L) of the sequence if some of the DNA is repeated, as illustrated in this example.

E.g. consider 1000 bp DNA.

- 500 bp is sequence a, present in a single copy.
- 500 bp is sequence b (100 bp) repeated 5 times:

a b b b b b

|_____|__|__|__|__|__| L = length = 1000 bp = a + 5b

N = complexity = 600 bp = a + b

Some viral and bacteriophage genomes have almost no repeated DNA, and L is approximately equal to N. But for many genomes, repeated DNA occupies 0.1 to 0.5 of the genome, as in this simple example. The key result for genome analysis is that **less complex DNA sequences renature faster** than do more complex sequences. Thus determining the rate of renaturation of genomic DNA allows one to determine how many kinetic components (sequences of different complexity) are in the genome, what fraction of the genome each occupies, and the repetition frequency of each component.

Before investigating this in detail, let's look at an example to illustrate this basic principle, i.e. the inverse relationship between reassociation kinetics and sequence complexity.

Inverse Relationship between Reassociation Kinetics and Sequence Complexity

Let a, b, ... z represent a string of base pairs in DNA that can hybridize (see Figure 4.2.). For simplicity in arithmetic, we will use 10 bp per letter.

- DNA 1 = ab (This is very low sequence complexity, 2 letters or 20 bp)
- DNA 2 = cdefghijklmnopqrstuv. (This is 10 times more complex (20 letters or 200 bp)).
- DNA 3 =izyajczkblqfreighttrainrunninsofastelizabethcottonqwftzxvbifyoudontbelieveimleavingyoujustcountthe

  daysimgonerxcvwpowentdowntothecrossroadstriedtocatchariderobertjohnsonpzvmwcomeonhomeintomykitchentrad.

(This is 100 times more complex (200 letters or 2000 bp).

A solution of 1 mg DNA/ml is 0.0015 M (in terms of moles of bp per L) or 0.003 M (in terms of nucleotides per L). We'll use 0.003 M = 3 mM, i.e. 3 mmoles nts per L. (nts = nucleotides).

Consider a 1 mg/ml solution of each of the three DNAs. For DNA 1, this means that the sequence ab (20 nts) is present at 0.15 mM or 150 mM (calculated from 3 mM / 20 nt in the sequence). Likewise, DNA 2 (200 nts) is present at 15 mM, and DNA 3 is present at 1.5 mM. Melt the DNA (i.e. dissociate into separate strands) and then allow the solution to reanneal, i.e. let the complementary strand reassociate.

Since the rate of reassociation is determined by the rate of the initial encounter between complementary strands, the higher the concentration of those complementary strands, the faster the DNA will reassociate. So for a given overall DNA concentration, the simple sequence (ab) in low complexity DNA 1 will reassociate 100 times faster than the more complex sequence (izyajcsk ....trad) in the higher complexity DNA 3. **Fast reassociating DNA is low complexity.**

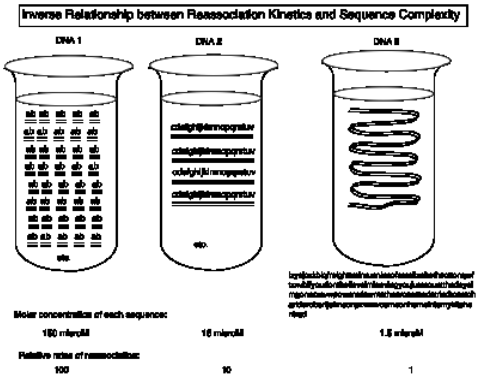


Figure 4.2.

Kinetics of renaturation

In this section, we will develop the relationships among rates of renaturation, complexity, and repetition frequency more formally.

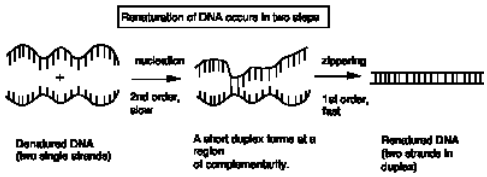


Figure 4.3.

The time required for half renaturation is inversely proportional to the rate constant. Let C = concentration of single-stranded DNA at time t (expressed as moles of nucleotides per liter). The rate of loss of single-stranded (ss) DNA during renaturation is given by the following expression for a second-order rate process:

$$\frac{-dC}{dt} = kC^2 \tag{4.1.1}$$

or

$$\frac{dC}{C^2} = -kdt \tag{4.1.2}$$

Integration and some algebraic substitution shows that

$$\frac{C}{C_o} = \frac{1}{1 + kC_o t} \tag{4.1.3}$$

Thus, at half renaturation, when

$$\frac{C}{C_o} = 0.5 \text{ at } t = t_{1/2} \tag{4.1.4}$$

one obtains:

$$C_o t_{1/2} = \frac{1}{k} \tag{4.1.5}$$

where k is the rate constant in in liters (mole nt)$^{-1}$ sec^{-1}

> *The rate constant for renaturation is inversely proportional to sequence complexity.*

The rate constant, k, shows the following proportionality:

$$k \propto \frac{\sqrt{L}}{N} \tag{4.1.6}$$

where

- L = length and

- N = complexity.

Empirically, the rate constant *k* has been measured as

$$k = 3x10^5 \frac{\sqrt{L}}{N} \tag{4.1.7}$$

in 1.0 M Na^+ at $T = T_m - 25^oC$

The time required for half renaturation (and thus $C_o t_{1/2}$) is directly proportional to sequence complexity.

From Equations 4.1.5 and 4.1.6,

$$C_0 t_{1/2} \propto \frac{N}{\sqrt{L}} \tag{4.1.8}$$

For a renaturation measurement, one usually shears DNA to a constant fragment length L (e.g. 400 bp). Then L is no longer a variable, and

$$C_o t_{1/2} \propto N \quad . \tag{4.1.9}$$

The data for renaturation of genomic DNA are plotted as $C_0 t$ curves:

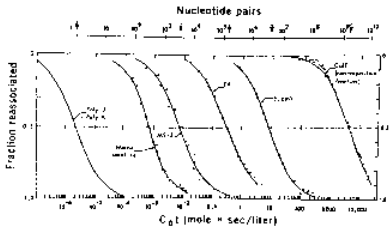


Figure 4.4.

Renaturation of a single component is complete (0.1 to 0.9) over 2 logs of $C_0 t$ (e.g., 1 to 100 for *E. coli* DNA), as predicted by Equation 4.1.3.

> *Sequence complexity is usually measured by a proportionality to a known standard*

If you have a standard of known genome size, you can calculate N from $C_0 t_{1/2}$:

$$\frac{N^{unknown}}{N^{known}} = \frac{C_0 t_{1/2}^{unknown}}{C_0 t_{1/2}^{known}} \quad (6) \tag{4.1.10}$$

A known standard could be

- *E. coli* with N = 4.639 x 10^6 bp
- pBR322 with N = 4362 bp

More complex DNA sequences renature more slowly than do less complex sequences. By measuring the rate of renaturation for each component of a genome, along with the rate for a known standard, one can **measure the complexity** of each component.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

4.2: Analysis of Renaturation curves with Multiple Components

In this section, the analysis in Section 4.2 is applied quantitatively in an example of renaturation of genomic DNA. If an unknown DNA has a single kinetic component, meaning that the fraction renatured increases from 0.1 to 0.9 as the value of C_0t increases 100-fold, then one can calculate its complexity easily. Using equation (6), all one needs to know is its $C_0t_{1/2}$, plus the $C_0t_{1/2}$ and complexity of a standard renatured under identical conditions (initial concentration of DNA, salt concentration, temperature, etc.).

The same logic applies to the analysis of a genome with multiple kinetic components. Some genomes reanneal over a range of C_0tvalues covering many orders of magnitude, e.g. from 10^{-3} to 10^{4}. Some of the DNA renatures very fast; it has low complexity, and as we shall see, high repetition frequency. Other components in the DNA renature slowly; these have higher complexity and lower repetition frequency. The only new wrinkle to the analysis, however, is to treat each kinetic component independently. This is a reasonable approach, since the DNA is sheared to short fragments, e.g. 400 bp, and it is unlikely that a fast-renaturing DNA will be part of the same fragment as a slow-renaturing DNA.

Some terms and abbreviations need to be defined here.

- *f* = fraction of genome occupied by a component
- $C_0t_{1/2}$ for pure component = (*f*) ($C_0t_{1/2}$ measured in the mixture of components)
- R = repetition frequency
- G = genome size. G can be measured chemically (e.g. amount of DNA per nucleus of a cell) or kinetically (see below).

One can read and interpret the C_0t curve as follows. One has to estimate the number of components in the mixture that makes up the genome. In the hypothetical example in Figure 4.5, three components can be seen, and another is inferred because 10% of the genome has renatured as quickly as the first assay can be done. The three observable components are the three segments of the curve, each with an inflection point at the center of a part of the curve that covers a 100-fold increase in C_0t (sometimes called 2 logs of \(C_0t). The fraction of the genome occupied by a component, *f*, is measured as the fraction of the genome annealing in that component. The measured$C_0t_{1/2}$ is the value of C_0t at which half the component has renatured. In Figure 4.5, component 2 renatures between C_0t values of 10^{-3} and 10^{-1}, and the fraction of the genome renatured increased from 0.1 to 0.3 over this range. Thus *f* is 0.3-0.1=0.2. The C_0t value at half-renaturation for this component is the value seen when the fraction renatured reached 0.2 (i.e. half-way between 0.1 and 0.3; this C_0t value is 10^{-2}, and it is referred to as the $C_0t_{1/2}$for component 2 (measured in the mixture of components). Values for the other components are tabulated in Figure 4.5.

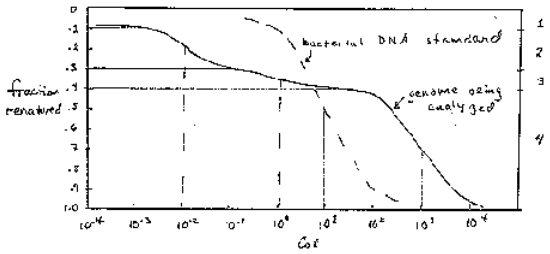


Figure 4.5.

All the components of the genome are present in the genomic DNA initially denatured. Thus the value for C_0 is for all the genomic DNA, not for the individual components. But once one knows the fraction of the genome occupied by a component, one can calculate the C_0 for each individual component, simply as C_0 ´ *f*. Thus the $C_0t_{1/2}$ for the individual component is the $C_0t_{1/2}$ (measured in the mixture of components) ´ *f*. For example the $C_0t_{1/2}$ for individual (pure) component 2 is 10^{-2} ´ 0.2 = 2 ´ 10^{-3} .

Knowing the measured $C_0t_{1/2}$for a DNA standard, one can calculate the complexity of each component.

Nn= C_0t_{1/2}_{pure}, n

$$N^{nn} = \frac{C_0t_{1/2}^{nn}}{C_0t_{1/2}^{nn}} \times \frac{3 \times 10^9 bp}{10}$$

- where n refers to the particular component, i.e. (1, 2, 3, or 4)

The repetition frequency of a given component is the total number of base pairs in that componentdivided by the complexity of the component. The total number of base pairs in that component is given by *fn* ´ *G*.

Rn =

For the data in Figure 4.5, one can calculate the following values:

Component	f	$C_0t_{1/2}$, mix	$C_0t_{1/2}$, pure	N (bp)	RR
1 foldback	0.1	< 10^{-4}	< 10^{-4}		
2 fast	0.2	10^{-2}	2 x 10^{-3}	600	10^5
3 intermediate	0.1	1	0.1	3 x 10^4	10^3
4 slow (single copy)	0.6	103	600	1.8 x 10^8	1
std bacterial DNA			10	3 x 10^6	1

The genome size, *G*, can be calculated from the ratio of the complexity and the repetition frequency.

$$G = \frac{N^{tot.}}{f^{tot.}} = \frac{1.8 \times 10^8}{0.6} = 3 \times 10^8 bp$$

E.g. If G = 3 x 108 bp, and component 2 occupies 0.2 of it, then component 2 contains 6 x 107 bp. But the complexity of component 2 is only 600 bp. Therefore it would take 105 copies of that 600 bp sequence to comprise 6 x 10^7 bp, and we surmise that R = 105.

Exercise 4.1

If one substitutes the equation for *Nn* and for *G* into the equation for *Rn*, a simple relationship for *R* can be derived in terms of $C_0t_{1/2}$ values measured for the mixture of components . What is it?

Types of DNA in each kinetic component for complex genomes

Eukaryotic genomes usually have multiple components, which generates complex *C0t* curves. Figure 4.6 shows a schematic *C0t* curve that illustrates the different kinetic components of human DNA, and the following table gives some examples of members of the different components.

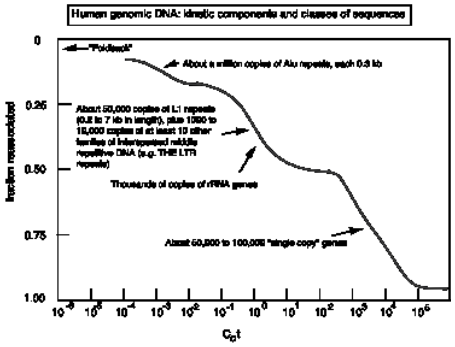


Figure 4.6.

Table 4.2. Four principle kinetic components of complex genomes

Renaturation kinetics	*C0t* descriptor	Repetition frequency	Examples
too rapid to measure	"foldback"	not applicable	inverted repeats
fast renaturing	low *C0t*	highly repeated, > 10^5 copies per cell	interspersed short repeats (e.g. human *Alu* repeats); tandem repeats of short sequences (centromeres)
intermediate renaturing	mid *C0t*	moderately repeated, 10-10^4 copies per cell	families of interspersed repeats (e.g. human L1 long repeats); rRNA, 5S RNA, histone genes
slow renaturing	high *C0t*	low, 1-2 copies per cell, "single copy"	most structural genes (with their introns); much of the intergenic DNA

N, *R* for repeated DNAs are averages for many families of repeats. Individual members of families of repeats are similar but not identical to each other.

The emerging picture of the human genome reveals approximately 30,000 genes encoding proteins and structural or functional RNAs. These are spread out over 22 autosomes and 2 sex chromosomes. Almost all have introns, some with a few short introns and others with very many long introns. Almost always a substantial amount of intergenic DNA separates the genes.

Several different families of repetitive DNA are interspersed throughout the the intergenic and intronic sequences. Almost all of these are repeats are vestiges of transposition events, and in some cases the source genes for these transposons have been found. Some of the most abundant families of repeats transposed via an RNA intermediate, and can be called **retrotransposons**. The most abundant repetitive family in humans are ***Alu* repeats**, named for a common restriction endonuclease site within them. They are about 300 bp long, and about 1 million copies are in the genome. They are probably derived from a modified gene for a small RNA called 7SL RNA. (This RNA is involved in translation of secreted and membrane bound proteins). Genomes of species from other mammalian orders (and indeed all vertebrates examined) have roughly comparable numbers of short interspersed repeats independently derived from genes encoding other short RNAs, such as transfer RNAs.

Another prominent class of repetitive retrotransposons are the long**L1 repeats**. Full-length copies of L1 repeats are about 7000 bp long, although many copies are truncated from the 5' end. About 50,000 copies are in the human genome. Full-length copies of recently transposed L1s and their sources genes have two open reading frames (i.e. can encode two proteins). One is a multifunctional protein similar to the *pol* gene of retroviruses. It encodes a functional reverse transcriptase. This enzyme may play a key role in the transposition of all retrotransposons. Repeats similar to L1s are found in all mammals and in other species, although the L1s within each mammalian order have features distinctive to that order. Thus both short interspersed repeats (or SINEs) and the L1 long interspersed repeats (or LINEs) have expanded and propogated independently in different mammalian orders.

Both types of retrotransposons are currently active, generating *de novo*mutations in humans. A small subset of SINEs have been implicated as functional elements of the genome, providing post-transcriptional processing signals as well as protein-coding exons for a small number of genes.

Other classes of repeats, such as L2s (long repeats) and MIRS (short repeats named mammalian interspersed repeats), appear to predate the mammalian radiation, i.e. they appear to have been in the ancestral eutherian mammal. Other classes of repeats are transposable elements that move by a DNA intermediate.

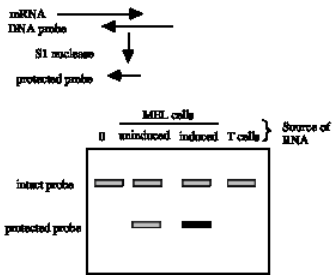
Other common interspersed repeated sequences in humans

4.3: RNA Abundance

The availability of cloned DNA probes for many genes has greatly facilitated the analysis of amounts of RNAs in different cells or under different conditions. For instance, it is very common to label a DNA probe that will hybridize to mRNA; the DNA comes from either a cDNA clone or a genomic clone containing an exon. The labeled probe is then hybridized to total or polyA-containing RNA (the latter is called polyA+ RNA, and is roughly equivalent to mRNA) from a cell. The concentration of the probe is much greater than the concentration of the target mRNA for the specific gene, thus the probe is in vast excess and all mRNA from the gene of interest should be driven into a duplex with the probe. The amount of probe protected from digestion by a single-strand specific nuclease such as nuclease S1 gives a measure of the amount of the specific mRNA that is in the cell. (This situation differs in some important aspects from the materialon estimating numbers of genes expressed and abundance from the kinetics of RNA-driven reactions. In that material, one was looking at entire populations of mRNAs, whereas in this situation, one is looking at only one mRNA - the one complementary to the labeled probe.)

[Two technical notes: The diagnostic assay here measures the amount of labeled DNA in duplex and the unhybridized DNA is digested. If the DNA probe is originally double-stranded, it is initially denatured prior to hybridization, but now how do you distinguish between nuclease protection arising from DNA-mRNA duplexes versus those that arise from the two strands of DNA reannealing? The cleanest approach is to just synthesize and label the strand of DNA complementary to the mRNA; this can be done by appropriate choices of primers for synthesis of DNA from plasmids carrying the DNA used as a probe. Alternatively, a labeled duplex DNA probe can be prepared that extends past the mRNA coding portion of a gene, so that the DNA-DNA duplex resulting from reannealing is larger than the DNA-RNA duplex resulting from hybridization to mRNA. Also, hybridization conditions with high concentrations of salt and formamide are used that favor DNA-RNA duplexes over DNA-DNA duplexes. (2) An equivalent approach is to synthesize an RNA probe derived from the cloned DNA; this "complementary RNA" forms a stronger duplex with the mRNA than does cDNA; RNA-RNA duplexes are stronger than RNA-DNA duplexes under conditions of high salt and formamide concentrations. The fragments protected from digestion by RNases are then detected.]

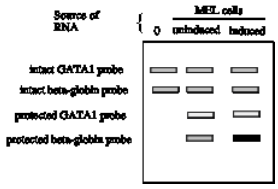
a) Murine erythroleukemia (MEL) cells are equivalent to proerythroblasts, immortalized by the Friend virus complex so that they can grow continuously in culture. Treatment with small organic compounds like dimethylsulfoxide (DMSO) will induce them to mature on to erythroblasts, with a substantial increase in the expression of erythroid specific genes (the mechanism for this induction is still unknown). Let's say that you isolated total RNA from both uninduced (untreated) cells and an equal number of DMSO-induced cells. The RNA samples were hybridized to an excess of a radiolabeled DNA probe from a mouse b-globin gene, and the amount of probe hybridized to the mRNA was determined by treatment of the samples with nuclease S1, electrophoresis on a denaturing polyacrylamide gel, and measuring the amount of radioactivity in the fragment resulting from the mRNA-DNA duplex. An illustration of the heteroduplex, the nuclease S1 treatment, and the resultant autoradiograph of the gel are shown below. The protected fragment from uninduced cells had 10,000 cpm, and the protected fragment from induced cells had 500,000 cpm. A negative control with RNA from a T-lymphocytic cell line, which produces no globin mRNA, gave no protection, i.e. 0 cpm for the diagnostic fragment. The expression of this b-globin gene is induced how much in MEL cells treated with DMSO?



b) The previous assay gives the relative amounts of the mRNA under the two conditions, and this is an extremely powerful and widely used assay. But what does this mean in terms of mRNA molecules per cell, i.e. how does the abundance change upon induction? One can alter this assay somewhat to get a measure of abundance, similar in principle to the calculations in Section VIIF. First, one needs a measure of the number of mRNA molecules per cell. Let's say that you harvested 107 MEL cells and isolated 3 mg of polyA+ RNA (essentially mRNA). What is the total number of mRNA molecules per MEL cell, assuming an average length of mRNA of 2000 nucleotides?

c) If one labels the RNA in the MEL cells, e.g. by growing the cells in the presence of [3H] uridine, which is incorporated only into RNA, then the isolated, labeled polyA+ RNA can be hybridized to an excess of the (now unlabeled) DNA complementary to the

mRNA of interest. RNA in duplex with DNA can be detected by its protection from digestion by nucleases such as RNase A and RNase T1; the resulting autoradiograph would look something like that shown below, with bands containing more radioactivity represented as a darker fill. Since the DNA is still in excess, all the mRNA complementary to the probe should be driven into duplex, and one can readily measure the fraction of polyA+ RNA complementary to each probe. The following table provides some representative, idealized data for polyA+ RNA from uninduced and induced MEL cells, including the total input RNA (not treated with nucleases) and the amount protected from nuclease digestion by hybridization with an excess of b-globin gene DNA, DNA encoding the erythroid transcription factor GATA1, and DNA encoding ovalbumin (which is not expressed in MEL cells, i.e. it is a negative control). What fraction of the mRNA (or polyA+ RNA) is composed of mRNA from these three genes, and what is their abundance in uninduced and induced cells?



DNA probe	cpm protected uninduced MEL cells	cpm protected induced MEL cells
[input labeled RNA]	[1,000,000]	[1,000,000]
b-globin	5,000	250,000
GATA1	25	25
ovalbumin	0	0

d) In general, what is the distribution of mRNAs in a particular type of differentiated cell, i.e. how abundant are the different complexity classes of mRNA?

Use of databases of sequences, mutations, and functional data

4.6 We used arginine biosynthesis to illustrate complementation analysis and construction of a pathway. The steps involved in arginine synthesis are also part of the urea cycle. One of the enzymes catalyzes the formation of citrulline from carbamoyl phosphate and ornithine. Let's find out more about this enzyme, called ornithine transcarbamoylase, or OTC.

Use your favorite Web browser to go to the URL for NCBI (National Center for Biotechnology Information).

- http://www.ncbi.nlm.nih.gov/
- Click on the Entrez button. Entrez provides a portal to many types of information at this server. Let's start with DNA and protein sequences.
- Click on the Nucleotides button.
- Enter "X00210" and press the Search button. Do not enter the quotation marks, and those are zeros and a one, not O or l.
- You should get a report on the gene for OTC in *E. coli*, called *argI*.

1. How large is the protein-coding region, from translation initiation codon to the termination codon? How big is the encoded protein?
2. Where is the *argI* gene on the *E. coli*chromosome? Go back to the Entrez server (where you clicked on Nucleotides before). Click on Genomes, and then select *Escherichia coli*. Enter "argI" in the Search window (don't enter the quotes, and that is the letter I "eye" not a "one").

4.7 Is the *E. coli* OTC protein related to any other proteins in the sequence databases? You need to get the protein sequence, which you can do by clicking on *argI*while you are at the genome map, or you can go back to the entry for the gene (accession number X00210). If you are at the GenBank Report for entry X00210, you need to click on the Protein button at the top of the page, and then select FastA Report from the next page. (If you take the default path the GenPept Report, that is OK, you can get the FastA Report from there as well.) Make a copy of this OTC sequence in FastA format (you may want to save it in another program, e.g. your favorite word processor, for convenience).

Now click on the Blast button at the top of the page, and at the next page select Basic Blast search. At the Blast server, select blastp from the pull-down menu next to Program (this aligns protein sequences; the default blastn aligns nucleotide sequences), and paste

the *E. coli*OTCsequence in FastA format into the input window. Note that the pull-down menu gives you the option of entering the accession number (40962) instead of the sequence. The default sequence databases are nr, the non-redundant compilation of databases from the US, Europe and Japan. We'll use that, but note that a pull-down menu allows you to select other databases.

a) Click on the Submit Query button. When the job finally runs (this can take a minute or more when the Server is busy) what do you see?

b) Is the *E. coli* OTC protein related to any human protein? Scroll down the table of hits, past many bacterial OTCs (*Neisseria*, *Pyrococcus*...) until you run into some mammalian hits. With a score of 172, you should find a hyperlink to sp|P00480|OTC_HUMAN ORNITHINE CARBAMOYLTRANSFERASE PRECURSOR. Click on this hyperlink.

4.8 The entry for human OTC (P00480, which is the same as 400687) is quite long.

a) What occupies much of the feature table? What does this tell you about the *OTC*gene in humans?

b) Using either the features table for the GenBank entry 400687 (or P00480) or better yet, go back to the home page for NCBI and click on the OMIM button to go to the On-Line Medelian Inheritance In Man (from Victor McKusick, M.D.). Where is the gene? What happens in OTC deficiency?

4.9 What do the aligned amino acid sequences of the bacterial and human proteins tell you? Do conserved regions correlate with functional regions? For instance, does mutation of any amino acids in the conserved regions lead to a phenotype in humans?

Since the Blast search generated so many hits with higher scores than the *E. coli*- human pair, we will have to use a different tool to see the alignment. At the Blast server top page (where you selected Basic Blast search before), select Blast 2 sequences. This utility allows you to enter any two sequences and generate a pairwise alignment by the program Blast2. You should use the human and *E. coli* OTC protein sequences or their accession numbers, and be sure to choose blastp as the program. When doing this in July of 1998, I ran into a problem with the utility making a duplicate of each sequence I entered (I don't know if that was a problem at my end or theirs); this is likely a temporary condition. If you encounter a problem, try a different Server, such as the Sequence Analysis Server at genome.cs.mtu.edu/sas.html. Choose Pairwise Sequence Alignment, enter your sequences and run GAP or SIM on protein sequences.

Chromatin

4.10 One of the important early pieces of evidence that helped define the structure of the nucleosome was the pattern of nuclease cleavage in chromatin. In this experiment, chromatin was treated briefly with an enzyme, micrococcal nuclease, that degrades DNA, then all protein was removed and the and the purified DNA resolved by electrophoresis. A regular pattern of broad bands was seen; the average sizes of the DNA fragments were multiples of 200 bp, i.e. 200, 400, 600, 800 bp, etc. What does this result tell you about chromatin structure? The bands of DNA bands were thick and spread out rather than sharp; what does this tell you about the positions of cleavage by micrococcal nuclease?

4.11 Which histones are in the core of the nucleosome? What are the protein-protein interactions in the core? What protein domains mediate these interactions?

4.12 The mammalian virus SV40 has minichromosomes in which the circular duplex DNA is packaged into nucleosomes. When histones are removed from the minichromosomes, the resulting DNA is found to be negatively supercoiled. What does this tell you about the state of the DNA in the minichrosomes and the path of the DNA around the nucleosome?

4.13 Are the following statements true or false?

1. The DNA coils around the histones about 1.65 turns per nucleosomal core.
2. The DNA in chromatin containing actively transcribed genes is usually more sensitive to DNases than is the DNA in nontranscribed chromatin.

4.14 The packing ratio of a nucleic acid-protein complex is the ratio between the length of the naked DNA in normal B form to the length of the protein-DNA structure. For instance, if a set of proteins folded a DNA molecule of 100 Å into a structure that is 25 Å long, this structure has a packing ratio of 4.

a) Given the dimensions of the nucleosome structure, what is the packing ratio for the DNA in the nucleosome core? Note that the pitch is the distance between the midpoints of the DNA duplex as it turns around the histones in the core.

b) If the nucleosomes are tight-packed into a solenoid with 6 nucleosomes per turn, what is the packing ratio now? Assume that each turn of the solenoid translates 110 Å, i.e. the distance between the midpoints of nucleosomes in successive turns of the

solenoid is 110 Å.

4.15 How close are the edges of the DNA as it curves around the surface of the nucleosomal core?

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

4.4: Genome Analysis by Large Scale Sequencing

Whole genomes can be sequenced both by random shot-gun sequencing and by a directed approach using mapped clones.

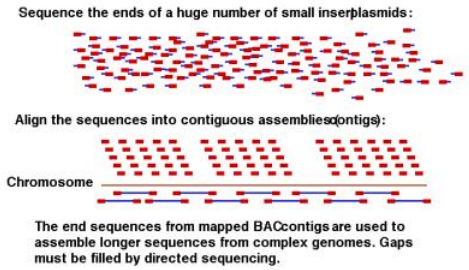


Figure 4.11. Directed sequencing of BAC contigs.

The results of the Celera and public collaboration on the fly sequence was published in early 2000, and descriptions of the human genome sequence were published separately by Celera and IHGSC in 2001. Neither genome is completely sequenced (as of 2001), but both are highly sequenced and are stimulating a major revolution in the life sciences.

The wisdom of which approach to take is still a matter of debate, and depends to some extent on how thoroughly one needs to sequence a complex genome. For instance, a publicly accessible sequence of the mouse genome at 3X coverage was recently generated by the shotgun approach. Other genomes will likely be “lightly sequenced” at a similar coverage. But a full, high quality sequence of mouse will likely use aspects of the more directed approach. Also, the Celera assembly (primarily shotgun sequence) used the public data on the human genome sequence as well. Thus current efforts use both the rapid sequencing by shotgun methods and as well as sequencing mapped clones.

Survey of sequenced genomes

The genome sequences are available for many species now, covering an impressive phylogenetic range. This includes more than 28 eubacteria, at least 6 archaea, a fungus (the yeast *Saccharomyces cerevisiae)*, a protozoan (*Plasmodium falciparum*), a worm (the nematode *Caenorhabditis elegans*), an insect (the fruitfly *Drosophila melanogaster*), two plants (*Arabadopsis*and rice (soon)), and two mammals (human *Homo sapiens*and mouse *Mus domesticus*). Some information about these is listed in Table 4.4.

Table 4.4.Sequenced genomes. This table is derived from the listing of “Complete Genomes Mapped on the KEGG Pathways (Kyoto Encyclopedia of Genes and Genomes)” at

www.genome.ad.jp/kegg/java/org_list.html

Additional genomes have been added, but only samples of the bacterial sequences are listed.

Genes encoding

Species	Genome Size (bp)	Protein	RNA	Total Enzymes	Category
Eubacteria					
Escherichia coli	4,639,221	4,289	108	1,254	gram negative
Haemophilus influenzae	1,830,135	1,717	74	571	gram negative
Helicobacter pylori	1,667,867	1,566	43	394	gram negative
Bacillus subtilis	4,214,814	4,100	121	819	gram positive
Mycoplasma genitalium	580,073	467	36	202	gram positive
Mycoplasma pneumoniae	816,394	677	33	226	gram positive
Mycobacterium tuberculosis	4,411,529	3,918	48	-	gram positive
Aquifex aeolicus	1,551,335	1,522	50	-	hyperthermophilic bacterium
Borrelia burgdorferi	1,230,663	1,256	23	176	lyme disease Spirochete

Synechocystis sp.	3,573,470	3,166	49	702	cyanobacterium
Archaebacteria					
Archaeoglobus fulgidus	2,178,400	2,407	49	439	S-metabolizing archaea
Methanococcus jannaschii	1,739,934	1,735	43	441	archaea
Methanobacterium thermoautotrophicum	1,751,377	1,871	47	558	archaea
Eukaryotes					
Saccharomyces cerevisiae	12,069,313	6,064	262	861	fungi
Caenorhabditis elegans	97,000,000	18,424		-	nematode
Drosophila melanogaster	180,000,000	13,601			insect, fly, 120 Mb sequenced
Arabidopsis thaliana	115,500,000	25,706			plant, complete
Homo sapiens	3,200,000,000	30,000-40,000			human, draft + finished
Mus domesticus	3,000,000,000				mouse, draft

Genome size

Bacterial genomes range in size from 0.58 to almost 5 million bp (Mb). *E. coli and B. subtilis*, two of the most intensively studied bacteria, have the largest genomes and largest numbers of genes. The genome of the yeast *Saccharomyces cerevisiae* is only 2.6 times as large as that of *E. coli*. The genome of humans is almost 700 times larger than that of *E. coli*. However, genome size is not a direct measure of genetic content over long phylogenetic distances. One needs to examine the fraction of the genome that codes for protein or contains other important information. Let's look at sizes and numbers of genes in different genomes.

Gene size and number

The average gene size is similar among bacteria, averaging around 1100 bp. Very little DNA separates most bacterial genes; in *E. coli*there is an average of only 118 bp between genes. Since the gene size varies little, then the number of genes varies over as wide a range as the genome size, from 467 genes in *M. genitalium*to 4289 in *E. coli*. Thus within bacteria, which have little noncoding DNA, the number of genes is proportional to the genome size.

Saccharomyces cerevisiaehas one gene every 1900 bp on average, which could reflect both an increase in size of gene as well as somewhat greater distance between genes. Both bacteria and yeast show a much denser packing of genes than is seen in more complex genomes.

Data on a large sample of human genes shows that they are much larger than bacterial genes, with the median being about 14 times larger than the 1 kb bacterial genes. This is not because most human proteins are substantially larger; both bacterial proteins average about 350 amino acids in length, which is similar to the median size of human proteins. The major difference is the large amount of intronic sequence in human genes.

Table 4.5.Average size of human genes and parts of genes. This is based on information in the IHGSC paper in Nature, and derived from analysis of 1804 human genes.

	Median	Mean
Internal exon	122 bp	145 bp
Number of exons	7	8.8
Length of each intron	1023 bp	3365 bp
3' UTR	400 bp	770 bp
5' UTR	240 bp	300 bp
Coding sequence	1100 bp	1340 bp
Length of protein encoded	367 amino acids	447 amino acids
Genomic extent	14,000 bp	27,000 bp

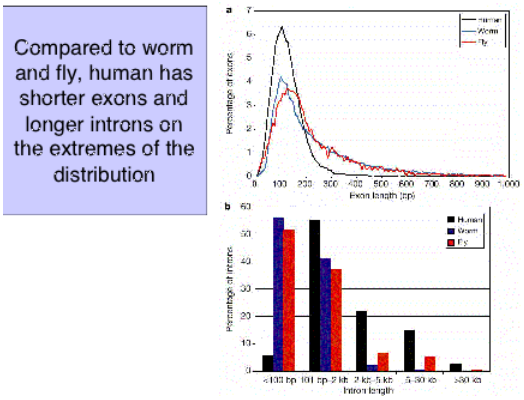


Figure 4.14. Genome size and number of genes in species ranging from bacteria to humans.

Alternative splicing is common in human genes

A previous lower estimate is that alternative splicing occurs in 35% of human genes. However, recent data show this fraction is larger.

For Chromosome 22:

- 642 transcripts cover 245 genes, 2.6 txpts/gene
- 2 or more transcripts for 145 (59%) of genes

For Chromosome 19:

- 1859 transcripts cover 544 genes, 3.2 txpts/gene

This contrasts with the situation in worm, in which alternative splicing occurs in 22% of genes. The increased genetic diversity from alternative splicing may contribute considerably to the greater complexity of humans, not just the increase in the number of genes.

Estimates of number of human genes

The estimated number of human genes has varied greatly over recent years. Some of these numbers have been widely quoted, and it may be useful to list some of the sources of these estimates.

- mRNA complexity (association kinetics): 40,000 genes
- Avg size of gene 30,000 bp: 100,000 genes
- Number of CpG islands: 70,000 to 80,000
- Unigene clusters of ESTs: 35,000 to 125,000
- More rigorous EST clustering: 35,000 genes
- Comparison to pufferfish: 30,000 genes
- Extrapolate from gene counts on chromosomes 21 and 22 (which are finished): 30,000 to 35,500 genes

Using the draft human sequence from Juy 2000, the IHGSC constructed an Initial Gene Index for human. They use the Ensembl system at the Sanger Centre. They started with ab initio predictions by Genscan, then confirmed by similarity to proteins, mRNAs, ESTs, and protein motifs (Pfam database) from any organism. This led to an initial set of 35,500 genes and 44,860 transcripts in the Ensemble database. After reducing fragmentation, merging with known genes, and removing contaminating bacterial sequences, they were left with 31,778 genes. After taking into account residual fragmentation, and the rate at which true genes are found by a similar analysis, the estimate remains about 32,000 genes. However, it is an estimate and is subject to change as more annotation is completed..

Starting with this estimate that the human genome contains about 32,000 genes, one can calculate how much of the genome is coding and how much is transcribed. If the average coding length is 1400 bp, then **1.5%**of human genome consists of coding sequence. If the average genomic extent per gene is 30 kb, then **33%** of human genome is “transcribed”.

Summary of number of genes in eukaryotic species:

- Human: 32,000 “still uncertain”
- Fly: 13, 338

- Worm: 18,266
- Yeast: 6,144
- Mustard weed: 25,706
- Human: 2x number of genes in fly and worm
- Human: more alternative splicing, perhaps 5x number of proteins as in fly or worm

Assignment of functions to genes

Genes encoding proteins and RNAs can be detected with considerable accuracy using compuational tools. Note that even for an extensively studies organism like *E. coli*, the number of genes found by sequence analysis (4289 encoding proteins) is far greater than the number that can be assigned as encoding a particular enzyme (1254). The discrepancy between genes found in the sequence versus those with known function (i.e. assigned as encoding an enzyme) is greater for some poorly characterized organisms such as the lyme-disease causing Spirochete *Borrelia burgdorferi*.

The many genes with unassigned function present an exciting challenge both in bioinformatics and in biochemistry/cell biology/genetics. Large collaborations have been initiated for a comprehensive genetic and expression analysis of some organisms. For instance, projects are underway to make mutations in all detected genes in *Saccharomyces cerevisiae* and to quantify the level of stable RNA from each gene in a variety of growth conditions, through the cell cycle and in other conditions. Databases are already established that record the changes in RNA levels for all yeast genes when the organism is shifted from glucose to galactose as a carbon source. These large scale expression analysis use high density microchip arrays that contain characteristic sequences for all 6064 yeast genes. These gene arrays are then hybridized with fluorescently labeled RNA or cDNA from cells grown under the two different conditions. The hybridization signals are quantitated and compared automatically, analyzed. The plan is to store the results in public databases. Useful websites include:

- SGD
- MIPS: a database for genomes and protein sequences

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

4.5: Sizes of genomes - The C-value paradox

The C-value is the amount of DNA in the haploid genome of an organism. It varies over a very wide range, with a general increase in C-value with complexity of organism from prokaryotes to invertebrates, vertebrates, plants. The **C-value paradox** is basically this: how can we account for the amount of DNA in terms of known function? Very similar organisms can show a large difference in C-values (e.g. amphibians). The amount of genomic DNA in complex eukaryotes is much greater than the amount needed to encode proteins. For example: Mammals have 30,000 to 50,000 genes, but their genome size (or C-value) is 3 x 10^9 bp.

$$\frac{3 \times 10^9 \text{ base pairs}}{3000 \text{ base paires (average gene size)}} = 1 \times 10^6 \text{ (“gene capacity”).} \tag{4.5.1}$$

Drosophila melanogaster has about 5000 mutable loci (~genes). If the average size of an insect gene is 2000 bp, then

$$\frac{1 \times 10^8 \text{base pairs}}{2 \times 10^3 \text{ base pairs}} => 50,000 \text{ “gene capacity”.} \tag{4.5.2}$$

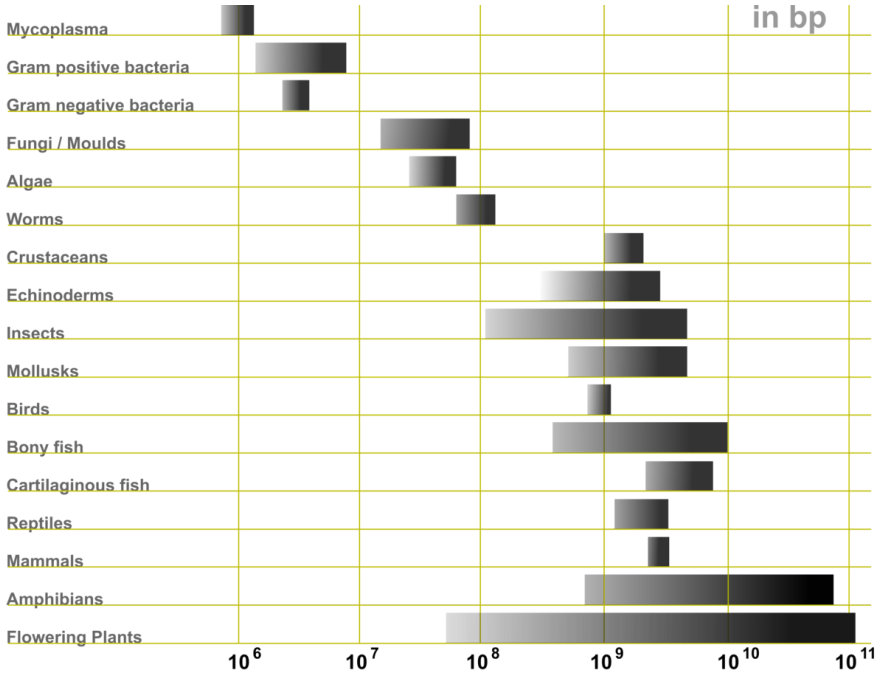


Figure 4.5.1: Genome size ranges (in base pairs) of various life forms. (CC BY-SA 4.0; Abizar).

Our current understanding of complex genomes reveals several factors that help explain the classic C-value paradox:

- Introns in genes
- Regulatory elements of genes
- Pseudogenes
- Multiple copies of genes
- Intergenic sequences
- Repetitive DNA

The facts that some of the genomic DNA from complex organisms is highly repetitive, and that some proteins are encoded by families of genes whereas others are encoded by single genes, mean that the genome can be considered to have several distinctive components. Analysis of the kinetics of DNA reassociation, largely in the 1970's, showed that such genomes have components that can be distinguished by their repetition frequency. The experimental basis for this will be reviewed in the first several sections of this chapter, along with application of hybridization kinetics to measurement of complexity and abundance of mRNAs. Advances in genomic sequencing have provided more detailed views of genome structure, and some of this information will be reviewed in the latter sections of this chapter.

Table 4.5.1: Distinct components in complex genomes

R= repetition frequency

R= repetition frequency		
Highly repeated DNA	*R* (repetition frequency) ≥ 100,000	Almost no information, low complexity
Moderately repeated DNA	10<<em>R<10,000>	Little information, moderate complexity
“Single copy” DNA	*R*=1 or 2	Much information, high complexity

LTR-containing retrotransposons

- MaLR: mammalian, LTR retrotransposons
- Endogenous retroviruses
- MER4 (MEdium Reiterated repeat, family 4)

Repeats that resemble DNA transposons

MER1 and MER2

Mariner repeats

Some of the repeats are clustered into tandem arrays and make up distinctive features of chromosomes (Figure 4.5.1). In addtion to the interspersed repeats discussed above, another contributor to the moderately repetitive DNA fraction are the thousands of copies of rRNA genes. These are in extensive tandem arrays on a few chromosomes, and are condensed into heterochromatin. Other chromosomal structures with extensive arrays of tandem repeats are centromeres and telomeres.

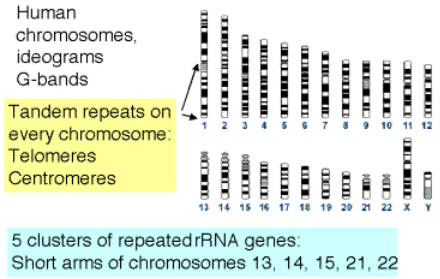


Figure 4.5.2: Clustered repeated sequences in the human genome.

The common way of finding repeats now is by sequence comparison to a database of repetitive DNA sequences, RepBase (from J. Jurka). One of the best tools for finding matches to these repaats is RepeatMasker (from Arian Smit and P. Green, U. Wash.). A web server for RepeatMasker can be accessed at: ftp.genome.washington.edu/cgi-bin/RepeatMasker

Exercise 4.5.2

Try Repeat Masker on *INS*gene sequence. You can get the *INS*sequence either from NCBI (GenBank accession gi|307071|gb|L15440.1 or one can use LocusLink, query on) or from the course website.

Very little of the nonrepetive DNA component is expressed as mRNA

Hybridization kinetic studies of RNA revealed several important insights. First, saturation experiments, in which an excess of unlabeled RNA was used to drive labeled, nonrepetitive DNA (tracer) into hybrid, showed that only a small fraction of the nonrepetitive DNA was present in mRNA. Classic experiments from Eric Davidson’s lab showed that only 2.70% of total nonrepetitive DNA correspondss to **mRNA** isolated from sea urchin gastrula (this is corrected for the fact that only one strand of DNA is copied into RNA; the actual amount driven into hybrid is half this, or 1.35%; Figure 4.8). The complexity of this nonrepetitive fraction is (*N*sc) is 6.1 x 108 bp, so only 1.64 x 107 bp of this DNA is present as mRNA in the cell. If an "average" mRNA is 2000 bases long, there are ~8200 mRNAs present in gastrula.

In contrast, if the nonrepetitive DNA is hybridized to **nuclear** RNA from the same tissue, 28% of the nonrepetitive fraction corresponds to RNA (Figure 4.8). The nuclear RNA is heterogeneous in size, and is sometimes referred to as heterogeneous nuclear RNA, or hnRNA. Some of it is quite large, much more so than most of the mRNA associated with ribosomes in the cytoplasm. The latter is called polysomal mRNA.

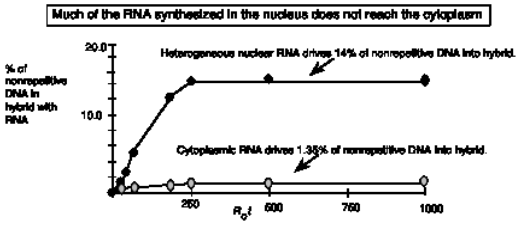


Figure 4.5.3

These data show that a substantial fraction of the genome (over one-fourth of the nonrepetitive fraction) is transcribed in nuclei at the gastrula stage, but much of this RNA never gets out of nucleus (or more formally, many more sequences from the DNA are represented in nuclear RNA than in cytoplasmic RNA). Thus much of the complexity in nuclear RNA stays in the nucleus; it is not processed into mRNA and is never translated into proteins.

Factors contributing to an explanation include

1. Genes may be transcribed but the RNA is not stable. (Even the cytoplasmic mRNA from different genes can show different stabilities; this is one level of regulation of expression. But there could also be genes whose transcripts are so unstable in some tissues that they are never processed into cytoplasmic mRNA, and thus never translated. In this latter case, the gene is transcribed but not expressed into protein.)
2. Intronic RNA is transcribed and turns over rapidly after splicing.
3. Genes are transcribed well past the poly A addition site. These transcripts through the 3' flanking, intergenic regions are usually very unstable.
4. Not all of this "extra" RNA in the nucleus is unstable. For instance, some RNAs are used in the nucleus, e.g.:
5. U2-Un RNAs in splicing (small nuclear RNAs, or snRNAs).

RNA may be a structural component of nuclear scaffold (S. Penman)

Thus, although 10 times as much RNA complexity is present in the nucleus compared to the cytoplasm, this does not mean that 10 times as many genes are being transcribed as are being translated. Some fraction (unknown presently) of this "excess" nuclear RNA may represent genes that are being transcribed but not expressed, but many other factors also contribute to this phenomenon.

mRNA populations in different tissues show considerable overlap:

- Housekeeping genes encode metabolic functions found in almost all cells.
- Specialized genes, or tissue-specific genes, are expressed in only 1 (or a small number of) tissues. These tissue-specific genes are sometimes expressed in large amounts.

Estimating numbers of genes expressed and mRNA abundance from the kinetics of RNA-driven reactions

Using principles similar to those for analysis of repetition classes in genomic DNA, one can determine from the kinetics of hybridization between a preparation of RNA and single copy DNA both the average number of genes represented in the RNA, as well as the abundance of the mRNAs. The details of the kinetic analysis will not be presented, but they are similar to those already discussed. Highly abundant RNAs (like high copy number DNA) will hybridize to genomic DNA faster than will low abundance RNA (like low copy number DNA). Only a few mRNAs are highly abundant, and they constitute a low complexity fraction. The bulk of the genes are represented by lower abundance mRNA, and these many mRNAs constitute a high complexity, slowly hybridizing fraction.

An example is summarized in Table 4.5.2. an excess of mRNA from chick oviduct wash ybridized to a tracer of labeled cDNA (prepared from oviduct mRNA). Three principle components were found, ranging from the highly abundant ovalbumin mRNA to much rarer mRNAs from many genes.

Table 4.5.2

Component	Kinetics of hybridization	N (nt)	# mRNAs	Abundance	Example
1	fast	2,000	1	120,000	Ovalbumin
2	medium	15,000	7-8	4,800	Ovomucoid, others
3	slow	2.6 x 107	13,000	6-7	Everything else

Preparation of normalized cDNA libraries for ESTs

Just like the mRNA populations used as the templates for reverse transcriptase, the cDNAs from a particular tissue or cell type will be composed of many copies of a very few, abundant mRNAs, a fairly large number of copies of the moderately abundant mRNAs, and a small number of copies of the rare mRNAs. Since most genes produce low abundance mRNA, a corresponding small number of cDNAs will be made from most genes. In an effort to obtain cDNAs from most genes, investigators have normalized the cDNA libraries to remove the most abundant mRNAs.

The cDNAs are hybridized to the template mRNA to a sufficiently high *Rot* (concentration of RNA ´ time) so that the moderately abundant mRNAs and cDNAs are in duplex, whereas the rare cDNAs are still single-stranded. The duplex mRNA-cDNA will stick to a hydroxyapatite column, and the desired single-stranded, low abundance cDNA will elute. This procedure can be repeated a few times to improve the separation. The low abundance, high complexity cDNA is then ligated into a cloning vector to construct the cDNA library.

This normalization is key to the success of a random sequencing approach. **Random cDNA clones**, hundreds of thousands of them, have been picked and **sequenced**. A single-pass sequence from one of these cDNA clones is called an **expressed sequence tag**, or **EST** (Figure 4.9). It is called a “tag” because it is a sequence of only part of the cDNA, and since it is in cDNA, which is derived from mRNA, it is from an expressed gene. If the cDNA libraries reflected the normal abundance of the mRNAs, then this approach would result in re-sequencing the abundant cDNAs over and over, and most of the rare cDNAs would never be sequenced. However, the normalization has been successful, and many genes, even with rare mRNAs, are represented in the EST database.

As of May, 2001, over 2,700,000 ESTs individual sequences of human cDNA clones have been deposited in dbEST. They are grouped into nonredundant sets (called Unigene clusters). Over 95,000 Unigene clusters have been assembled, and almost 20,000 of them contain known human genes. The estimated number of human genes is less than the number of Unigene clusters, presumably because some large genes are still represented in more than one Unigene cluster. It is likely that most human genes are represented in the EST databases. Exceptions include genes expressed only in tissues which have not been sampled in the cDNA libraries. For more information, see **www.ncbi.nlm.nih.gov/UniGene/index.html**

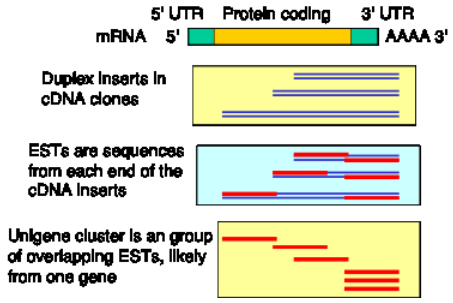


Figure 4.5.4. cDNA clones from normalized libraries are sequenced to generate ESTs.

Databases for genomic analysis

NCBI: http://www.ncbi.nlm.nih.gov

- Nucleic acid sequences
- genomic and mRNA, including ESTs
- Protein sequences
- Protein structures
- Genetic and physical maps

Organism-specific databases

- MedLine (PubMed)
- Online Mendelian Inheritance in Man (OMIM)

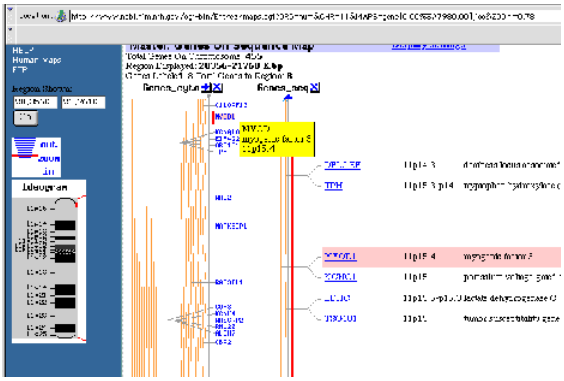


Figure 4.15. Example of mapping information at NCBI. Genetic map around MYOD1, 11p15.4

Sequences and annotation of the human genome

- Human Genome Browser
- http://genome.ucsc.edu/goldenPath/hgTracks.html

Ensemble (European Bioinformatics Institute (EMBL) and Sanger Centre)

http://www.ensembl.org/

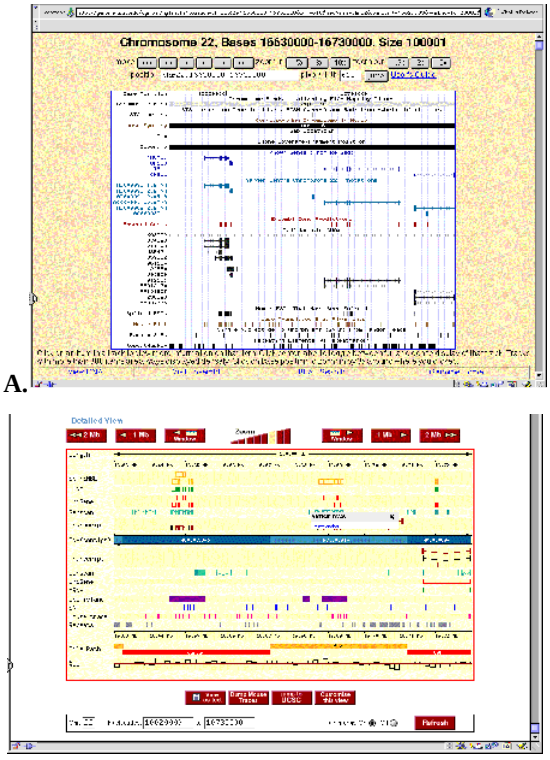


Figure 4.16. Sample views from servers displaying the human genome. (A) View from the Human Genome Browser. The region shown is part of chromosome 22 with the genes *PNUTL1*, *TBX1*and others. Extensive annotation for exons, repeats, single nucleotide polymorphisms, homologous regions in mouse and other information is available for all the sequenced genome. (B) Comparable information in a different format is available at the ENSEMBL server.

Programs for sequence analysis

- BLAST to search rapidly through sequence databases
- PipMaker (to align 2 genomic DNA sequences)
- Gene finding by ab initio methods (GenScan, GRAIL, etc.)
- RepeatMasker

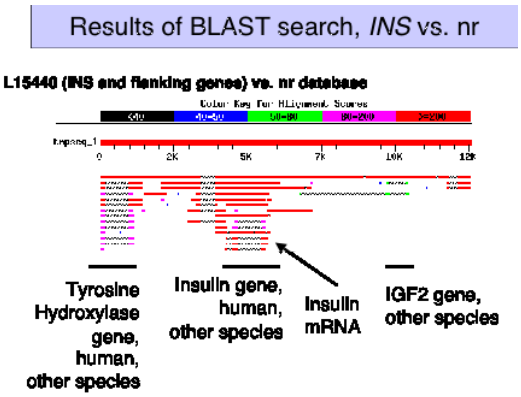


Figure 4.18. Results of BLAST search, *INS*vs. nr

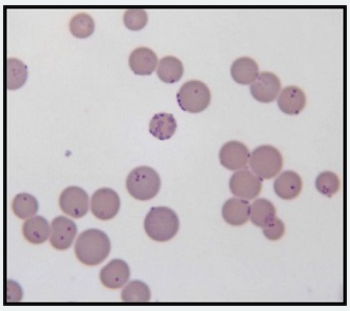
Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

4.6: Large Scale Genome Organization

How to get by with the smallest possible genome

The *Mycoplasma* species have the smallest genomes of any free-living species. They are most related to the *Bacillaceae* family, but have lost their cell walls and many other functions in a process of reductive evolution. They are obligate parasites, e.g. living in the lungs of humans. Their genomes encode many transport proteins, so that amino acids, sugars, etc. can be taken up from their hosts. They have very little metabolic capacity, utilizing only glycolysis in the case of *M. genitalium*. There is very little biosynthetic capacity, depending largely on uptake from the host for these nutrients.



Mycoplasma Haemofelis, Wright-Geiemsa Staining 100X. (CC BY-SA 3.0; Nr387241).

One might have thought that the Mycoplasmal species would retain only the most highly conserved genes in bacteria, under the premise that these are the most critical genes. However, they have retained a proportion of conserved and variable genes that is quite similar to the proportion seen in *E. coli.* This indicates that these bacteria are maintaining a balance between conserved and variable genes that perhaps reflects an equilibrium between the stability of major physiological processes and the need for environmental adaptability.

More information from *E. coli*

The complete sequence of the *E. coli* genome provides an overview of genome structure within a well-understood context. For more information, see Blattner et al. (1997) Science, vol. 277, pp. 1453- 1462.

(1) Organization with respect to direction of replication

Since replication proceeds bidirectionally from the origin (*oriC*) and ends at the terminus, one can divide the genome into two "replicores." The replication fork proceeds clockwise in Replicore 1 and counter-clockwise in Replicore 2 (Figure 4.19).

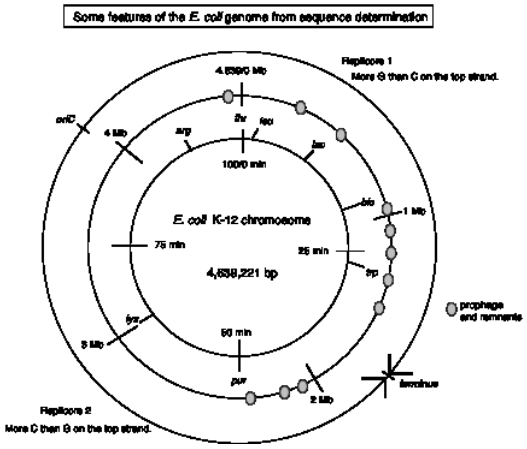


Figure 4.19.

Several features of the genome are oriented with respect to replication. All the rRNA genes, 53 of 86 tRNA genes, and 55% of the protein coding genes are transcribed in the same direction as the replication fork moves. In other species, such as the *Mycoplasma*, the transcriptional polarity is even more pronounced, and it also corresponds to the direction of replication.

These replicores show a pronounced skew in base composition, such that an excess of G over C is seen on the top strand (i.e. the one presented in the sequence file) in Replicore 1 and the opposite in Replicore 2. This nucleotide bias is striking and unexpected. As will be appreciated more after we study DNA synthesis in Part Two, this means that the leading strand for both replication forks is richer in G than C. Such an nucleotide bias may reflect differential mutation in the leading and lagging strands as a result of the asymmetry inherent in the DNA replication mechanism.

The recombination hotspot chi (GCTGGTGG) also shows a prominent strand preference, being more abundant on the leading strand of each replicore. The role of chi sites in recombination is covered in Chapter 8.

(2) Repeats, prophage and transposable elements

The *E. coli* chromosome contains several prophages and remnants of prophage, including lambda and three lambdoid prophages. The genome is peppered with at least 18 families of repeated DNA. The longest are the 5 *Rhs*elements, which are 5.7 to 9.6 kb in length. Others are as short as the 581 copies of the 40 bp palindromic REP repeat. Several families of insertion sequences, which are transposable elements, are found. Note that repetitive elements are common in bacteria as well as in eukaryotes.

(3) General categories of genes

Many of the genes are similar to other genes in *E. coli*. Homologous genes that have diverged because of gene duplications are **paralogous**. The genes that encode proteins of similar but not necessarily identical function are referred to as a paralogous family. About 1/3 of the *E. coli* genes (1345) have at least one paralogous sequence in the genome. Some paralogous groups are quite large, the largest being the ABC transporters with 80 members. The larger number of genes in *E. coli* could reflect some redundancy in function as well as greater diversification of function compared to other bacteria with fewer genes.

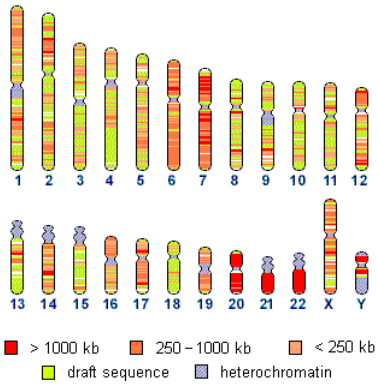
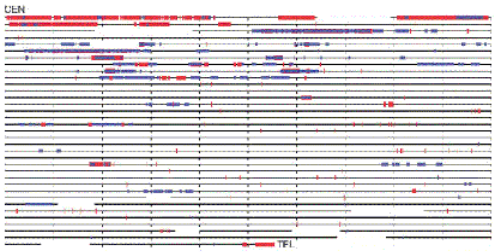


Figure 4.20. Human chromosomes, and the status of their sequencing.

Based on current understanding of the function of the gene products, about 1/4 are involved in small-molecule metabolism, about 1/8 are used in large-molecule metabolism, and at least 1/5 are associated with cell structure and processes. A specific function has not been assigned to the products of about 40% of the *E. coli* genes. Segmental duplications are common, as illustrated in Figure 4.21 for chromosomes 22.



The size and location of intrachromosomal (blue) and interchromosomal (red) duplications are depicted for chromosome 22q, using the PARASIGHT computer program (Bailey and Eichler, unpublished). Each horizontal line represents 1 Mb (ticks, 100-kb intervals). Pairwise alignments with > 90% nucleotide identity and > 1 kb long are shown.

Figure 4.21.Segmental duplications on chromosome 22.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

4.7: Comparative Genome Analysis

Paralogous Genes

- Genes that are similar because of descent from a common ancestor are **homologous.**
- Homologous genes that have diverged after speciation are **orthologous.**
- Homologous genes that have diverged after duplication are **paralogous.**

One can identify **paralogous groups** of genes encoding proteins of similar but not identical function in a species e.g., ABC transporters: 80 members in *E. coli*

Core proteomes vary little in size

Proteome: all the proteins encoded in a genome

To calculate the Core proteome:

Count each group of paralogous proteins only once

Number of distinct protein families in each organism

Species	Number of genes	Core proteome
Haemophilus	1709	1425
Yeast	6241	4383
Worm	18424	9453
Fly	13601	8065

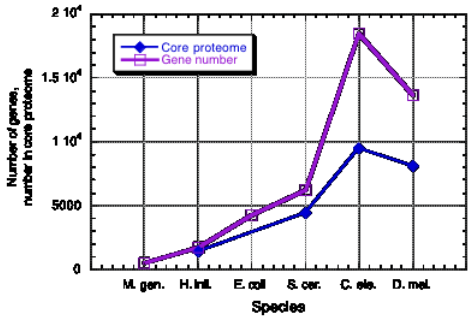


Figure 4.22.Little change in core proteome size in eukaryotes

Core proteomes are conserved

- Many of the proteins in the core proteomes are shared among eukaryotes
- 30% of fly genes have orthologs in worm
- 20% of fly genes have orthologs in both worm and yeast
- 50% of fly genes have likely orthologs in mammals

Function of proteins in flies (and worms and yeast) provides strong indicators of function in humans. Flies have orthologs to 177 of the 289 human disease genes

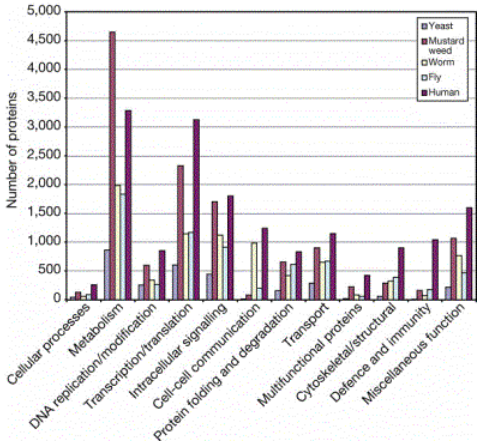


Figure 4.23. Functional categories in eukaryotic proteomes

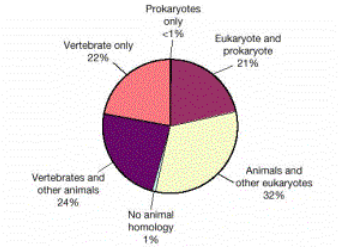


Figure 4.24. Distribution of the homologues of the predicted human proteins

Conserved Segments in the Human and Mouse Genomes

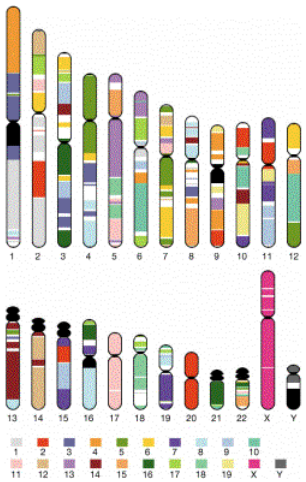


Figure 4.25. Regions of human chromosomes homologous to regions of mouse chromosomes (indicated by the colors). For example, virtually all of human chromosome 20 is homologous to a region on mouse chromosome 2, and almost all of human chromosome 17 is homologous to a region on mouse chromosome 11. More commonly, segments of a given human chromosomes are homologous to different mouse chromosomes. Chromsosomes from mouse have more rearrangements relative to humans than do chromosomes from many mammals, but the homologous relationships are still readily apparent.

CHROMOSOMES and CHROMATIN

Chromosomes are the cytological package for genes. Genomes are much longer than the cellular compartment they occupy compartment dimensions length of DNA

- Phage T4:

$$0.065 \times 0.10\, mm 55\, mm = 170\, kb \tag{4.7.1}$$

- E. coli:

$$1.7 \times 0.65\, mm\, 1.3\, mm = 4.6 \times 10^3\, kb \tag{4.7.2}$$

- Nucleus (human):

$$6mm\ diam.\ 1.8\ m = 6 \times 10^6\ kb \tag{4.7.3}$$

Definition: Packing ratio

$$\text{Packing ratio} = \frac{\text{length of DNA}}{\text{length of the unit that contains it}}. \tag{4.7.4}$$

The smallest human chromosome contains about

$$46 \times 10^6\ bp = 14,000\ mm = 1.4\ cm\ DNA. \tag{4.7.5}$$

When condensed for mitosis, this chromosome is about. 2 mm long. The packing ratio is therefore about 7000!

Loops, matrix and the chromosome scaffold

When DNA is released from *mitotic*chromosomes by removing most of the proteins, long loops of DNA are seen, emanating from a central scaffold that resembles the remnants of the chromosome.

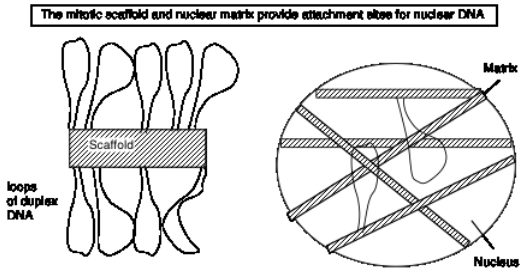


Figure 4.26: EM analysis of intact nuclei shows network of fibers called a matrix.

Biochemical preparations using salt and detergent to remove proteins and nuclease to remove most of the DNA leaves a "matrix" or "scaffold" preparation. Similar DNA sequences are found in these preparations; these sequences are called **matrix attachment regions** = MARs (or scaffold attachment regions = SARs). They tend to be A+T rich and have sites for cleavage by topoisomerase II. Topoisomerase II is one of the major components of the matrix preparation; but the composition of the matrix is still in need of further study.

Since it is attached at the base to the matrix, each loop is a separate topological domain and can accumulate supercoils of DNA.

From the measured sizes of loops, and calculations based on the amount of nicking required to relax DNA within the loops, we estimate that the average size of these loops is about 100 kb (85 kb based on nicking frequency for relaxation).

Some evidence suggests that replication and possibly some transcriptional control may be exerted at the bases of the loops.

Interphase chromatin and mitotic chromosomes

During interphase, i.e. between mitotic divisions, the highly condensed mitotic **chromosomes**spread out through the nucleus to form **chromatin**. Interphase chromatin is not very densely packed in most of the nucleus (euchromatin). In some regions it is very densely packed, comparable to a mitotic chromosome (heterochromatin).

Both interphase chromatin and mitotic chromosomes are made of a 30 nm fiber. The mitotic chromosome is much more coiled than interphase chromosomes.

Most transcription occurs in euchromatin.

- Constitutive heterochromatin = nonexpressed regions that are condensed (compact) in all cells (e.g. centromeric simple repeats)
- Facultative heterochromatin = inactive in only some cell lineages, active in others.

One example of heterochromatin is the inactive X chromosome in female mammals. The choice of which X chrosomosome to inactivate is random in various cell lineages, leading to a mosaic phenotypes for some X-linked traits. For instance, one genetic determinant of coat color in cats is X-linked, and the patchy coloration on calico cats results from this random inactivation of one of the X chromosomes, leading to the lack of expression of this determinant in some but not all hair cells.

Cytologically visible bands in chromosomes

G bands and R bands in mammalian mitotic chromosomes (Figure 4.27)

Giemsa-dark (G) bands tend to be A+T rich, with a large number of L1 repeats.

Giemsa-light bands tend to be more G+C rich, with very few L1 repeats and many Alu repeats.

(R bands are about the same as Giemsa-light bands. They are visualized by a different preparative procedure so that the "reverse" of the Giemsa-stained images are seen.)

T bands are adjacent to telomeres, do not stain with Giemsa, and are extremely G+C rich, with lots of genes and myriad Alu repeats.

The functional significance of these bands is still under active investigation.

One can **localize**a gene to a particular region of a chromosome by *in situ*hybridization with a radioactive or, now more commonly, fluorescent probe for the gene. The region of hybridization is determined by simultaneously viewing the stained banding pattern and the hybridization pattern. Many spreads of mitotic chromosomes are viewed and scored, and the gene is localized to the chromosomal region with a significantly greater incidence of hybridization signal than that seen to the rest of the chromosomes.

Another common method of mapping the location of genes is by hybridization to DNA isolated from a panel of somatic cell hybrids, each hybrid cell carrying a small subset of, e.g., human chromosomes on a hamster background. Some hybrid cells carry broken human chromosomes, which allows even more precise localization (see Figure 1.8.2, "J-1 series").

Polytene chromosomes are visible in several Drosophilatissues

These contain many copies of the chromosomes, side by side in register. Thus most chromosomal regions are highly **amplified** in these tissues. Chromosomal stains reveal characteristic banding pattern, which is the basis for the cytological map. The cytological map (of polytene bands) combined with the genetic map gives a **cytogenetic map**, which is a wonderful guide to the Drosophila genome. One can localize a gene to a particular region by in situ hybridization (in fact the technique was invented using *Drosophila*polytene chromoomes.

Multiple genes per band on mammalian chromosomes

Figure 4.27 gives a view of human chromosome 11 at several different levels of resolution. The region 11p15 has many genes of interest, including genes whose products regulate cell growh (*HRAS*), determination and differentiation of muscle cells (*MYOD*), carbohydrate metabolism (*INS*), and mineral metabolism (*PTH*). The b-globin gene (*HBB*) and its closely linked relatives are also in this region. A higher resolution view of 11p15, based on a compilation of genetic and physical mapping (Cytogenetics and Cell Genetics, 1995) is shown next to the classic ideogram (banding pattern). This is in a scale of millions of base pairs, and one can start to get a feel for gene density in this region. Interestingly, it varies quite a lot, with the gene-dense sub-bands near the telomeres; these may correspond to the T-bands discussed above. Other genes appear to be more widely separated. For instance, each of the b-like globin genes is separated by about 5 to 8 kb from each other (see the map of the YAC, or yeast artificial chromosome, carrying the b-like globin genes), and this gene cluster is about 1000 kb (i.e. 1 Mb) from the nearest genes on the map. However, further mapping will likely find many other genes in this region. Now even more information is available at the web sites mentioned earlier.

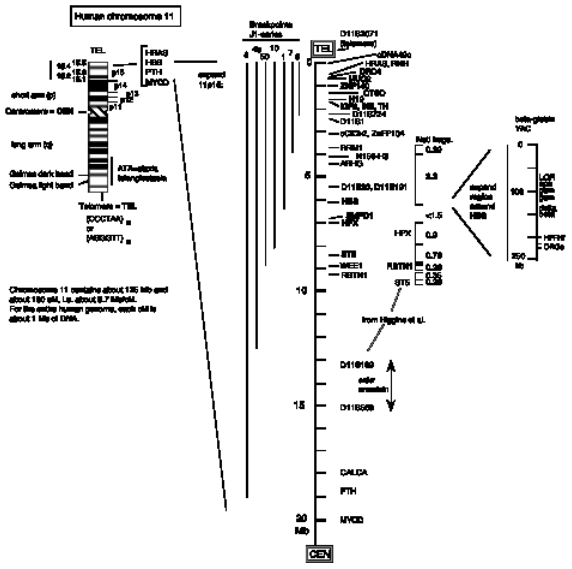


Figure 4.27.

The relationship between recombination distances and physical distances varies substantially among organisms. In human, one centiMorgan (or cM) corresponds to roughly 1 Mb, whereas in yeast 1 cM corresponds to about 2 kb, and this value varies at least 10-fold along the different yeast chromosomes. This is a result of the different frequencies of recombination along the chromosomes.

Specialized regions of chromosomes

Centromere: region responsible for segregation of chromosomes at mitosis and meiosis. The centromere is a constricted region (usually) toward the center of the chromosome (although it can be located at the end, as with mouse chromosomes.) It contains a kinetochore, a fibrous region to which microtubules attach as they pull the chromosome to one pole of the dividing cell. DNA sequences in this region are highly repeated simple sequences (in *Drosophila*, the unit of the repeat is about 25 bp long, repeated hundreds of times). Specific proteins are at the centromere, and are now intensely investigated.

Telomere: forms the ends of the linear DNA molecule that makes up the chromosome. The telomeres are composed of thousands of repeats of CCCTAA in human. Variants of this sequence are found in the telomeres in other species. Telomeres are formed by **telomerase**; this enzyme catalyzed the synthesis of more ends at each round of replication to stabilize linear molecules.

The Principal Proteins in Chromatin are Histones

Composition of chromatin: Various biochemical methods are avialable to isolated chromatin from nuclei. Chemical analysis of chromatin reveals proteins and DNA, with the most abundant proteins being the **histones**. A complex set of less abundant histones are referred to as the nonhistone chromosomal proteins.

The histones and DNA present in equal masses.

Mass Ratio DNA: histones: nonhistone proteins: RNA = 1: 1: 1: 0.1

Histones are small, basic (positively charged), highly conserved proteins. They bind to each other to form specific complexes, around which DNA wraps to form **nucleosomes**. The nucleosomes are the fundamental repeating unit of chromatin.

There are **5 histones, 4 in the core of the nucleosome and one outside the core**.

H3, H4: Arg rich, most conserved sequence ü

ý CORE Histones

H2A, H2B: Slightly Lys rich, fairly conservedþ

H1: very Lys rich, most variable in sequence between species.

X-ray diffraction studies of histone complexes and the nucleosome core have provided detailed insight into how histones interact with each other and with DNA in this fundamental entity of chromatin structure.

Key reference: "Crystal structure of the nucleosome core particle at 2.8 Å resolution" by Luger, K. Mader, A., Richmond, R.K., Sargent, D.F. & Richmond, T.J. in **Nature** 389: 251-260 (1997)

Histone Interactions via the Histone fold

The core histones have a highly positively charged amino-terminal tail, and most of the rest of the protein forms an a-helical domain. Each core histone has at least 3 a-helices.

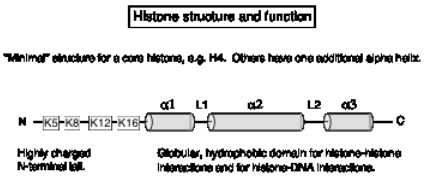


Figure 4.28

The a-helical domain forms a characteristic **histone fold**, in which shorter a1 and a3 helices are perpendicular to the longer a2 helix. The a-helices are separated by two loops, L1 and L2. The histone fold is the dimerization domain between pairs of histones, mediating the formation of crescent-shaped heterodimers H3-H4 and H2A-H2B. The histone-fold motifs of the partners in a pair are antiparallel, so that the L1 loop of one is adjacent to the L2 loop of the other.

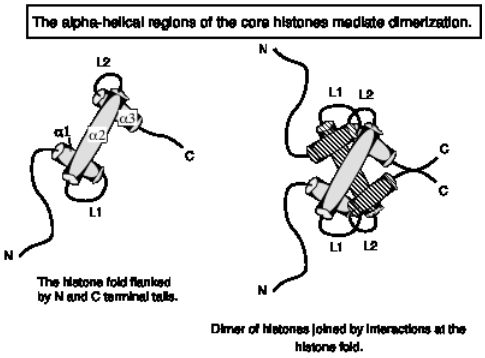


Figure 4.29

A structure very similar to the histone fold has now been seen in other nuclear proteins, such as some subunits of TFIID, a key component in the general transcription machinery of eukaryotes. It also serves as a dimerization domain for these proteins.

Two H3-H4 heterodimers bind together to form a tetramer.

Nucleosomes are the Subunits of the Chromatin Fiber

The most extended chromatin fiber is about 10 nm in diameter. It is composed of a series of histone-DNA complexes called *nucleosomes*.

Principal lines of evidence for this conclusion are:

1. Observations of this 10 nm fiber in the electron microscope showed a series of bodies that looked like beads on a string. We now recognize the beads as the nucleosomal cores and the string as the linker between them.
2. Digestion of DNA in chromatin or nuclei with micrococcal nuclease releases a series of products that contain DNA of discrete lengths. When the DNA from the products of micrococcal nuclease digestion was run on an agarose gel, the it was found to be a series of fragments of 200 bp, 400 bp, 600 bp, 800 bp, etc. , i.e. integral multiples of 200 bp. This showed that cleavage by this nuclease, which has very little sequence specificity, was restricted to discrete regions in chromatin. Those regions of cleavage are the linkers.
3. Physical studies, including both both neutron diffraction and electron diffraction data on fibers and most recently X-ray diffraction of crystals, have provided more detailed structural information.

2. The **nucleosomal core** is composed of an octamer of histones with 146 bp of duplex DNA wrapped around it in 1.65 very tight turns. The octamer of histones is actually a tetramer H32H42 at the central axis, flanked by two H2A-H2B dimers (one at each end of the core.

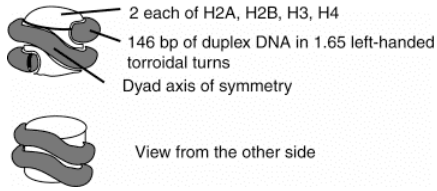


Figure 4.30. Schematic views of the nucleosomal core

The 10 nm fiber is composed of a string of nucleosomal cores joined by linker DNA. The length of the linker DNA varies among tissues within an organism and between species, but a common value is about 60 bp. The **nucleosome** is the **core plus the linker**, and thus contains about 200 bp of DNA.

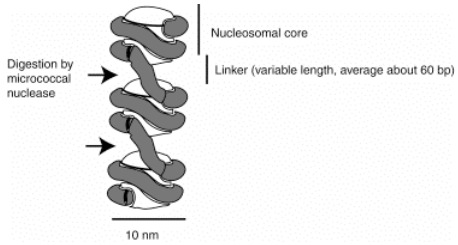


Figure 4.31. A string of nucleosomes

Detailed structure of the nucleosomal core.

Path of the DNA and tight packing

The 146 bp of DNA is wrapped around the histone octamer in 1.65 turns of a flat, left-handed torroidal superhelix. Thus 14 turns or "twists" of the DNA are in the 1.65 superhelical turns, presenting 14 major and 14 minor grooves to the histone octamer. Pancreatic DNase I will cleave DNA on the surface of the core about every 10 bp, when each twist of the DNA is exposed on the surface.

The DNA superhelix has an average radius of 41.8 Å and a pitch of 23.9 Å. This is a very tight wrapping of the DNA around the histones in the core - note that the duplex DNA on one turn is only a few Å from the DNA on the next turn! The DNA is not uniformly bent in this superhelix. As the DNA wraps around the histones, the major and then minor grooves are compressed, but not in a uniform manner for all twists of the DNA. G+C rich DNA favors the major groove compression, whereas A+T rich DNA favors the minor groove compression. This is an important feature in translational positioning of nucleosomes and could also affect the affinity of different DNAs for histones in nucleosomes.

The DNA phosphates have high mobility when not contacting histones; the DNA phosphates facing the solvent are much more mobile than is seen with other protein-DNA complexes.

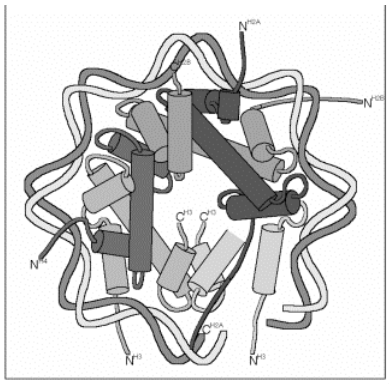


Figure 4.32. A cross-sectional view of the nucleosome core showing histone heterodimers and contacts with DNA. This images corresponds to the proteins and DNA in about one half of the nucleosome.

The left-handed torroidal supercoils of DNA in nucleosomal cores is the equivalent of a right-handed, hence negative, supercoil. Thus the DNA in nucleosomes is effectively underwound.

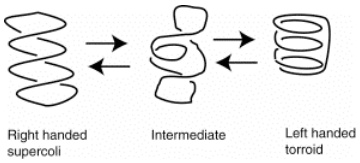


Figure 4.33.

Histones in the nucleosome core particle

The protein octamer is composed of four dimers (2 H2A-H2B pairs and 2 H3-H4 pairs) that interact through the "histone fold". The two H3-H4 pairs interact through a 4-helix bundle formed between the two H3 proteins to make the H32H42 tetramer. Each H2A-H2B pair interacts with the H32H42 tetramer through a second 4-helix bundle between H2B and H4 histone folds.

The histone-fold regions of the H32H42 tetramer bind to the center of of the DNA covering a total of about 6 twists of the DNA, or 3 twists of DNA per H3-H4 dimer. Those of the H2A-H2B dimers cover a comparable amount of DNA, 3 twists per dimer. Additional helical regions extend from the histone fold regions and are an integral part of the the core protein within the confines of the DNA superhelix.

Histone-DNA interactions in the core particle.

The histone-fold domain of the heterodimers (H3-H4 and H2A-H2B) bind 2.5 turns of DNA double helix, generating a 140˚ bend. The interaction with DNA occurs at two types of sites:

1. The L1 plus L2 loops at the narrowly tapered ends of each heterodimer form a similar DNA binding site for each histone pair. The L1-L2 loops interact with DNA at each end of the 2.5 turns of DNA.
2. The a1 helices of each partner in a pair form the convex surface in the center of the DNA binding site. The principal interactions are H-bonds between amino acids and the **phosphate** backbone of the DNA (there is little sequence specificity to histone-DNA binding). However, there are some exceptions, such a hydrophobic contact between H3Leu65 and the 5-methyl in thymine. An Arg side chain from a histone fold enters the minor groove at 10 of the 14 times it faces the histone octamer. The other 4 occurrences have Arg side chains from tail regions penetrating the minor groove.

Histone Tails

The histone N- and C-termial tails make up about 28% of the mass of the core histone proteins, and are seen over about 1/3 of their total length in the electron density map - i.e. that much of their length is relatively immobile in the structure.

The tails of H3 and H2B pass through channels in the DNA superhelix created by 2 juxtaposed minor grooves. One H4 tail segment makes a strong **interparticle** connection, perhaps relevant to the higher-order structure of nucleosomes.

The most N-terminal regions of the histone tails are not highly ordered in the X-ray crystal structure. These regions extend out from the nucleosome core and hence could be involved in **interparticle** interactions. *The sites for acetylation and de-acetylation of specific lysines are in these segments of the tails that protrude from the core*. Post-translational modifications such as acetylation have been implicated in "chromatin remodeling" to allow or aid transcription factor binding. It seems likely that these modifications are affecting interactions between nucleosomal cores, but not changing the structure of the core particle.

Outside Links

- Some excellent **resources are available on the World Wide Web**for visualizing and further investigating chromatin structure and its involvment in nuclear processes.
- Dmitry Pruss maintains a site with many good images, including dynamic, step-by-step view of the nuclesomal core beginning with the histone fold domains and ending with a complete core, with DNA. **www.average.org/~pruss/nucleosome.html**
- Another good site is from J.R. Bone: **rampages.onramp.net/~jrbone/chrom.html**

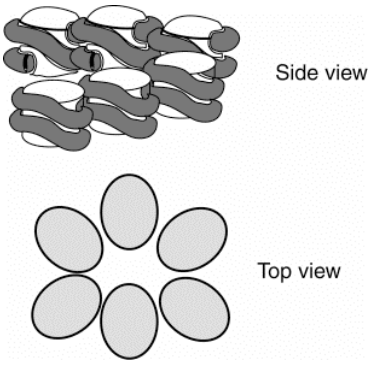
Higher order chromatin structure

1. The 10 nm fiber composed of nucleosomal cores and spacers is folded into higher order structures for much of the DNA in chromatin. In fact, the 10 nm fiber with the beads-on-a-string appearance in the electron microscope was prepared at very low salt concentrations and is free of histone H1.
2. **In the presence of H1 and at more physiological salt concentrations, chromatin forms a 30 nm fiber.** The exact structure of this fiber remains a point of considerable debate, and one cannot rule the possibility of multiple structure in this fiber.

3. One reaonable model is that the 10 nm fiber coils around itself to generate a **solenoid that is 30 nm in diameter, with 6 nucleosomes per turn of of the solenoid**.

Histone H1 binds to the outer surface of the nucleosomal core, interacting at the points of DNA entry and exit. H1 molecules can be cross-linked to each other with chemical reagents, indicating that the H1 proteins also interact with each other. Interactions between H1 proteins, each bound to a nucleosomal core, may be one of the forces driving the formation of the 30 nm fiber.

Figure 4.34. Model for one turn of the solenoid in the 30 nm fiber.



4. Each level of chromatin structure produces a more compact arrangment of the DNA. This can be described in terms of a packing ratio, which is the length of the DNA in an extended state divided by the length of the DNA in the more compact state.

For the 10 nm fiber, the packing ratio is about 7, i.e. there are 7mm of DNA per mm of chromatin fiber. The packing ratio in the core is higher (see problems), but this does not include the additional, less compacted DNA in the spacer. In the 30 nm fiber, the packing ratio is about 40, i.e. there 40mm DNA per mm of chromatin fiber.

5. The 30 nm fiber is probably the basic constituent of both interphase chromatin and mitotic chromosomes. It can be compacted further by additional coils and loops. One of the key issues in gene regulation is the nature of the chromating fiber in transcriptionally acative euchromatin. Is it the 10 nm fiber? the 30 nm fiber? some modification of the latter? or even some higher order structure? These are topics for current research.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

4.E: Genomes and Chromosomes (Exercises)

4.3

(BPA) Answer the following questions with reference to the figure below.

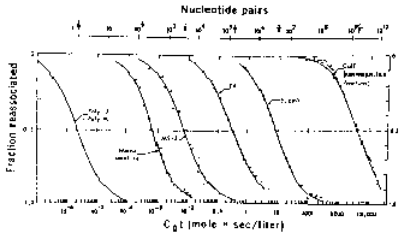


Figure for 4.3 Reassociation of nucleic acids, sheared to 500-nucleotide fragments, from various sources [Derived fom R. J. Britten and D. Kohne, *Science*, **161,**529 (1968).]

a. How many of these DNA preparations contain more than one frequency class of sequences? Explain your answer.
b. If the genome size of *E. coli*is taken to be 4.5 x 106 nucleotide pairs, what is the genome size of T4?
c. What is the complexity of mouse satellite DNA?
d. Mouse satellite DNA represents 10% of the mouse genome. What is the repetition number for mouse satellite sequences, given that the haploid genome size is 3.2 x 109 nucleotide pairs?
e. The calf genome is the same size as the mouse genome. What fraction of the calf genome is composed of unique sequences?

4.4

Let's imagine that you obtained a DNA sample from an armadillo and measured the kinetics of renaturation of the genomic DNA. A standard of bacterial DNA (N= 3 x 106 bp) was also renatured under identical conditions. Three kinetic components were seen in the armadillo DNA $C0t$ curve, renaturing fast, medium or slow. The fraction of the genome occupied by each component (f) and the $C0t$value for half-renaturation ($Cot1/2$(measured)) are as follows:

Component	f	*Cot*1/2(measured)
fast	0.2	10^{-4}
medium	0.4	10^{-1}
slow	0.4	10^{4}

a. Use the information provided to calculate the $Cot1/2$(pure), the complexity (N), and the repetition frequency (R) for each component. Assume that the slowly renaturing component is single copy.
b. Calculate the genome size (G) of the armadillo under the assumption that the slowly renaturing component is single copy.
c. Which of the following sequences could be a member of the fast renaturing component?

GACTCAGACTCAGACTCA

ATATATATATATATATAT

ACTGCCACGGGATACTGC

GCGCGC

4.S: Genomes and Chromosomes (Summary)

Additional Readings

1. Britten RJ, Kohne DE. (1968) Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science 161:529-540
2. Wetmur and Davidson (1968)The rate constant for renaturation is inversely proportional to sequence complexity. J. Molecular Biology 34:349-370.
3. Davidson EH, Hough BR, Amenson CS, Britten RJ. (1973) General interspersion of repetitive with non-repetitive sequence elements in the DNA of Xenopus. J. Molecular Biology 77:1-23.
4. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 269:496-512
5. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M,
6. Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M,
7. Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC. (2000) The genome sequence of Drosophila melanogaster. Science 287:2185-2195
8. International Human Genome Sequencing Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. Nature **409**: 860-921.
9. Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., Cherry, J. M., Henikoff, S., Skupski, M. P., Misra, S., Ashburner, M., Birney, E., Boguski, M. S., Brody, T., Brokstein, P., Celniker, S. E., Chervitz, S. A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R. F., Gelbart, W. M., George, R. A., Goldstein, L. S., Gong, F., Guan, P., Harris, N. L., Hay, B. A., Hoskins, R. A., Li, J., Li, Z., Hynes, R. O., Jones, S. J., Kuehl, P. M., Lemaitre, B., Littleton, J. T., Morrison, D. K., Mungall, C., O'Farrell, P. H., Pickeral, O. K., Shue, C., Vosshall, L. B., Zhang, J., Zhao, Q., Zheng, X. H., Zhong, F., Zhong, W., Gibbs, R., Venter, J. C., Adams, M. D. and Lewis, S. (2000). Comparative genomics of the eukaryotes. Science 287: 2204-15.
10. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C. (2001). The sequence of the human genome. Science 291: 1304-1351.

11. The Arabidopsis Genome Initiative (2000) Sequence of the Arabidopsis thaliana genome. Nature 408:796-815.

SECTION OVERVIEW

Unit II: Replication, Maintenance and Alteration of the Genetic Material

Unit 2 covers the structures of nucleic acids (DNA and RNA) and methods for analyzing them biochemically. In addition, this chapter explores some of the insights into gene structure and function, especially in eukaryotes, that the use of these techniques has provided. This includes the separation of mRNA-coding regions into exons, production of multiple proteins from a single gene by differential splicing of the exons in RNA, and the duplication of genes to form gene families with both active and inactive copies.

5. DNA replication I: Enzymes and mechanism

6. DNA replication II: Start, stop and control

7: Mutation and Repair of DNA

8: Recombination of DNA

9. Transposition of DNA

Thumbnail: DNA Polymerase I: Klenow Fragment (PDB 1KLN EBI). (Public Domain; Jawahar Swaminathan and MSD staff at the European Bioinformatics Institute).

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

CHAPTER OVERVIEW

5. DNA replication I: Enzymes and mechanism

A fundamental property of living organisms is their ability to reproduce. Bacteria and fungi can divide to produce daughter cells that are identical to the parental cells. Sexually reproducing organisms produce offspring that are similar to themselves. On a cellular level, this reproduction occurs by mitosis, the process by which a single parental cell divides to produce two identical daughter cells. In the germ line of sexually reproducing organisms, a parental cell with a diploid genome produces four germ cells with a haploid genome via a specialized process called meiosis. In both of these processes, the genetic material must be duplicated prior to cell division so that the daughter cells receive a full complement of the genetic information. Thus accurate and complete replication of the DNA is essential to the ability of a cell organism to reproduce.

In this chapter and the next, we will examine the process of replication. After describing the basic mechanism of DNA replication, we discuss the various techniques researchers have used to achieve a more complete understanding of replication. Indeed, a theme of this chapter is the combination of genetic and biochemical approaches that has allowed us to uncover the mechanism and physiology of DNA replication. In the remaining sections of the chapter, we focus on the enzymes that mediate DNA replication. In these descriptions, you will encounter several cases of structure suggesting a particular function. We will point out parallels and homologies between bacterial and eukaryotic replication components. This chapter covers the basic process and enzymology of DNA synthesis, and the next chapter will cover regulation of DNA replication.

Topic hierarchy

Further readings

- A. Kornberg and T. Baker (1992) **DNA Replication, 2nd Edition**, W.H. Freeman and Company, New York.
- A. Kornberg, I. R. Lerman, M. J. Bessman, and E. S. Simms (1956) "Enzymic synthesis of deoxyribonucleic acid" Biochimica et Biophysica Acta **21**:197-198.
- M. Meselson and F. W. Stahl (1958) "The replication of DNA in *Escherichia coli*." Proceedings of the National Academy of Sciences, USA **44**:671-682.
- R. Okazaki, T. Okazaki, K. Sakabe, K. Sugimoto, and A. Sugino (1968) "Mechanism of DNA Chain Growth, I. Possible Discontinuity and Unusual Secondary Structure of Newly Synthesized Chains" Proceedings of the National Academy of Sciences, USA **59**: 598-605.
- P. De Lucia and J. Cairns (1969) Isolation of an *E.* colistrain with a mutation affecting DNA polymerase. Nature **224**:1164-1166.
- J. Gross and M. Gross (1969) Genetic analysis of an *E. coli*strain with a mutation affecting DNA polymerase. Nature **224**:1166-1168
- R. Sousa (1996) Trends in Biochemical Sciences **21**:186-190. Similarities in structure among DNA polymerases
- Herendeen and Kelly (1996) Cell **84**:5-8. Subunits and mechanism of DNA polymerase III.

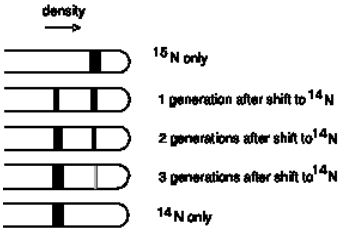
Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

5.E: DNA replication I: Enzymes and Mechanism (Exercises)

Question 5.8

Imagine you are investigating the replication of a bacterial species called *B. mulligan*. The bacteria is grown for several generations in medium containing a heavy density label, [^{15}N] NH_4Cl. The bacteria are then shifted to medium containing normal density [^{14}N] NH_4Cl. DNA is extracted after each generation and analyzed on a CsCl gradient. From the results shown below, what is the mode of replication in *B. mulligan*? Explain your conclusion.



Question 5.9

How many turns must be unwound during replication of the *E. coli* chromosome? The chromosome contains 4.64 x 106 base pairs.

Question 5.10

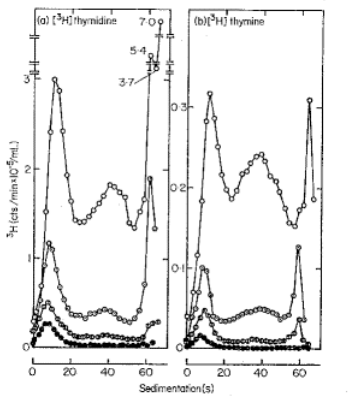
Which of the following comments about Okazaki fragments are true or false? Okazaki fragments:

a. are short segments of newly synthesized DNA.
b. are formed by synthesis on the leading strand of DNA.
c. have a short stretch of RNA, or a mixture of ribonucleotides and deoxyribonucleotides, at their 5' end.
d. account for overall synthesis of one DNA strand in a 3' to 5' direction.

Question 5.11.

The following experimental results are from A. Sugino and R. Okazaki (1972) "Mechanisms of DNA Chain Growth VII. Direction and rate of growth of T4 nascent short DNA chains" J. Mol. Biol. 64: 61-85.

a. *E. coli*cells were infected with bacteriophage T4 and then chilled to 4^{o}C to slow the rate of replication. Replicating DNA in the infected cells was pulse-labeled with [3H]-thymidine (a) or [^{3}H]-thymine (b) for 5 sec (black-filled circles), 30 sec (open circles with vertical line), 60 sec (open circles with dot) or 300 sec (open circles). The pulse labeling was stopped with potassium cyanide and ice, and the DNA was extracted, denatured in NaOH, and separated on an alkaline sucrose gradient. Fractions from the gradient were collected and assayed for the amount of ^{3}H in the DNA (as material that bound to a filter after washing in (a) and as acid-insoluble material in (b)). The sedimentation value in Svedbergs (S) is given along the x-axis; faster sedimenting material is toward the right. What do these data tell you about the sizes of nascent (newly synthesized) DNA at the various pulse labeling times?



(b) Sugino and Okazaki used a method to break the isolated short nascent chains (completed Okazaki fragments) randomly and recover only the oligonucleotides from the 5Ã ends. They found that at very short labeling times (e.g. 5 sec) the [^{3}H] thymidine was not at the 5' ends of the DNA (hence it was internal and at the 3' ends). After longer labeling times, the [^{3}H] thymidine was

found in the oligonucleotides at the 5' end. What do you conclude is the direction of chain growth of the nascent chains? Explain your logic.

Question 5.12

We have covered two experiments from the Okazaki lab using pulse labeling for increasing times to follow the synthesis of new DNA. How would you design a pulse-chase experiment to monitor not only the initial production of Okazaki fragments, but also their incorporation into larger DNA molecules?

Question 5.13

Which enzymes, substrates, and cofactors are used in common and which ones are distinctive for synthesis of leading strands and lagging strands of DNA at the replication fork of *E. coli*?

Question 5.14

Which subunit or complex within *E. coli* DNA polymerase III holoenzyme has each the following functions?

a. Catalyzes 5'to 3' polymerization of new DNA.
b. Has the proofreading function (3' to 5' exonuclease).
c. Dimerizes the two catalytic cores.
d. Forms the clamp that is thought to account for its high processivity.
e. Loads and unloads the sliding clamp.

Question 5.15

What are the components of the multiprotein complex known as the primosome in *E. coli*? What does it do? In what direction does it travel?

Question 5.16

Which eukaryotic nuclear DNA polymerase(s) is (are) thought to account for leading strand and lagging strand synthesis?

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Basic Mechanisms of Replication

DNA Replication is Semiconservative

We begin our investigation by describing the basic model for how nucleotides are joined in a specific order during DNA replication. By the early 1950's, it was clear that DNA was a linear string of deoxyribonucleotides. At that point, one could postulate three different ways to replicate the DNA of a cell. First, a cell might have a DNA-synthesizing "machine" which could be programmed to make a particular string of nucleotides for each chromosome. A second possibility is that the process of replication could break the parental DNA into pieces and use them to seed synthesis of new DNA.

A third model could be proposed from the DNA structure deduced by Watson and Crick. When they described the double-helical structure of DNA in a one-page article in Nature in 1953, they included this brief statement of a third model:

> "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."

A subsequent paper elaborated on this mechanism. The complementarity between base pairs (A with T and G with C) not only holds the two strands of the double helix together, but the sequence of one strand is sufficient to determine the sequence of the other. Hence a third possibility for a mechanism of DNA replication was clear - one parental strand could serve as atemplate directing synthesis of a complementary strand in the daughter DNA molecules. This 1953 paper is of course most famous for its description of the double-helical structure of DNA held together by base complementarity, but it is also important because the proposed structure suggested a testable model for how a particular process occurs, in this case replication.

These three models make different predictions about the behavior of the two strands of the parental DNA during replication (Figure 5.1). In the first, programmed machine model, the two strands of the parental DNA can remain together, because they are not needed to determine the sequence of the daughter strands. This model of replication is called **conservative**: the parental DNA molecules are the same in the progeny as in the parent cell. In the second model, the each strand of the daughter DNA molecules would be a combination of old and new DNA. This type of replication is referred to as **random** (or dispersive). The third model, in which one strand of the parental DNA serves as a template directing the order of nucleotides on the new DNA strand, is a **semiconservative** mode of replication, because half of each parent duplex (i.e. one strand) remains intact in the daughter molecules.

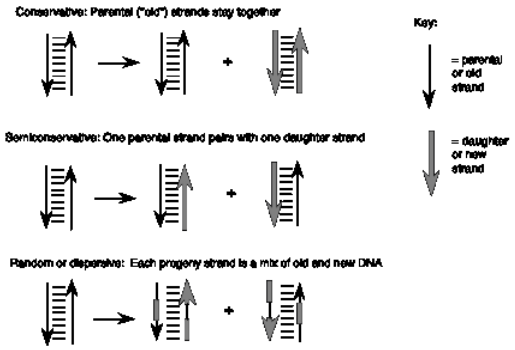
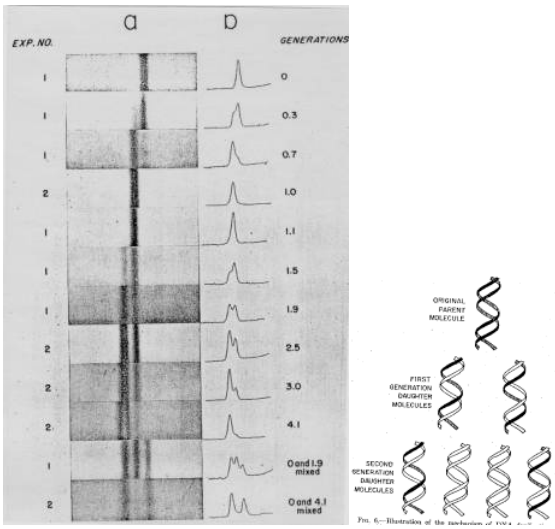


Figure 5.1.Possible models of replication of a duplex nucleic acid.

When they were graduate students at the California Institute of Technology, Matthew Meselson and Franklin Stahl realized that they could test these three models for replication by distinguishing experimentally between old and new strands of DNA. They labeled the old or parental DNA with nucleotides composed of a heavy isotope of nitrogen (^{15}N) by growing *E. coli*cells for several generations in media containing [^{15}N] NH4Cl. Ammonia is a precursor in the biosynthesis of the purine and pyrimidine bases, and hence this procedure labeled the nitrogen in the nucleotide bases in the DNA of the *E. coli*cells with ^{15}N. The cells were then shifted to grow in media containing the highly abundant, light isotope of nitrogen, ^{14}N, in the NH4Cl, so that newly synthesized DNA would have a "light" density. The labeled, heavy (old) DNA could be separated from the unlabeled, light (new) DNA on a CsCl density gradient, in which the DNA bands at the position on the gradient where the concentration of CsCl has a density equal to that of the macromolecule. At progressive times after the shift to growth in [^{14}N] NH4Cl, samples of the cells were collected, then DNA was isolated from the cells and separated on a CsCl gradient.



A. B.

Figure 5.2. *Results of the Meselson and Stahl experiment demonstrating semiconservative replication of DNA.* ***A.*** *The left panel (a) shows ultraviolet absorption photographs of DNA after equilibrium sedimentation in a CsCl gradient, as a function of the number of generations from the shift from media that labeled DNA with a high density (^{15}N-labeled) to a medium in which the DNA is normal, or light density (^{14}N-DNA). The density of the CsCl gradient increases to the right. The panel on the right (b) shows a trace of the amount of DNA along the gradient. The number of generations since the shift to the media with ^{14}N substrates is shown at the far right. Mixing experiments at the bottom show the positions of uniformly light and heavy DNA (generations 0 and 4.1 mixed) and the mixture of those plus hybrid light and heavy DNA (generations 0 and 1.9 mixed). Parental DNA forms a band at the heavy density (^{15}N-labeled), whereas after one generation in light (^{14}N) media, all the DNA forms a band at a hybrid density (between heavy and light). Continued growth in light media leads to the synthesis of DNA that is only light density.* ***B.*** *The interpretation of the experimental results as demonstrating a semiconservative model of replication. Part A of this figure is Figure 4 and Part B is Figure 6 from M. Meselson and F. Stahl (1958) "The Replication of DNA in Escherichia coli" Proceedings of the National Academy of Sciences, USA* ***44****:671-682.*

The results fit the pattern expected for semiconservative replication (Figure 5.2). To quote from Meselson and Stahl, "until one generation time has elapsed, half-labeled molecules accumulate, while fully labeled DNA is depleted. One generation time after the addition of ^{14}N, these half-labeled or 'hybrid' molecules alone are observed. Subsequently, only half-labeled DNA and completely unlabeled DNA are found. When two generation times have elapsed after the addition of ^{14}N, half-labeled and unlabeled DNA are present in equal amounts." A conservative mode of replication is ruled out by the observation that all the DNA formed a band at a hybrid density after one generation in the [14N] NH4Cl-containing medium. However, it is consistent with either the semiconservative or random models. As expected for semiconservative replication, half of the DNA was at a hybrid density and half was at a light density after two generations in [14N] NH4Cl-containing medium. Further growth in the 14N medium resulted in an increase in the amount of DNA in the LL band.

Exercise

Question 5.1: What data from this experiment rule out a random mode of replication?

These experiments demonstrated that each parental DNA strand is used as a template directing synthesis of a new strand during DNA replication. The synthesis of new DNA is directed by base complementarity. The enzymes that carry out replication are not programmed "machines" with an inherent specificity to synthesize a given sequence, but rather the template strand of DNA determines the order of nucleotides along the newly synthesized DNA strand (Figure 5.3).

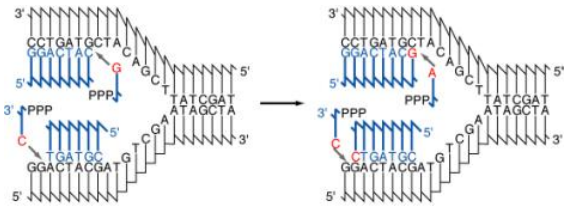


Figure 5.3. Diagram of the addition of nucleotides in a new strand of DNA during semiconservative replication. The parental DNA strands are shown in black and the new DNA strands and deoxyribonucleoside triphosphates are in blue. The DNA strands are shown using the convention that vertical lines are the deoxyribose portion of each deoxyribonucleotide, and the connecting lines represent the phosphodiester linking the 3' hydroxyl of one deoxyribonucleotide with the 5' hydroxyl of the next. The part of the connecting line representing the 3' end of the phosphodiester attached to the vertical (deoxyribose) line about 1/3 of the way along it, and the part of the connecting line representing the 5' end of the phosphodiester is attached at the end of the vertical line. Bases are abbreviated by a single letter. The bases on the deoxyribonucleotides that are being added are in red. Two rounds of addition of nucleotides are shown. In this diagram, each strand of the parental DNA is serving as the template for synthesis of a new DNA strand. The chemistry of the synthesis reaction, the enzymes needed for separating the two parental strands, and other features of replication will be discussed later in the chapter.

The association of a parental DNA strand with a newly synthesized DNA strand observed in this important experimental result is consistent with the use of each parental DNA strand as a template to direct the replication machinery to place nucleotides in a particular order. Watson and Crick proposed that base complementarity would guide the replication machinery to insert an A opposite a T, a T opposite an A, a G opposite a C, and C opposite a G (Figure 5.3). This was verified once the enzymes carrying out DNA synthesis were isolated, and the chemical composition of the products of replication was compared with that of the templates. These enzymes are discussed in detail later in the chapter, as will be the chemistry of the process of adding individual nucleotides to the growing DNA chain (a process called **elongation**). You may recall that these enzymes were also used to demonstrate the antiparallel arrangement of the DNA strands predicted by Watson and Crick (recall problem 2.5). With this understanding of how the sequence of nucleotides is specified, we can examine the types of DNA structures found during replication.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Biochemical and Genetic Identifcation of Enzymes

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

DNA Primers for Synthesis

Making primers for DNA synthesis

The enzyme **primase** catalyzes the synthesis of the primers from which DNA polymerases can begin synthesis (Figure 5.21). Primers are short oligonucleotides, ranging from 6 to 60 nucleotides long. They can be made of ribonucleotides or a mixture of deoxyribonucleotides and ribonucleotides. The principal primase in *E. coli* is the 60 kDa protein called **DnaG protein**, the product of the *dnaG* gene. The major primase in eukaryotic cells is **DNA polymerase α**.

The primase of *E. coli*, DnaG protein, cannot synthesize primers by itself, but rather it is part of much larger complex called the **primosome**. The primosome acts repeatedly during lagging strand synthesis, finding a primer-binding site on the SSB-coated single-stranded template strand and synthesizing a primer. Identification of the components of the primosome was aided by the convenient model system of *in vitro*synthesis of fX174 DNA. fX174 is a single-stranded bacteriophage; the DNA found in the virus is termed the plus strand. After infection of *E. coli*, this plus strand is converted to a double-stranded replicative form (Figure 5.24 A). The conversion of single-stranded phage DNA to duplex DNA occurs by the synthesis of several Okazaki fragments, and hence it is a good model for discontinuous synthesis on the lagging strand. This reaction can be carried out *in vitro*, which allowed the biochemical dissection of the various steps in primosome assembly and movement.

Table 5.3. Components of the primosome in E. coli

protein	gene	activities and functions
PriA	*priA*	helicase, 3' to 5' movement, site recognition
PriB	*priB*	
PriC	*priC*	
DnaT	*dnaT*	needed to add DnaB-DnaC complex to preprimosome
DnaC	*dnaC*	forms complex with DnaB
DnaB	*dnaB*	helicase, 5' to 3' movement. DNA dependent ATPase.
DnaG	*dnaG*	synthesize primer

Five different proteins are found in a **prepriming complex**, PriA, PriB, PriC, DnaT, and DnaB (Table 5.4). A sixth protein, DnaC, is needed for the assembly of this complex. In the case of fX174 viral DNA template coated with SSB, PriA (Figure 5.24 B) recognizes a primer assembly site. The proteins **PriB**and **PriC**are then added to form a complex. The hexameric protein **DnaB**is in a complex with six molecules of **DnaC**when it is not on the DNA. In an ATP-dependent process, and with help from **DnaT**, DnaB is transferred to the template and DnaC is released.

The prepriming complex is now ready for the **primase, DnaG**, to bind and make the active primosome. Although the role of each of the proteins in the primosome is not yet clear, information is available on some of the steps in primosome action. The preferred binding site on the template for primase is CTG, with the T being used as the template for the first nucleotide of the primers. A high affinity complex between DnaB and ATP forms or stabilizes a secondary structure in the single-stranded template DNA that is used by primase; this is thought to be how DnaB "activates" the primase to begin synthesis. After ATP hydrolysis by DnaB, the low affinity ADP-DnaB complex dissociates from the template. The primosome can now move to the next site for primer synthesis.

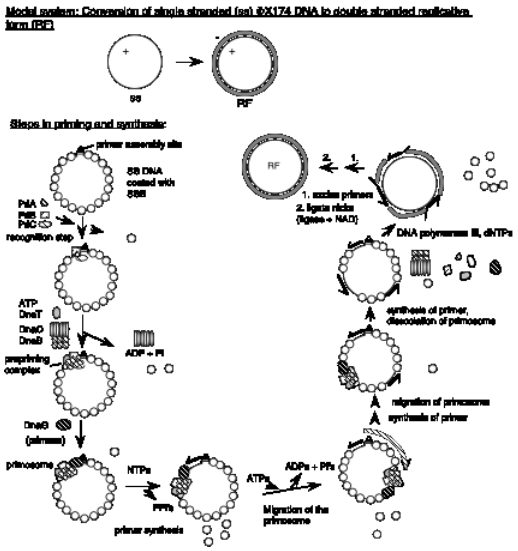


Figure 5.24. Assembly and migration of the primosome. After assembly of the prepriming complex, DnaG joins the complex to complete the primosome. After synthesis of the primer (dark black line), the primosome moves to the next site for synthesis. This tracking along the SSB-coated single stranded DNA requires ATP hydrolysis and causes dissociation of some of the SSB. In the diagram, the primosome is shown moving in a 5' to 3' direction along the template strand (clockwise), which is the opposite of the direction of primer synthesis. This would be the direction of movement catalyzed by DnaB. The primosome also contains PriA, which catalyzes movement along single-stranded DNA in the opposite direction. Once primers have been synthesized, DNA polymerase III can synthesize Okazaki fragments from them, the primers are excised and gaps repaired by DNA polymerase I, and then the fragments are ligated together.

The primosome contains two helicases than can move along single-stranded DNA with opposite polarity. PriA moves in a 3' to 5' direction, whereas DnaB moves in a 5' to 3' direction. When tested *in vitro*with a substrate similar to that shown in Figure 5.22, fragments from each end are displaced, indicating that the primosome moved in one direction on some molecules and in the other direction on others. Figure 5.24 B shows the migration as driven by the DnaB helicase, but movement can also occur in the other direction as well.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

DNA Synthesis is Semi-discontinuous

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

DNA topology during replication

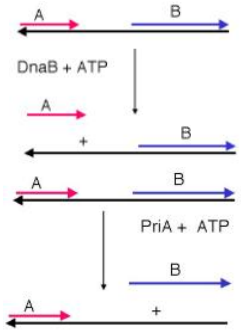


Figure 5.22: An assay for direction of helicase tracking along a single stranded region.

Figure 5.22B shows the results of the tracking assay for a helicase called PriA. In what direction does it track along the single-stranded DNA?

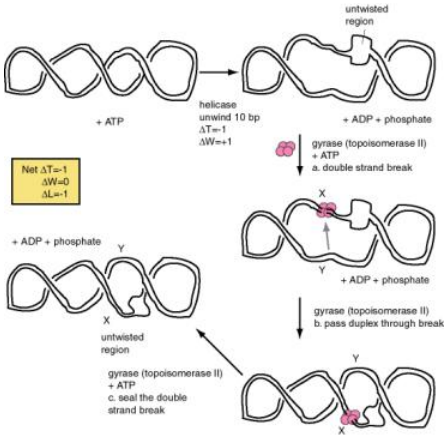


Figure 5.23. Changes in topology of the DNA during replication. The ATP-dependent untwisting (negative DT) catalyzed by DNA helicases causes a compensatory change in writhing (positive DW), which is relieved by the action of a topoisomerase II, such as DNA gyrase. Gyrase catalyzes an ATP-dependent, three-step reaction, cleaving the two strands of DNA, passing another part of the DNA duplex through the break and re-sealing the break. The action of gyrase generates a negative DW to balance the positive DW from the action of helicase. X and Y mark two different regions of the DNA molecule. A gray arrow indicates the direction of duplex movement through the break. Gyrase is a tetramer, and is shown as four pink balls.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Eukaryotic Replication Proteins

Eukaryotic replication proteins have functions analogous to those found in bacteria.

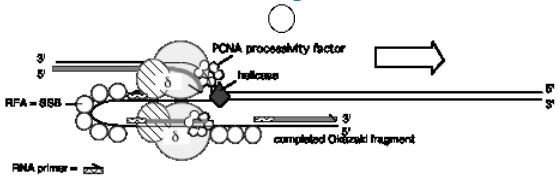


Figure 5.28. Similar structures of processivity factors for DNA replication. The mammalian protein, PCNA (top), is a trimer, each monomer of which has two similar domains. T he trimer forms a circle that surrounds DNA, hence serving as a sliding clamp. The b subunit of DNA polymerase III from E. coli is a dimer (bottom), each monomer of which has three similar domains. These domains have a very similar structure to those of PCNA, despite having only limited sequence similarity. Thus functionally analogous sliding clamps in eukaryotes and prokaryotes have similar structures.

The template-primer junctions are recognized by the multisubunit **replication factor C, or RFC**. Like the g complex in *E. coli*, this enzyme is an ATPase, and it helps to load on the processivity factor PCNA. Thus RFC is carrying out a similar function to the bacterial g-complex.

One of the first eukaryotic polymerases to be isolated was **DNA polymerase a,** which is now recognized as a catalyst of primer synthesis. This enzyme contains four polypeptide subunits, one with a polymerase activity (170 kDa), two that comprise a primase activity (50 and 60 kDa), and another subunit of (currently) undetermined function (70 kDa). DNA polymerase a has low processivity but high fidelity. This high fidelity is surprising because no 3' to 5' exonuclease is associated with the enzyme. Polymerase a, possibly with additional primases, catalyzes the synthesis of short segments of DNA and RNA that serve as primers for the replicative polymerases.

DNA polymerase e is related to polymerase d, and it may play a role in lagging strand synthesis. It is also dependent on PCNA, *in vivo*. However, no requirement has been identified for it in viral replication systems *in vitro.*

The compound aphidicolin will block the growth of mammalian cells. It does this by preventing DNA replication, and the targets of this drug are DNA polymerases a and d (as well as e). The fact that inhibition of these DNA polymerases with aphidicolin also stops DNA replication in mammalian cells argues that indeed, a and d are responsible for replication of nuclear DNA in eukaryotic cells. This conclusion is strongly supported by the phenotype of conditional loss-of-function mutations in the genes encoding the homologs to these polymerases in yeast. Such mutants do not grow at the restrictive temperature, indicating that d and a are the replicative polymerases. The biochemical evidence implicates polymerase a in primer formation, and d appears to be the major polymerases used to synthesize the new strands of DNA.

Table 5.4: Analogous components of the replication machinery in E. coliand eukaryotic cells.

Function	Bacterial (*E. coli*)	Number of subunits	Eukaryotic replication (SV40)	Number of subunits
Leading and lagging strand synthesis	asymmetric dimer, *E. coli*polymerase III	10 (3 in core)	polymerase d	2
Sliding clamp	b subunit	2	PCNA	3
Clamp loader	g-complex	6	RFC	multiple
Primase	DnaG	1	Polymerase a	4
Helicase	DnaB	6	T-antigen (SV40)	6
Bind single-stranded DNA	SSB	1	RFA	3
Swivel	Gyrase	4, A_2B_2	Topo I or Topo II	1 2 (homodimer)

The parallels between bacterial and eukaryotic DNA replication are striking. The overall strategy of synthesis is similar, and analogous proteins carry out similar functions, as listed in Table 5.4. It is difficult to determine whether the proteins carrying out similar functions are actually homologous proteins, i.e. encoded by genes descended from the same gene in the last common

ancestor. The protein sequence identities are marginal, and frequently the analogous proteins have different numbers of subunits. These differences complicate the analysis considerably, because different subunits in bacteria or mammals may have similar functions. However, the functional similarities are convincing.

Several other DNA polymerases have been isolated from eukaryotic cells. **DNA polymerase b** and **e** are involved in repair of nuclear DNA. DNA polymerase b is a single polypeptide of 36 kDa, and has no 3' to 5' exonuclease. **DNA polymerase g** replicates mitochondrial DNA.

Reverse transcriptaseis frequently referred to as an RNA-dependent DNA polymerase because it can use RNA as a template, but in fact it can use either RNA or DNA as a template. It is encoded by retroviruses, and hence it is present in cells infected with a retrovirus. This enzyme has widespread use in the laboratory for making complementary copies of RNA, called cDNA. Active copies of LINE1 repetitive elements (in mammals) or Ty1 repeats (in yeast), also encode reverse transcriptase. Thus in cells where these retrotransposable elements are being transcribed, active reverse transcriptase is also present. Reverse transcriptase also has an RNase H activity, which will digest away RNA from an RNA-DNA duplex.

In contrast to the other DNA polymerases discussed in this chapter, **terminal deoxynucleotidyl transferase**does not require a template. It adds dNTPs (as dNMP) to the 3' end of DNA, using that 3' hydroxyl as a primer. It is found in differentiating lymphocytes, and appears to be used physiologically to introduce somatic mutations into immunoglobulin genes. In the laboratory, it is used to add "homopolymer tails" to the ends of DNA molecules by incubating a linear DNA with one particular dNTP and terminal deoxynucleotidyl transferase.

As will be discussed in more detail in the next chapter, the ends of linear chromosomes (telomeres) must be expanded at each replication or they will eventually become shortened. The enzyme **telomerase**catalyzes the addition of many tandem copies of a simple sequence to the ends of the chromosomes. The template for this reaction is an RNA that is a component of the enzyme. Thus telomerase is a reverse transcriptase that only makes copies of the template that it carries, using the 3' end of a chromosomal DNA strand as the primer.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Non-polymerases Enzyees needed for Replication

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Polymerases

Of all the enzymatic functions needed for replication of DNA, the ability to catalyze the incorporation of deoxynucleotides into DNA is most central. Enzymes that catalyze this reaction, DNA polymerases, have been isolated from many species, and many species have multiple DNA polymerases. Our earliest and most complete understanding of the mechanism of these enzymes comes from studies of the first DNA polymerase isolated, called DNA polymerase I.

Mechanism of nucleotide addition by DNA polymerases

In 1956 Arthur Kornberg and his co-workers isolated a protein from *E. coli*that has many of the properties expected for a DNA polymerase used in replication. In particular, it catalyzes synthesis of DNA from deoxynucleotides, it requires a template and it synthesizes the complement of the template. It is a single polypeptide chain of 928 amino acids, and it is the product of the *pol*Agene. We now understand that this an abundant polymerase, but rather than synthesizing new DNA at the replication fork, it is used during the process of joining Okazaki fragments after synthesis and in DNA repair. Detailed studies of DNA polymerase I have been invaluable to our understanding of the mechanisms of polymerization. Although DNA polymerase I is not the replicative polymerase in *E. coli*, homologous enzymes are used in replication in other species. Also, the story of how the replicative DNA polymerases were detected in *E. coli*is a classic illustration of the power of combining biochemistry and genetics to achieve a more complete understanding of an important cellular process.

DNA polymerase I catalyzes the polymerization of dNTPs into DNA. This occurs by the addition of a dNTP (as dNMP) to the 3' end of a DNA chain, hence chain growth occurs in a 5' to 3' direction (Figure 5.11). In this reaction, the 3' hydroxyl at the end of the growing chain is a nucleophile, attacking the phosphorus atom in the a-phosphate of the incoming dNTP. The reaction proceeds by forming a phosphoester between the 3' end of the growing chain and the 5' phosphate of the incoming nucleotide, forming a phosphodiester linkage with the new nucleotide and liberating pyrophosphate (abbreviated PPi). Thus in this reaction, a phosphoanhydride bond in the dNTP is broken, and a phosphodiester is formed. The free energy change for breaking and forming these covalent bonds is slightly unfavorable for the reaction as shown. However, additional noncovalent interactions, such as hydrogen bonding of the new nucleotide to its complementary nucleotide and base-stacking interactions with neighboring nucleotides, contribute to make a total free energy change that is favorable to the reaction in the synthetic direction. Nevertheless, at high concentrations of pyrophosphate, the reaction can be reversed. In the reaction in the reverse direction, nucleotides are progressively removed and released as dNTP in a *pyrophosphorolysis*reaction. This is unlikely to be of large physiological significance, because a ubiquitous pyrophosphatase catalyzes the hydrolysis of the pyrophosphate to molecules of phosphate. This latter reaction is strongly favored thermodynamically in the direction of hydrolysis. Thus the combined reactions of adding a new nucleotide to a growing DNA chain and pyrophosphate hydrolysis insure that the overall reactions favors DNA synthesis. The basic chemistry of addition of nucleotides to a growing polynucleotide chain outlined in Figure 5.11 is common to virtually all DNA and RNA polymerases.

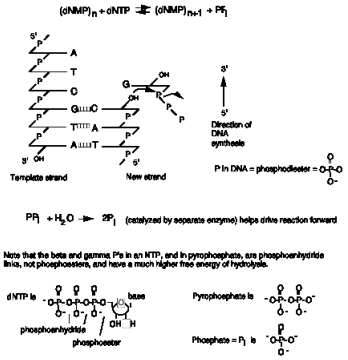


Figure 5.11.Reaction catalyzed by DNA polymerases.

The DNA synthesis reaction catalyzed by DNA polymerase I requires Mg2+, which is a cofactor for catalysis, and the four deoxynucleoside triphosphates (dNTPs), which are the monomeric building blocks for the growing polymer. The reaction also requires a template strand of DNA to direct synthesis of the new strand, as predicted by the double helical model for DNA and confirmed by the Meselson and Stahl experiment.

This reaction also requires a **primer**, which is a molecule (usually a chain of DNA or RNA) that provides the 3’ hydroxyl to which the incoming nucleotide is added. DNA polymerases cannot start synthesis on a template by simply joining two nucleotides. Instead, they catalyze the addition of a dNTP to a pre-existing chain of nucleotides; this previously synthesized chain is the primer. The primer is complementary to the template, and the 3’ end of the primer binds to the enzyme at the active site for polymerization (Figure 5.12). When a new DNA chain is being made, once a new nucleotide has been added to the growing chain, its 3' hydroxyl is now the end of the primer. The polymerase moves forward one nucleotide so that this new primer end is at the active site for polymerization. The alternative view, that the DNA primer-template moves while the DNA polymerase remains fixed, is also possible. In both cases the last nucleotide added is now the 3' end of the primer, and the next nucleotide on the template is ready to direct binding of another nucleoside triphosphate.

For the initial synthesis of the beginning of a new DNA chain, a primer has to be generated by a different enzyme; this will be discussed in more detail later in the chapter. For example, short oligoribonucleotides are the primers for the Okazaki fragments; these are found at the 5' ends of the Okazaki fragments and are made by an enzyme called primase. The RNA primers are removed and replaced with DNA (by DNA polymerase I) before ligation.

These requirements for Mg2+, deoxynucleotides and two types of DNA strands (template and primer) were discovered in studies of DNA polymerase I. We now realize that they are also required by all DNA polymerases.

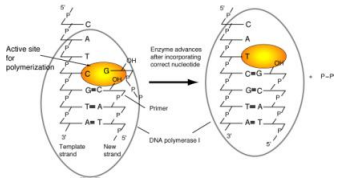


Figure 5.12.Chain elongation by DNA polymerase I. Binding of the correct incoming deoxynucleoside triphosphate to the active site for polymerization is directed by the deoxynucleotide on the template strand. The polymerase catalyzes formation of a phosphodiester bond with the new deoxynucleotide, and then it effectively moves forward so that the next deoxynucleotide on the template can direct binding of the next deoxynucleoside triphosphate to the active site. During elongation, the new DNA strand is also the primer.

The polymerization active site for DNA polymerase I has a specific dNTP-binding site (Figure 5.12), and the active site adjusts to the deoxynucleotide on the template strand to favor binding of the complementary deoxynucleotide at the active site. Thus the polymerase catalyzes addition to the growing chain of the deoxynucleotide complementary to the deoxynucleotide in the template strand.

In the reaction catalyzed by DNA polymerase I, and all other DNA polymerases studied, the incoming deoxynucleotide is activated. The phosphoanhydride bonds in the triphosphate form of the deoxynucleotide are high-energy bonds (i.e., they have a negative, or favored, free energy of hydrolysis), and the b- and g- phosphates make a good leaving group (as pyrophosphate) after the nucleophilic attack. In contrast, the end of the growing DNA chain is not activated; it is a simple 3'-hydroxyl on the last deoxynucleotide added. This addition of an activated monomer to an unactivated growing polymer is called a **tail-growth mechanism**. DNA polymerases using this mechanism can only synthesize in a 5' to 3' direction, and all known DNA and RNA polymerases do this. Some other macromolecules, such as proteins, are made by a **head-growth mechanism**. In this case, the nonactivated end of a monomer attacks the activated end of the polymer. The lengthened chain again contains an activated head (from the last monomer added).

Exercise

Question 5.4. Describe a hypothetical head-growth mechanism for DNA synthesis. In which direction does chain synthesis occur in this mechanism?

Proofreading the newly synthesized DNA by a 3’ to 5’ exonuclease that is part of the DNA polymerase

The protein DNA polymerase I has additional enzymatic activities related to DNA synthesis. One, a 3’ to 5’ exonuclease, is intimately involved in the accuracy of replication. **Nucleases** are enzymes that catalyze the breakdown of DNA or RNA into smaller fragments and/or nucleotides. An **exonuclease** catalyzes cleavage of nucleotides from the end of a DNA or RNA polymer. An **endonuclease** catalyzes cutting within a DNA or RNA polymer. These two activities can be distinguished by the ability of an endonuclease, but not an exonuclease, to cut a circular substrate. A 3’ to 5’ exonuclease removes nucleotides from the 3’ end of a DNA or RNA molecule.

DNA synthesis must be highly accurate to insure that the genetic information is passed on to progeny largely unaltered. Bacteria such as *E. coli*can have a mutation rate, as low as one nucleotide substitution in about 10^9 to 10^{10} nucleotides. This low error frequency is accomplished by a strong preference of the polymerase for the nucleotide complementary to the template, which allows about one substitution every 10^4 to 10^5 nucleotides. The accuracy of DNA synthesis is enhanced by a **proofreading** function in the polymerase that removes incorrectly incorporated nucleotides at the end of the growing chain. With proofreading, the accuracy of DNA synthesis is improved by a factor of 10^2 to 10^3, so the combined effects of nucleotide discrimination at the polymerization active site plus proofreading allows only about one substitution in 10^6 to 10^8 nucleotides. Further reduction in the error rate is achieved by mismatch repair (Chapter 7).

The proofreading function of DNA polymerase I is carried out by a 3' to 5' exonuclease (Figure 5.13). It is located in a different region of the enzyme from the active site for polymerization. When an incorrect nucleotide is added to the 3' end of a growing chain, the rate of polymerization decreases greatly. The primer-template moves to a different active site on the enzyme, the one with the 3’ to 5’ exonucleolytic activity. The incorrect nucleotide is cleaved, and the primer-template moves back to the polymerization active site to resume synthesis.The enzyme distinguishes between correct and incorrect nucleotides at the 3’ end of the primer, such that the 3' to 5' exonuclease much more active when the terminus of the growing chain is not base paired correctly, but the polymerase activity exceeds that of the 3' to 5' exonuclease activity when the correct nucleotide is added.

The polymerizing activity and the proofreading 3' to 5' exonuclease found in DNA polymerase I are also found in most other DNA polymerases. These are central activities to DNA replication.

Tail growth mechanisms allow proofreading and subsequent elongation. If the end of the growing chain were activated (as in head growth), then proofreading would eliminate the activated end and elongation could not continue.

D.

A.

C.

B.

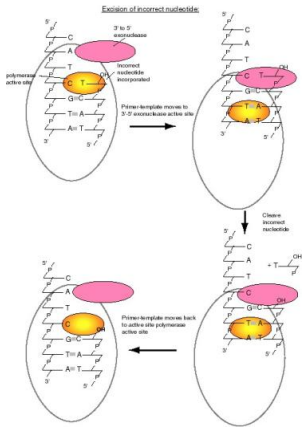


Figure 5.13.Excision of an incorrect nucleotide by DNA polymerase I. Incorporation an incorrect nucleotide (e.g., a T opposite a C, lower panel) (A) causes the primer-template to shift to the 3'-5' exonuclease active site (B) where the incorrect nucleotide is excised (C). The primer-template then can move back to the polymerase active site to resume synthesis (D).

Exercise

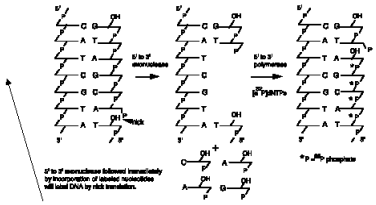
Question 5.4 Removal of a nucleotide from the 3’ end of the growing chain by a 3’ to 5’ exonuclease is not the reverse of the polymerase reaction. Can you state what the difference is?

Removal of nucleotides by a 5' to 3' exonuclease that is part of DNA polymerase I

In addition to the polymerase and 3’ to 5’ exonuclease common to most DNA polymerases, DNA polymerase I has an unusual 5' to 3' exonucleolytic activity. This enzyme catalyzes the removal of nucleotides in base-paired regions and can excise either DNA or RNA. It is used by the cell to remove RNA primers from Okazaki fragments and in repair of damaged DNA.

This 5' to 3' exonuclease, in combination with the polymerase, has useful applications in the laboratory. One common use is to label DNA *in vitro* by **nick translation** (Figure 5.14). In this process, DNA polymerase I will remove the DNA from a nicked strand by the 5' to 3' exonuclease, and then use the exposed 3' hydroxyl at the nick as a primer for new DNA synthesis by the 5' to 3' polymerase, thereby replacing the old DNA. The result is also a movement, or translation, of the nick from one point on the DNA to another, hence the process is called nick translation. If the reaction is carried out in the presence of one or more radiolabeled deoxynucleoside triphosphates (e.g., [a^{32}P] dNTPs), then the new DNA will be radioactively labeled.

A similar process can be used to repair DNA in a cell. As will be discussed in Chapter 7, specific enzymes recognize a damaged nucleotide and cleave upstream of the damage. One way to remove the damaged DNA and replace it with the correct sequence is with the 5' to 3' exonuclease of DNA polymerase I and accompanying DNA synthesis.



[$a^{32}P$]

Figure 5.14.*The 5' to 3' exonuclease of DNA polymerase I can be used in nick translation to label DNA in vitro.*

Structural domains of DNA polymerase I

Further understanding of the mechanism of the three enzymatic functions of DNA polymerase can be obtained from a study of the three-dimensional (3-D) structure of the protein. Much of our knowledge of the structure of DNA polymerase I has come from biochemical characterization and more recently by determination of the 3-D structure using X-ray crystallography. These studies have shown that distinct structural domains of DNA polymerase I contain the different catalytic activities. Also, the 3-D structure provided the first look at what is now recognized as a common structure for many polymerases.

Mild treatment with the protease subtilisin cleaves DNA polymerase I into two fragments. The small fragment contains the 5' to 3' exonuclease, and the larger, or "Klenow," fragment (named for the biochemist who did the cleavage analysis) contains the polymerase and the proofreading 3' to 5' exonuclease (Figure 5.15). Thus the two activities common to most polymerases are together in the Klenow fragment, whereas the distinctive 5' to 3' exonuclease is in a separable domain. The fact that a mild treatment with a protease without a precise sequence specificity indicates that an exposed, readily cleaved domain connects the large and small fragments. Both these observations suggest that the 5' to 3' exonuclease was an active domain added to a polymerase plus proofreading domain during the evolution of *E. coli*. The Klenow polymerase is used in several applications in the laboratory, e.g., labeling the ends of restriction fragments by filling in the overhangs and sequencing by the dideoxynucleotide chain termination method.

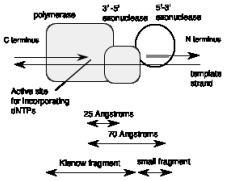


Figure 5.15. DNA polymerase I from *E. coli* has three active sites in three structural domains in one polypeptide.

The 3-D structure of the large fragment of DNA polymerase I, determined by crystallography, provides additional insight into the enzymatic functions of key structural components. The large fragment has a deep cleft, about 30 Å deep, into which the template strand and primer bind. This cleft resembles a "cupped right hand" as illustrated in Figure 5.16. The "palm" is formed by a series of b-sheets and the thumb and fingers are made by a-helices. The polymerase active site has been mapped within the deep cleft, with contributions from the b-sheets that form the palm and the a-helices forming the fingers. You can see more detailed views of the structure of the Klenow fragment at the Course/Book web site (currently www.bmb.psu.edu/courses/bmb400/default.htm. Click on the link to kinetic images, download the MAGE program and the kinemage file for DNA polymerase I, and view them on your own computer.)

The 3' to 5' proofreading exonuclease is located in another part of the structure of the Klenow fragment, about 25 Å from the polymerase active site. Thus the primer terminus has to move this distance in order for the enzyme to remove misincorporated nucleotides.

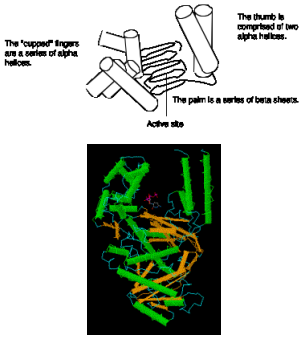


Figure 2 of Kim et al. (1995) Crystal structure of Thermus aquaticus DNA polymerase, Nature 376:612-616, which shows all 3 domains.

The large Klenow fragment of the *E. coli*DNA polymerase I lacks the 5' to 3' exonuclease, so the 3-D structure of the Klenow fragment gives no information about that exonuclease. However, the 5' to 3' exonuclease domain can be seen in the structure of DNA polymerase from the thermophilic bacterium *Thermus aquaticus*. This protein structure is very similar to that of DNA polymerase I of *E. coli*in the polymerase and 3' to 5' exonuclease domains, and it has an additional 5' to 3' exonuclease domain located about 70 Å from the polymerase active site. This is a large distance, but remember that this exonuclease is working on a different region of the DNA molecule than the polymerase. The 5' to 3' exonuclease uses one part of the DNA molecule as a substrate for excising primers or removing damaged DNA, whereas the polymerase uses a different part of the DNA molecule as a template to direct synthesis of a new strand.

Curiously, a region homologous to the proofreading 3' to 5' exonuclease domain of DNA polymerase I is present in the *Thermus aquaticus*polymerase structure, but it is no longer functional. The absence of proofreading accounts for the elevated error rate in this polymerase used very commonly for amplification of DNA by PCR. Of course, this polymerase is used in PCR because it is stable at the high temperatures encountered during the cycles of PCR. Some other thermostable polymerases with a lower error rate have become available more recently for use in PCR.

Similar "cupped right hand" structures occur in the tertiary structure of T7 RNA polymerase and the HIV reverse transcriptase. Thus DNA polymerase I was the first member described in what we now realize is a large class of nucleic acid polymerases. This family includes single unit polymerases for both RNA and DNA synthesis. You can access a tutorial on the T7 DNA polymerase at www.clunet.edu/BioDev/omm/exh...s.htm#displays. This structure has some similarities to that of DNA polymerase I.

Physiological role of DNA polymerase I

Although studies of DNA polymerase I have provided much information about the mechanism of DNA synthesis, genetic analysis has shown that the polymerase function of this enzyme is not required for DNA replication. DNA polymerase I is encoded by the *polA*gene in *E. coli*.However, no mutant allele of *polA*was isolated in screens for conditional mutants defective in DNA replication. The most compelling argument that this polymerase is not required for replication came from an examination of thousands of *E. coli*mutants, assaying them for DNA polymerase I activity. A mutant *polA* strain was isolated (Figure 5.16). This mutant allele, called *polA1*, contained a nonsense codon, leading to premature termination of synthesis of the product polypeptide and hence a loss of polymerase function. However, the mutant strain grew at a normal rate, which shows that DNA polymerase I is *not*required for DNA synthesis. The most striking phenotype of the *polA1*mutant was its strongly reduced ability to repair DNA damage. Further investigation led to the isolation of conditional lethal alleles of the *polA*gene. The mutant DNA polymerase I proteins encoded by these

conditional lethal alleles are defective in the 5' to 3' exonuclease activity, demonstrating that this activity is required for cell viability. The 5' to 3' exonuclease activity removes RNA primers during synthesis of the lagging strand at the replication fork, and it is used in DNA repair.

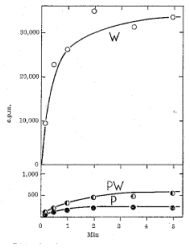


Figure 5.16. Extracts from a *polA1* mutant strain are defective in DNA polymerase activity. DNA polymerization in extracts of *E. coli*cells was measured by the incorporation of radiolabeled dTTP into DNA. The wild type strain (line labeled W) showed high activity. Mutants of this strain were systematically screened for the loss of this DNA polymerizing activity, and one was found (line labeled P; note the change in scale from the upper panel). This mutant strain has less than 1% of the wild type activity, as shown by the mixing 1 part of wild-type extract with 99 parts of the mutant extract (a 100-fold dilution; results are shown as line PW). The mutation was mapped to *polA*, which encodes DNA polymerase I. The mutant strain grows as well as the wild type, showing that DNA polymerase I is not required for DNA replication. This figure is from De Lucia and Cairns (1969) Nature 224:1164-1166.

DNA polymerase III is a highly processive, replicative polymerase

The conclusion that DNA polymerase I is not the replicative polymerase for *E. coli*led to the obvious question of what enzyme is actually used during replication. Investigation of the genes isolated in screens for mutants that are conditionally deficient in replication led to the answer. The replicative polymerase in *E. coli*is DNA polymerase III.

DNA polymerase I is more abundant than other polymerases in *E. coli*and obscures their activity. Thus the depletion of DNA polymerase I activity in *polA1*mutant cells (Figure 5.17) provided the opportunity to observe the other DNA polymerases. DNA polymerases II and III were isolated from extracts of *polA1*cells, named in the order of their discovery.

DNA polymerase IIis a single polypeptide chain whose function is uncertain. Strains having a mutated gene for DNA polymerase II (*polB1*) show no defect in growth or replication. However, the activity of DNA polymerase II is increased during induced repair of DNA, and it may function to synthesize DNA opposite a deleted base on the template strand.

Genetic evidence clearly shows that **DNA polymerase III**is used to replicate the *E. coli* chromosome. This enzyme is composed of multiple polypeptide subunits. Several of the genes encoding these polypeptide subunits were identified in screens for conditional lethal mutants defective in DNA replication. Loss of function of these *dna*genes blocks replication, showing that their products are required for replication.

Low abundance and high processivity of DNA polymerase III

DNA polymerase III has many of the properties expected for a replicative polymerase. One of the complications to studies of DNA polymerase III is that different forms were isolated by various procedures. We now realize that these forms differ in the number of subunits present in the isolated enzyme. For enzymes with multiple subunits, we refer to the complex with all the subunits needed for its major function as the **holoenzyme** or **holocomplex**. The DNA polymerase holoenzyme has ten subunits, which will be discussed in detail in the next section.

It is the DNA polymerase holoenzyme that has the properties expected for a replicative polymerase, whereas DNA polymerase I does not (see comparison in Table 5.1). It is *less abundant*than DNA polymerase I, but large number of replicative DNA polymerases are not needed in the cell. Only one or two polymerases can be used at each replication fork, so the 10 molecules of the DNA polymerase III holoenzyme will suffice. DNA polymerase III *catalyzes DNA synthesis at a considerably higher rate*than DNA polymerase I, by a factor of about 70. The elongation rate measured for the DNA polymerase III holoenzyme (42,000 nucleotides per min) is close to the rate of replication fork movement measured *in vivo*in *E. coli*(60,000 nucleotides per min).

A key property for a replicative DNA polymerase is *high processivity*, which is a striking characteristic of the DNA polymerase III holoenzyme. **Processivity** is the amount of polymerization catalyzed by an enzyme each time it binds to an appropriate template, or primer-template in the case of DNA polymerases. It is measured in nucleotides polymerized per binding event. In order to replicate the 4.5 megabase chromosome of *E. coli*in 30 to 40 min, DNA polymerase needs to synthesize DNA rapidly, and in a highly processive manner. DNA polymerase I synthesizes less than 200 nucleotides per binding event, but as the holoenzyme, DNA polymerase III is much more processive, exceeding the limits of the assay used to obtain the results summarized in Table 5.1. In contrast, the DNA polymerase III core, which has only three subunits (see next section), has very low processivity.

Table 5.1. Comparison of DNA polymerases I and III (Pol I and Pol III)

Property	Pol I	Pol III core	
molecules per cell	400	40	10
nucleotides polymerized min-1 (molecule enzyme)-1	600	9000	42,0(
processivity [nucleotides polymerized per initiation]	3-188	10	>105
5' to 3' polymerase	+	+	+
3' to 5' exonuclease, proofreading	+	+	+
5' to 3' exonuclease	+	-	-

Note: + and – refer to the presence or absence of the stated activity in the enzyme.

Question 5.6. If the rate of replication fork movement measured *in vivo* in *E. coli*is 60,000 nucleotides per min, how many forks are needed to replicate the chromosome in 40 min? Recall that the size of the *E. coli*chromosome is 4.64 ´ 10^6 bp.

Subunits and mechanism of DNA polymerase III

The DNA polymerase III enzyme has four distinct functional components, and several of these contain multiple subunits, as listed in Table 5.2 and illustrated in Figure 5.18. The a and e subunits contain the major polymerizing and proofreading activities, respectively. They combine with the q subunit to form the catalytic core of the polymerase. This core can be dimerized by the t_2 linker protein to form a subassembly called DNA polymerase III'. Addition of the third functional component, the g complex, generates another subassembly denoted DNA polymerase III*. All of these subassemblies have been isolated from *E. coli*and have been characterized extensively. The final component is the b2 dimer, which when combined with DNA polymerase III* forms the holoenzyme.

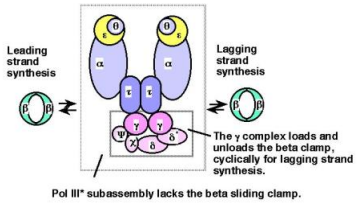


Figure 5.19.

The various activities of DNA polymerase III can be assigned to individual subunits (Table 5.2). For instance, the major polymerase is in the a subunit, which is encoded by the *dnaE*gene (also known as *polC*). The 3' to 5' exonuclease is in the e subunit, which is encoded by the *dnaQ*gene (also known as the *mutD*gene). However, maximal activity is obtained with combinations of subunits. The DNA polymerase III core is a complex of the a, e and q subunits, and the activity of the core in both polymerase and 3' to 5' exonuclease assays is higher in than in the isolated subunits.

Table 5.2. Subassemblies of DNA polymerase III, major subunits, genes and functions

Functional component	Subunit	Mass (kDa)	Gene	Activity or function
Core polymerase	a	129.9	*polC=dnaE*	5' to 3' polymerase
	e	27.5	*dnaQ=mutD*	3'-5' exonuclease
	q	8.6		Stimulates e exonuclease
Linker protein	t	71.1	*dnaX*	Dimerizes cores
Clamp loader	g	47.5	*dnaX*	Binds ATP
(or g complex)	d	38.7		Binds to b
(ATPase)	d'	36.9		Binds to g and b
	c	16.6		Binds to SSB
	y	15.2		Binds to c and g
Sliding clamp	b	40.6	*dnaN*	Processivity factor

The activities of the subunits can be measured *in vitro*by appropriate biochemical assays. In addition, the phenotype of mutations in the gene encoding a given subunit can show that subunit is required for a particular process. Mutant a subunits are the product of conditional lethal alleles discovered in screens for *dna*genes, but they also were discovered as the product of polymerase-defective alleles defining the *polC*gene. Thus the *dnaE*gene is the same as same as the *polC*gene, showing that this subunit with polymerase activity is needed in replication. Similarly, the phenotype of mutations in the gene encoding the e subunit shows that it is needed for proofreading. Mutant alleles of the *dnaQ*gene were identified in a screen for **mutator** genes, which generate a high frequency of mutants in bacteria when defective. These alleles defined a gene *mutD*, which was subsequently shown to be the same as *dnaQ*. The mutator phenotype of mutant *dnaQ/mutD*strains results from a lack of proofreading by the e subunit during replication, allowing more frequent incorporation of incorrect nucleotides into DNA.

The b2 dimer is the key protein that confers *high processivity*on DNA polymerase III. Association of the b2 dimer with DNA polymerase III increases the processivity from about 10 nucleotides polymerized per binding event to over 100,000 nucleotides polymerized per binding event (Table 5.1). This dimeric protein forms a ring through which the duplex DNA can pass; the ring will slide easily along DNA unless impeded, as, for example, by proteins bound to the template DNA. Thus the b2 dimer acts as a *sliding clamp*, holding the polymerase onto the DNA being copied. Once DNA polymerase III is associated with the clamp on DNA, it will polymerize until it reaches the next primer for an Okazaki fragment during lagging strand synthesis. For leading strand synthesis, the DNA polymerase presumably remains associated with the DNA via the b2 clamp until the chromosomal DNA is completely replicated. The 3-D structure of the b2 dimer, determined by X-ray crystallography, shows a macromolecular ring. This structure can be viewed at the web site for the course and at the Online Museum of Macromolecules (www.clunet.edu/BioDev/omm/exh...s.htm#displays).

The g-complex contains several subunits: two molecules of g subunits and one molecule each of d, d', c, and y. It *loads*the b_2 dimer clamp onto a primer-template, in a process that requires ATP hydrolysis (Figure 5.19). The catalytic core of DNA polymerase III will then link to the template-bound clamp and will initiate highly processive replication. The g-complex also serves to unload the clamp once an Okazaki fragment is completed during lagging strand synthesis; hence it is both a clamp loader and unloader, allowing the polymerase and the clamp to cycle repeatedly from one Okazaki fragment to another.

The g-complex carries out these opposite activities on different structures, loading on the clamp at a template-primer and unloading the clamp at the end of a completed Okazaki fragment. For instance, encountering the 5' end of the previously synthesized Okazaki fragment may be the distinctive structure that shifts the g-complex into its unloading mode. It does not unload the clamp while DNA polymerase III is catalyzing polymerization.

Figure 5.19 illustrates the proposed steps in this process. The g-complex in the ATP-bound form binds the b_2 clamp, whereas the g-complex in the ADP-bound form releases the b_2 clamp. Thus loading and unloading depend on a round of ATP hydrolysis. When the g-complex in the ATP-bound form binds the b_2 clamp, the DNA polymerase III holoenzyme is in a conformation that allows it to find a primer-template. The ring of the b_2 clamp is held open by the g-complex-ATP, allowing it to bind around a primer-template. Hydrolysis of ATP by the g-complex leaves it in an ADP-bound form. In this new, ADP-bound conformation of the g-complex, it dissociates from the b_2 clamp, thereby allowing the b_2 clamp to bind to the catalytic core of the holoenzyme and also close around the primer-template. The holoenzyme is now ready to catalyze processive DNA synthesis. Elongation continues until the holoenzyme encounters a previously synthesized Okazaki fragment. Now the g-complex binds ATP (presumably by an ADP-ATP exchange reaction) and shifts into the conformation for binding to the b_2 clamp and taking it off the DNA template. This half of the holoenzyme is now able to dissociate from the template and find the next primer-template junction to begin synthesis of another Okazaki fragment.

The clamp loading and unloading activities of the g-complex are a cycle of changes in protein associations. These changes occur because of the enzymatic activities of the complex, which in turn alter the conformations of the proteins and their preferred interactions. As shown in Figure 5.19, the g-complex is an **ATPase**, which is an enzyme that catalyzes the hydrolysis of ATP to ADP and phosphate. It is also an ATP-ADP exchange factor.

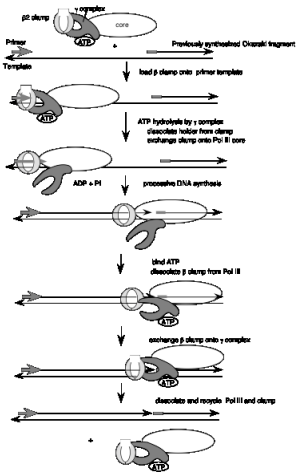


Figure 5.19. (Previous page)Loading and unloading of the b_2 clamp by the g-complex in the DNA polymerase III holoenzyme. The b_2 clamp is shown as a ring that can be bound and opened by the ATP-bound form of the g-complex (shown as a pincher shape, but its shape is not known). After initial binding of the b_2 clamp, ATP-hydrolysis by the g-complex causes this complex to shift into an ADP-bound conformation that allows it to release the b_2 clamp, so that the b_2 clamp is loaded onto the primer template and also linked to the catalytic core of the polymerase. After completion of the Okazaki fragment, the g-complex exchanges the ADP for ATP, and is now is a conformation to bind the b_2 clamp and unload it. Not all the steps in this model have been demonstrated, but it is useful to illustrate how cycles of ATP hydrolysis could be used in loading and unloading the b_2 clamp. The figure shows only the half of the DNA polymerase III holoenzyme engaged in synthesis of the lagging strand; the other half is thought to be engaged in synthesis of the leading strand, but is not shown here to keep the diagram relatively simple.

Changes in conformation and activity of proteins depending on whether they are bound to a nucleoside triphosphate (ATP or GTP) or a nucleoside diphosphate (ADP or GDP) is a common theme in biochemistry. The GTP-bound forms of proteins, which can be turned off by GTP-hydrolysis and reactivated by GDP-GTP exchange proteins, mediate critical cell signaling events. As will be seen in

Chapter 14, GTP- and GDP-bound forms of translation factors carry out opposite functions. Proteins assume different conformations depending on the cofactor bound (in this case a nucleotide), and each conformation has a distinct activity. The ability to change the conformation by a hydrolytic activity (converting ATP to ADP and phosphate) allows the protein to shift activities readily.

The two catalytic cores of DNA polymerase III are joined together by the t subunits to make an asymmetric dimer (see Figure 5.18). The half of the holoenzyme without the g complex is proposed to synthesize the leading strand of new DNA, and the core with the g complex is proposed to synthesize the lagging strand. Both of the cores in the asymmetric dimer are associated with a b2 clamp at the replication fork. In this model, synthesis of *both*the leading and lagging strands is catalyzed by the *same*DNA polymerase III complex, thereby coordinating synthesis of both new strands strand. Note that if the template for lagging strand synthesis is looped around the enzyme, then leading and lagging strand synthesis would be occurring in the same direction as replication fork movement (Figure 5.20), despite the opposite polarities of the two template strands. Thus the asymmetric dimer model suggests a means to couple both leading strand and lagging strand synthesis.

Replication machinery

C.

B.

A.

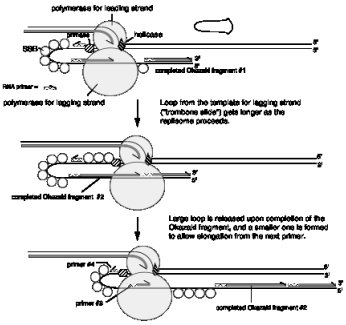


Figure 5.20. Simultaneous synthesis of both leading and lagging strands by an asymmetric dimer of DNA polymerase III. In this model, one of the catalytic cores synthesizes the leading strand and the other synthesizes the lagging strand. The template for the lagging strand may be looped around the polymerase, making this strand resemble the slide on a trombone (panel A); this model has been called the Trombone Slide model. When this is done, the synthesis of the Okazaki fragments is in the same direction as leading strand synthesis and fork movement (i.e. left to right in the diagram). Effectively, wrapping the template strand for lagging strand synthesis around the polymerase orients this strand at the active site in the same polarity as the template for leading strand synthesis. (A) The enzyme primase makes a primer in the enlarging single stranded loop, while the DNA polymerase III core catalyzes extension of the Okazaki fragment. (B) The Okazaki fragment is completed, and DNA polymerase III encounters the previously synthesized Okazaki fragment. (C) Now the loop with the completed Okazaki fragment is released and a new single stranded loop is formed. DNA polymerase III initiates replication at the primer in the new loop, and lagging strand replication resumes, again moving in the same direction as the fork (and the leading strand). SSB is single-strand binding protein and primase catalyzes the synthesis of primers (which are mainly RNA) for Okazaki fragments; these will be discussed in more detail later.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Specialized DNA Structures

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Welcome to the Biology Library. This Living Library is a principal hub of the LibreTexts project, which is a multi-institutional collaborative venture to develop the next generation of open-access texts to improve postsecondary education at all levels of higher learning. The LibreTexts approach is highly collaborative where an Open Access textbook environment is under constant revision by students, faculty, and outside experts to supplant conventional paper-based books.

CHAPTER OVERVIEW

6. DNA replication II: Start, stop and control

Regulation is largely exerted at the initiation of replication, and methods for finding origins and termini of replication will be covered. The proteins involved in control of replication initiation in *E. coli* and yeast will be discussed. One solution to the problem of completing the synthesis of linear DNAs in eukaryotes will be described - that of making telomeres. Some of the factors controlling the rate of initiation of replication will be discussed briefly.

Topic hierarchy

References

1. Jacob, F., Brenner, S. and Cuzin, F. (1963) On the regulation of DNA replication in bacteria. Cold Spring Harbor Symposium on Quantitative Biology **28:**329-348.
2. Danna and Nathans, D. (1972) (Proc. Natl. Acad. Sci., USA **69:**3097-3100.
3. Brewer and Fangman (1987) Cell **51:**463-471.
4. Dutta A and Bell SP (1997) Initiation of DNA replication in eukaryotic cells. Annu Rev Cell Dev Biol 13():293-332
5. Cimbora, Daniel M. and Groudine, Mark (2001) The control of mammalian DNA replication: A brief history of space and timing. Cell 104: 643-646.
6. Diffley, John F.X. (1995) Once and only once upon a time: specifying and regulating origins of DNA replication in eukaryotic cells. Genes & Development 10:2819-2830.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

6.1: The Replicon

It is critical that all the DNA in a cell be replicated once, and only once, per cell cycle. Jacob, Brenner and Cuzin defined a **replicon** as the unit in which the cell controls individual acts of replication. The replicon initiates and completes synthesis once per cell cycle. Control is exerted primarily at initiation. They proposed that an **initiator** protein interacted with a DNA sequence, called a **replicator**, to start replication. The replicator can be identified genetically as a DNA sequence required for replication, whereas the **origin** is defined by physical or biochemical methods as the DNA sequence at which replication begins. For many replicons, such as the *E. coli oriC* and the autonomously replicating sequences (or *ARS*) in yeast, the replicator is also an origin. However, this need not be the case: the replicon for amplified chorion genes in silkmoths has an origin close to, but separable from, the replicator. Initiator proteins have now been identified for some replicons, such as the DnaA protein in *E. coli* and the *Origin Recognition Complex* in the yeast *Saccharomyces cerevisiae*. In both cases, they bind to the replicators, which are also origins in these two species.

The replicator is a sequence of DNA needed for synthesis of the rest of the DNA in a replicon. It is a control element that affects the chromosome on which it lies. We say that this element acts in *cis*, since the replicator and the replicon are on the same chromosome. In contrast, the initiator is a protein that can be encoded on any chromosome in a cell. Thus is acts in *trans*, since it does not have to be encoded on the same chromosome as the replicon that it controls. In general, a *trans*-acting factor is an entity, usually a protein, that can diffuse through the cell to act in regulation of a certain target, whereas a *cis*-acting DNA sequence is on the same chromosome as the target of control. This pattern of a *trans*-acting protein binding to a *cis*-acting site on the DNA is also seen in transcriptional control.

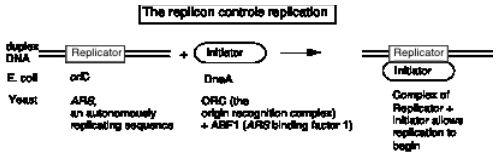


Figure 6.1. Components of a replicon, the unit in which the cell controls repliation.

Exercise

Although *E. coli* has a single origin in a single replicon, eukaryotic chromosomes have multiple origins, and multiple replicons. Consider a line of mammalian cells growing in culture that has an S phase of 5 hr, i.e. all the genome is replicated in 5 hr. The haploid genome size is 3 x 10^9 bp. If the rate of replication fork movement in this cell lines is 2000 bp per min, how many replication origins are required to replicate the entire haploid genome during S phase? Assume that two replication forks emerge from each origin (this is bidirectional replication, see below).

Experimental approaches to map origins and termini of replication and to distinguish between uni- and bidirectional replication

Several experimental techniques have been established for finding where replication begins and ends on chromosomes, and for distinguishing between unidirectional and bidirectional replication. We will cover two major techniques.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

6.2: Structural analysis of pulse-labeled DNA molecules

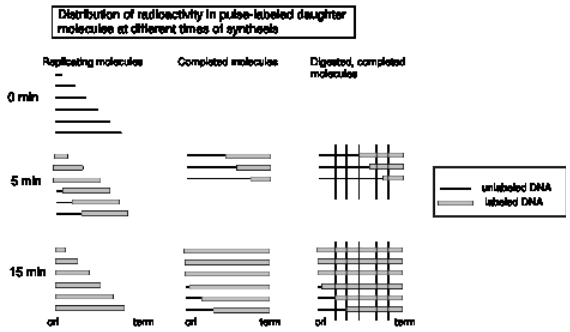


Figure 6.3 Restriction map of SV40

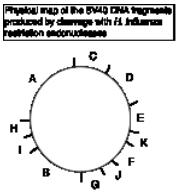


Table 6.1. Appearance of radiolabel into restriction fragments of completed SV40 DNA molecules. The relative amount of pulse label from each restriction fragment is given below (the relative amount of pulse label is the 3H/32P ratio of each fragment, corrected for thymidine content and normalized to 1 for fragment A).

	Relative amount of pulse label		
Fragment	5 min	10 min	15 min
A	1.0	1.0	1.0
B	3.9	3.0	2.3
C	0	0.75	0.75
D	0.92	0.86	1.1
E	1.8	2.0	1.7
F	4.0	3.1	2.4
G	5.4	4.2	2.6
H	1.7	2.5	2.0
I	2.7	3.0	2.2
J	4.9	3.7	2.6
K	2.4	2.9	1.9

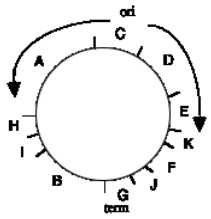


Figure 6.4.B.

Question 6.2: What would the pattern be for unidirectional replication?

Question 6.3: What would be the pattern if there were two origins, say in fragments E and H, with bidirectional replication from each?

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

6.3: Two-dimensional gels to analyze the number and position of replication origins

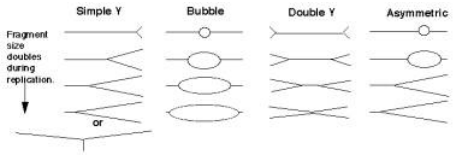
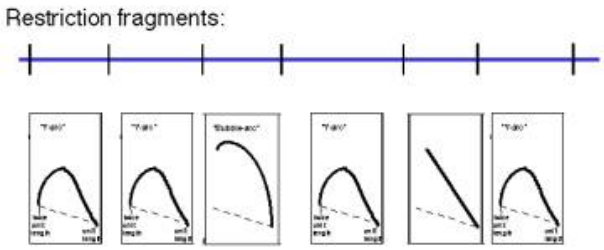


Figure 6.6.C.Summary of the patterns of fragments containing replication intermediates in two-dimensional gels.

Question 6.4. A restriction map is shown for a portion of a chromosome below, along with the patterns on two-dimensional gels for the replication intermediates formed by each fragment. Where are the origins and termini? Can you deduce the direction of replication fork movement?



Question 6.5. How can you calculate the position of an origin within a DNA fragment from an asymmetric fork/bubble pattern on a 2-D gel of replicating DNA molecules?

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

6.4: Replication landscape in E. coli

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

6.E: DNA replication II: Start, stop and control (Exercises)

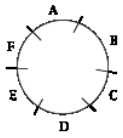
6.12

At what step is the rate of DNA replication in *E. coli* is regulated - initiation, elongation or termination?

6.13

The following problem further illustrates the analysis of replication by pulse-labeling, using a hypothetical virus and constructed data. Consider the replication of a circular viral DNA in infected cells. The infected cells were pulse labeled with [3H] thymidine for 1, 2, 3 and 4 min; it takes 4 min for the DNA molecules to be replicated in this system (from initiation to termination). Those DNA molecules that had completed synthesis at each time point were isolated, cut with a restriction endonuclease, and assayed for radioactivity in each fragment. This restriction endonuclease cleaves the circular DNA into 6 fragments, named A, B, C, D, E, and F in a clockwise orientation around the genome. The following results were obtained; a plus (+) means the fragment was radioactively labeled, and a minus (-) means it was not labeled.

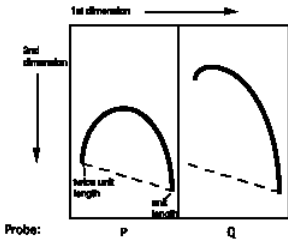
Fragment	Time of labeling (min)			
	1	2	3	4
A	-	-	+	+
B	-	-	-	+
C	-	-	+	+
D	-	+	+	+
E	+	+	+	+
F	-	+	+	+



a. What restriction fragment has the origin and which has the terminus of replication?
b. In which direction(s) does this viral DNA replicate?

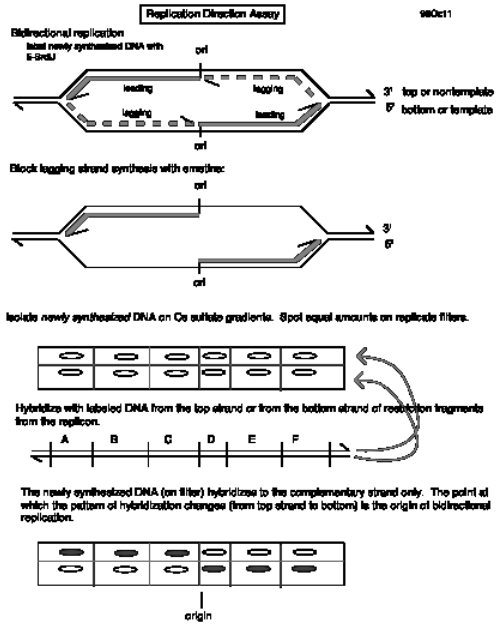
6.14

The two-dimensional gels developed by Brewer and Fangman were used to examine the origin of replication of a DNA molecule. In this system, *replicating* molecules are cleaved with a restriction endonuclease and separated in two dimensions. The first dimension separates on the basis of size, and the second separates on the basis of shape (more pronounced deviations from linearity move slower in the second dimension). After blotting the DNA onto a membrane, it is probed with fragments from the replicon under study. Restriction fragment P gives the pattern shown on the left, and the adjacent fragment Q gives the pattern shown on the right. The dotted line denotes the diagonal expected if all molecules were linear. Assuming both P and Q are in the same replicon, what can you conclude about the positions of origins of replication?



6.15

Dr. Howard Cedar and his colleagues at the Hebrew University in Jerusalem have developed a replication direction assay to map origins of replication on chromosomes (Kitsberg et al., Nature 366: 588-590, 1993). Growing cells are treated with the drug emetine to *inhibit lagging strand synthesis*. Leading strand synthesis continues, and this newly synthesized DNA is density labeled by incorporating 5-bromodeoxyuridylate (5-bromodeoxyuridine is added to the medium). The DNA is then sheared and denatured, and the newly synthesized leading strand DNA is separated from the rest of the DNA by sedimentation equilibrium on Cs2SO4 gradients. Samples of the heavy density DNA (containing 5-bromodeoxyuridylate) are spotted onto a membrane, and equal amounts are hybridized to labeled, separated strands of restriction fragments throughout a region.



Use of this approach to map replication origins in the human b-like globin gene cluster led to results like those below. The names of the genes are given above the line, and the names of the restriction fragments are given below the line. A + means that the leading strand (with 5-bromodeoxyuridylate incorporated) hybridized preferentially to a labeled probe corresponding to the designated strand, whereas a - means that the leading strand DNA did not hybridize to the designated probe. The genes are transcribed from left to right in this diagram, so the "top" strand reads the same as the mRNA in the coding regions (our convention is "nontemplate") and the "bottom" strand (abbreviated "bot") is complementary to the top strand ("template" or "antisense" strand).

e Gg Ag yh d b

|_|_-->_|____|_-->_|_-->_|____|__-->_|___|_-->_|___|__|_-->_|____|

A B C D E F G H I J K L M

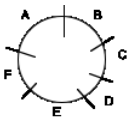
+ + + + + + + + + + + + - - top

- - - - - - - - - - - + + bot

a. Which restriction fragment(s) contain(s) the origin of replication?
b. Is replication from this origin uni- or bi-directional?
c. Explain how the data led you to your answers to a and b.
d. What direction is the replication fork moving for fragments A through K?
e. What direction is the replication fork moving for fragments L and M?
f. Name a possible target enzyme that could specifically block lagging strand synthesis when inhibited.
g. What cloning vector would be useful for generating the separated strands of the restriction fragments?

6.16

Let's imagine that you have isolated a new virus with a double-stranded, circular DNA that is 6000 bp long. The restriction endonuclease *Hha*I cleaves the DNA as shown below to generate 6 fragments.

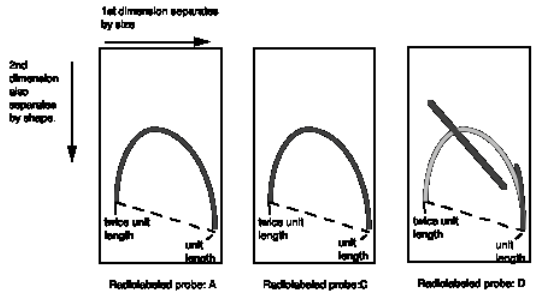


You initially use a pulse-labeling procedure to map the origin and terminus of replication. Infected cells were first allowed to incorporate [32P] phosphate into the DNA for several hours to uniformly label the DNA, and then [3H] thymidine was added for short periods of time (pulse labels), i.e. 5, 10 and 15 min. Completed viral DNA molecules were isolated, cut with *Hha*I, and separated on polyacrylamide gels. The amount of [32P] and [3H] in each fragment was determined for each period of pulse label and is tabulated below. The data are corrected for thymidine content and normalized so that fragment A has a ratio of 1.

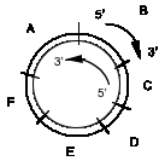
Relative amount of pulse label

Fragment	5 min	10 min	15 min
A	1.0	1.0	1.0
B	0.5	0.7	1.0
C	0	0.5	0.8
D	5.0	4.1	2.3
E	4.2	3.2	1.7
F	2.9	2.1	1.4

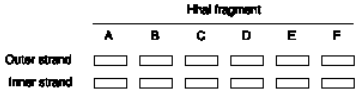
a. Which *Hha*I fragment(s) contain(s) the origin and terminus of replication?
b. What is the mode (uni- or bi-directional, or other) and direction(s) of replication (i.e. clockwise and/or counterclockwise)?
c. To confirm this result and map the origin and terminus more precisely, you analyzed the replicative intermediates on 2-dimensional gels. The DNA from infected cells, containing viral DNAs at all stages of synthesis, was digested with *Hha*I and then run initially on a gel that separates on the basis of size and then in a perpendicular direction in a gel that accentuates separations based on shape (Brewer and Fangman gels). The DNA in the gel was blotted onto a nylon membrane and hybridized with radiolabeled probes for the viral DNA fragments. The hybridization patterns obtained for *Hha*I fragments A, C and D are shown. The hypothetical line for linear intermediates of a fragment expanding from unit length to twice unit length is provided as a guide. How do you interpret these data, and what do you learn about the origin and terminus? Please indicate the significance of any transitions in the patterns.



d. (d) You also used a replication direction assay to examine the replication origin. Virally infected cells were treated with the drug emetine to inhibit lagging strand synthesis. Leading strand synthesis continued during the drug treatment, and this newly synthesized DNA was density labeled by incorporating 5-bromodeoxyuridylate (5-bromodeoxyuridine is added to the medium). The DNA was sheared and denatured, and the newly synthesized leading strand DNA was separated from the rest of the DNA by sedimentation equilibrium on Cs2SO4 gradients. Samples of the heavy density DNA (containing 5-bromodeoxyuridylate) were spotted onto a membrane, and equal amounts are hybridized to labeled, separated strands of restriction fragments throughout the virus. To keep track of strands and orientation in this problem, lets imagine the duplex circle to have an *outer*strand oriented 5' to 3' in a clockwise direction and an *inner*strand oriented 5' to 3' in a counterclockwise direction, as diagrammed below.



A grid of samples of heavy density DNA (containing 5-bromodeoxyuridylate, and enriched for leading strand DNA) immobilized on the filter is shown below, with each rectangle representing an equal loading of the heavy density DNA. What will be the pattern of hybridization to the indicated strands of each of the restriction fragments?



What does this experiment tell you about the origin and terminus of replication?

6.17

Are the following statements about the function of the DnaA protein true or false?

a. DnaA protein binds to 9-mer (nonamer) repeats at the origin for chromosomal replication.
b. DnaA protein catalyzes the formation of the primers for leading strand synthesis at the origin.
c. About 20 to 40 monomers of the DnaA protein form a large complex at the origin.
d. DnaA protein melts DNA at a series of 13-mer repeats at the origin.

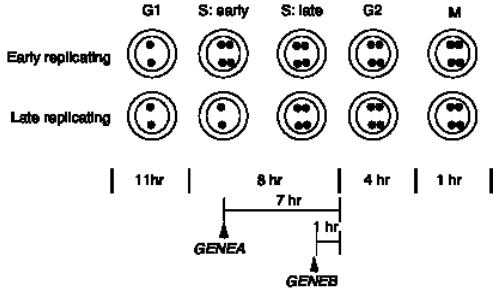
6.18

Consider a bacterium with a circular chromosome with one replication origin. It takes 30 min for bi-directional replication to copy its chromosome (the elongation time or C period) and 10 min from the end of DNA synthesis until the cell divides (the D period). How many replication forks are needed per chromosome to allow a culture of this bacterium to double in cell number every 20 min? Follow the molecules through a complete cell division cycle.

6.19

In many eukaryotes, actively transcribed genes are replicated early in S phase and inactive genes are replicated late. One assay to determine replication timing is *in situ* hybridization of cells with a gene-specific, fluorescent probe, followed by examination of the number of signals per nucleus. In diploid cells, an unreplicated gene will be seen as 2 fluorescent dots per nucleus, whereas a replicated gene will be seen as 4 dots. They look like 2 doublets, indicating that the replicated chromatids are close in the nucleus.

The types of pattern one can see at various stages of the cell cycle are shown below. Each dark dot is a fluorescent signal, the larger circle is the cell, and the smaller circle is the nucleus.



The fraction of cells in an asynchronous population with 2 dots or 4 dots is then tabulated. In an asynchronous population, the number of cells in each phase of the cell cycle is directly proportional to the length of that phase. If *GENEA*were replicated 1 hr after entry into S phase, and *GENEB*were replicated 1 hr before the end of S phase, what fraction of cells would show 4 dots (two doublets) for each? The length of each phase of the cell cycle is given in the figure, and the vertical arrowhead shows the time of synthesis. The time from synthesis of each gene until the beginning of G2 is shown above a horizontal line. Consider cells in M to have 4 dots (i.e., assume that the transition from 4 dots to 2 occurs at the M to G1 boundary).

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Control of initiation at oriC by methylation

A new round of replication will initiate on the *E. coli* chromosome at *oriC* only when the growth conditions permit it. The *dam* methylase and features of its sites of action are used to prevent premature re-initiation. The *dam* methylase of *E. coli* recognizes the tetranucleotide GATC in DNA and transfers a methyl group (from S-adenosyl methionine) to the amino group at position 6 of the adenine in that sequence. Note that GATC is a pseudopalindrome, so both strands read the same for these four nucleotides in DNA.

Palindrome and Pseudopalindromes

A nucleotide sequence is said to be a palindrome if it has an even number of base pairs and is equal to the reverse of its complementary sequence. For example, in a single strand of DNA the sequence of bases CCATTAATGG is palindromic because the sequence of bases in the complementary strand is GGTAATTACC, its reverse.

A pseudopalindrome is a DNA sequence with an odd number of base pairs yielding a symmetrical complement except at the central base-pair. For example, the DNA sequence ACCTGGT is pseudopalindromic, because its complement on the other strand is TGGACCA, which is its reverse except for the central element.

Thus a GATC in duplex DNA can be unmethylated on either strand, methylated on only one strand (referred to as **hemimethylated**) or methylated on both strands (referred to as **fully methylated**), as shown in Figure 1.

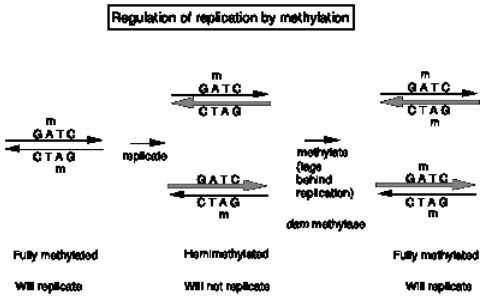


Figure 1

The methylation status of the 11 GATC motifs at *oriC*regulate whether replication can intitiate. When the GATCs are fully methylated, *oriC*DNA serves as an origin (in the presence of Dna A and the other proteins discussed above). However, when the GATCs are hemimethylated, it is not active as an origin. The reason for this is not fully known. One hint comes from the behavior of unmethylated *oriC*(from *dam*- strains). This unmethylated *oriC* is active, showing that methylation of the GATC is not a requirement for initiation, and further suggesting that some inhibitor of initiation recognizes the hemimethylated form.

1

How do these results lead to this conclusion? Let's explore this by posing the opposite hypotheses.

a. If methylation of the GATC motifs at *oriC* were needed for initiation, what would the result have been?
b. If some activator recognized the fully methylated form, what would the result have been.?

Re-methylation of *oriC*by the dam methylase is quite slow. Thus for some period the GATCs at *oriC*are hemimethylated, and the origin is inactive. This provides a means to delay the use of *oriC*to start another round of replication. Thus the methylation of the GATCs is part of a mechanism to regulate the timing of firing of *oriC*. In the next chapter, we will also see the use of methylation of GATCs in post-replicative repair.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)
- Edits attrbuted to Jeff Aronson

Linear Templates

The problem of linear templates

The requirement of DNA polymerases to have a primer causes a problem at the ends of linear templates. As illustrated in Figure 6.13, leading strand synthesis can proceed to the end of its template strand, but lagging strand synthesis cannot. As lagging strand synthesis nears the end of its template, at some point no binding site will be available for primase, and part of the 3’ end of the template for lagging strand synthesis will not be copied. Hence a 3’ overhang is left after the replication fork has finished, and part of the chromosome is not copied into new DNA. If nothing else were done, the chromosome would become progressively shorter after each round of replication.

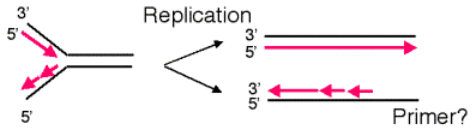


Figure 6.13: Lagging strand synthesis cannot copy the end of a linear chromosome.

At least three different types of solution to this problem have been discovered in various organisms. One, utilized by bacteriophage such as l and T4, is to convert the linear template to a circle. For instance, the linear chromosome of bacteriophage l has cohesive ends (complementary single strands at each end, generated by a phage endonuclease) that can anneal upon infection, thereby forming a *circular template* for replication. Other viruses, such as adenovirus, attach a protein to the end of unreplicated DNA to serve as a primer. Such an attached protein obviates the requirement for using the unreplicated DNA as a template, and the entire viral chromosome can be replicated.

A third solution is to make the ends a series of simple repeats that are synthesized in a process distinct from DNA replication. Indeed, the ends of the linear chromosomes of most (perhaps all) eukaryotes, called telomeres, are composed of many copies of a simple repetitive sequence. This sequence is distinctive for different organisms, but in all cases one strand is rich in G and the other is rich in C. The repeating unit for human telomeres is 5' AGGGTT 3' running from the centromeric end of the repeats to the telomeric end), and the repeating units for the ciliate *Tetrahymena* is 5' GGGGTT 3’.

New copies of the telomeric repeats can be synthesized each time the chromosome replicates (Figure 6.14). This re-synthesis of the telomeric repeats counteracts the progressive shortening of the linear chromosomes that would occur if only the replication forks were used to synthesize new chromosomes.

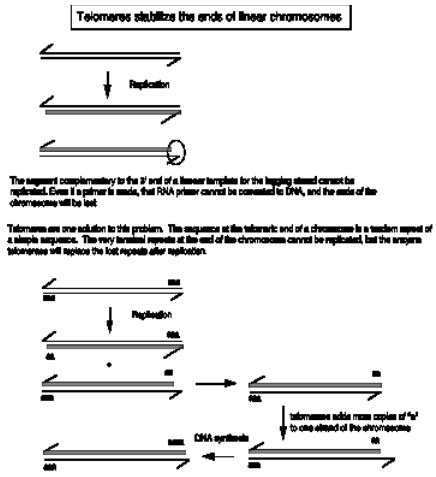


Figure 6.14. Addition of new telomeric repeats to the ends of replicated chromosomes.

In this figure, the string of "a" at the ends of the chromosome is the tandem repeat of simple sequence, in duplex form. For instance, for a human chromosome, "a" would be

CEN ... 5' AGGGTT 3' ... TEL

3' TCCCAA 5'

or for a Tetrahymena chromosome, "a" would be

CEN ... 5' GGGGTT 3' ... TEL

3' CCCCAA 5'

In each case, the "a" or monomer is repeated thousands of times in tandem.

Addition of new telomeric repeats is catalyzed by the enzyme **telomerase**. As illustrated in Figure 6.15, this enzyme catalyzes many successive rounds of synthesis, adding many copies of the simple repeat to the ends of the chromosomes. The enzyme is a ribonucleoprotein, i.e. it has both a polypeptide and an RNA component. The RNA serves as a template to direct addition of nucleotides to the 3' end of the G+T rich strand, and the polypeptide acts as a reverse transcriptase to make a DNA copy of a hexanuclotide segment of the RNA. For instance, the telomerase from *Tetrahymena* will copy the 3'CCCCAA in the RNA template into 5'GGGGTT telomeric repeat. Then the enzyme shifts over and synthesizes another hexanucleotide. The fact that the RNA serves as the template was demonstrated by exchanging the RNA component of isolated telomerase with the telomerase RNA from a second species. This exchange led to the addition of telomeres with sequences characteristic of that of the second species, showing that the telomerase RNA is the determinant of the sequence of the telomere. The protein component provides the reverse transcriptase activity.

Once many copies of the G+T-rich strand of telomeres have been synthesized by telomerase, the long single strand forms a specialized structure toward the 3' end. Some evidence indicates that a "G-quartet" is formed, in which four guanine nucleotides form a hydrogen-bonded complex. Examination of the ends of replicating chromosomes in the electron microscope show a circular structure. Although details of the structure at the end of this strand are not fully established, it is likely that a primer to support synthesis of the C+A-rich strand is made effectively by turning the G+T-rich strand around. Conventional synthesis by DNA polymerases can then copy the G+T-rich strand to make the complementary strand. Some processing, e.g. nucleases acting at the end, can convert the specialized structure or hairpin into a linear duplex.

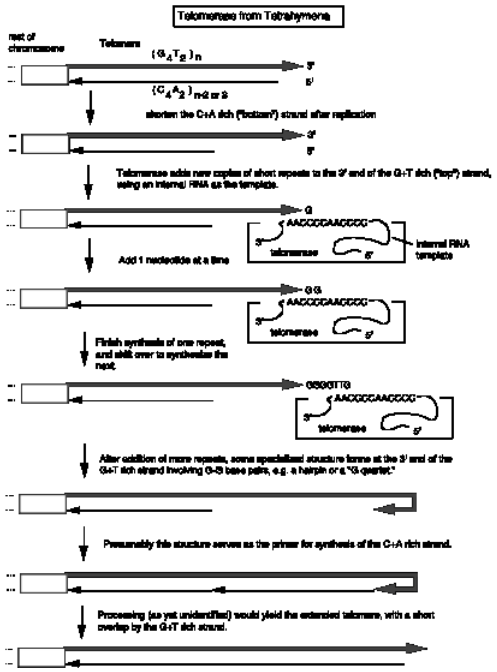


Figure 6.15. Synthesis of new telomeric repeats catalyzed by telomerase. This enzyme is a ribonucleoprotein complex. The RNA component is the template for synthesis of telomeric repeats.

How processive is telomerase?

Not all replicating cells have telomerase activity. This activity is higher in some transformed cells than in nontransformed cells. Also, older cells tend to have shorter telomeres. Thus telomeres are being actively investigated as possibly playing roles in both aging and in tumorigenic transformation.

Telomeres are important for stabilizing chromosomes. Some chromosomal deletions remove the ends of the chromosome, including the telomere, and these shortened chromosomes are less stable than their wild-type counterparts. Directed mutations have

been made in mice to eliminate telomerase activity. These mice are viable for several generations, but they show many broken and abnormal chromosomes, demonstrating the importance of this activity.

Replication in Bacteria

Cellular control of replication in bacteria

We have seen that the initiator protein DnaA and the replicator element *oriC*are needed for the initiation of replication, and that the slow rate of methylation at GATC motifs prevents re-initiation for some time. The bacterial cell can sense when the nutritional conditions, levels of nucleotide pools, and protein concentrations are adequate to support a round of replication. The details of this monitoring are beyond the scope of this presentation, and can be explored in references such as Niedhart et al. In general, initiation is triggered by the increase in cell mass. Initiation occurs at a constant ratio of cell mass to the number of origins. This suggests that a mechanism exists to titrate out some regulatory molecule as the cell mass increases, but the molecule and mechanism have not been elucidated.

The result of this monitoring and signalling is the formation of an active DnaA complex at *oriC*, followed by unwinding the DNA and the other events discussed above.

Depending on the growth conditions, bacteria can divide rapidly or slowly. In rich media, the cell number can double every 18 min, whereas when nutrients are scare, the doubling time can be long as 180 min. The bacterial cells accomplish this by varying the rate of re-initiation of replication. Re-initiation has to occur at the same frequency as the cell doubling time.

Although the frequency of re-initiation can be varied 10-fold, the time required for the replication cycle is constant. This cycle consists of two periods called C and D. The **elongation time**, or C period, is the time required to replicate the bacterial chromosome. From initiation to termination, this is about 40 min. The **division time**, or D period, is the time that elapses between completion of a round of DNA replication and completion of cell division. This is about 20 min. Hence the time for the replication cycle (C period plus D period) is essentially constant in bacterial cultures with doubling times shorter than 60 min.

In order to accommodate the variation in cell doubling time within the constraints of the constant time for replication (C+D), rapidly growing bacteria have chromosomes with multiple replication forks. The constant replication cycle time means that a round of replication must be initiated 60 min (i.e. C+D) before cell division. However, re-initiation can occur before 60 min has past. This is illustrated in Figure 6.16 for cells in a culture dividing every 30 min. When the cell doubling time is less than 60 min, a cycle of replication must initiate before the end of the preceding cycle. This results in chromosomes with more than one replication fork.

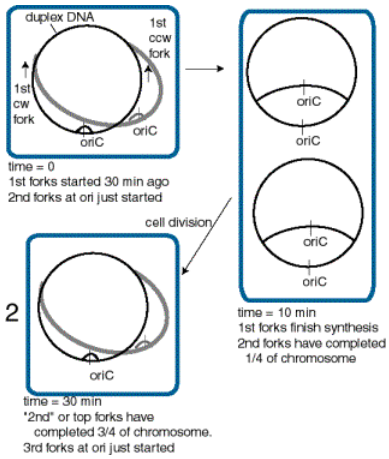


Figure 6.16. Multiple replication forks per chromosome allow bacteria to divide more rapidly than the replication cycle time. This diagram illustrates a bacterial cell dividing every 30 min, and hence initiating a new cycle of replication every 30 min.

Exercise 1

If the time required for two replication forks traveling in opposite directions to traverse the entire *E. coli* chromosome at 37oC is about 40 min, regardless of the culture conditions and the time required for cell division (D period) is 20 min, how many replication forks will be present on each DNA molecule in the culture?

Answer

TBA

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Replication in Eukaryotes

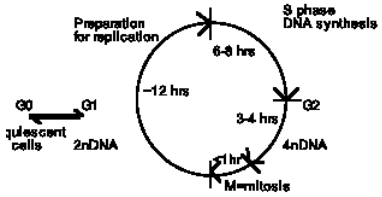


Figure 6.17. The eukaryotic cell cycle

Passage from one phase to the next is a highly regulated event. Critical control points, or checkpoints, are found at the G1 to S transition and at the G2 to M transition. The checkpoint in late G1 is the time for the cell to assess whether it has enough nucleotides, proteins and other materials to make two cells. The checkpoint prior to the G2/M transition allows any necessary repairs or corrections in the DNA to be made prior to mitosis. Loss of control of the G0 to G1 transition, or at the other checkpoints, generates cells that grow in an uncontrolled manner. This inappropriate expansion in the number of cells is fundamental to the progression of cancers, and hence the study of the molecular events at these checkpoints is an intensely active area of research in cell biology and biochemistry. A full treatment of this important topic is beyond the scope of this course. In general, cell cycle progression is regulated by environmental signals (such as extracellular growth factors) and intracellular monitors of metabolic state, intactness of DNA, and so forth. These disparate signals eventually impinge on highly regulated protein kinases. Activation of particular protein kinases is required for progression through each checkpoint. In general, two types of regulation have been seen.

1. Control of the amount of key proteins. The concentration of proteins called cyclins rise and fall through the cell cycle. Some of the cyclins are components of protein kinases whose activity regulates passage through the checkpoints. The cyclins must be present at a sufficiently high concentration for the kinase to be active.
2. Control of the state of phosphorylation. Proteins regulating the cell cycle (as is true of many regulatory proteins) can be covalently modified, e.g. by phosphorylation in a process catalyzed by protein kinases. The state of phosphorylation will determine the level of activity of the protein. So for instance a key protein kinase regulating passage through the G1 to S checkpoint must have its catalytic subunit in the correct state of phosphorylation, as well as having sufficient amounts of its cyclin subunit.

Many lines of investigation are being pursued to understand better the regulation of the cell cycle. One fundamental approach has been be isolation of scores of conditional yeast mutants that are defective in their progression through the cell cycle at the restrictive temperature. These mutants have particular phenotypes depending on which stage of the cell cycle they arrest in under nonpermissive conditions. The complementation groups defined by such mutants are called *CDC*, for cell division cycle phenotypes, followed by a number. For example, a protein kinase whose activity is needed for both the G1/S and the G2/M transition in *S. cerevisiae*is the product of the *CDC28* gene, and the polypeptide is called Cdc28p.

Once a cell has entered S phase, each origin of replication must fire once, but only once. As discussed above, the ORC is required for initiation of replication at an origin, but what determines when the origin fires? This is a matter of considerable current study, and many of the details are still unknown. In *S. cerevisiae*, the ORC binds to specific DNA sequences, the origins of replication, throughout the cell cycle, not just during S phase when the origins are active. During G1 phase, ORC recruits other proteins, such as Cdc6 and Mcm (minichromosome maintenace) proteins, to form a **prereplication complex**. At the G1/S transition, additional factors associate with this complex, and a cyclin-dependent kinase (CDK) activity stimulates intitiation of replication in S phase. After initiation, the Cdc6 and Mcm proteins are released from the prereplication complex, leaving the ORC still bound to the origin but unable to reinitiate replication until the next cell cycle. In mammals, an intact ORC is not stably bound to the origin, but rather one of the subunit, ORC1, is recruited to the origin at a defined time during G1. However, in both yeast and mammals, events in G1 involving the preinitiation complex mark an origin for firing in the next S phase.

As discussed above, many origins of replication, and hence many replicons, are used to replicate each chromosome. These origins do not all fire at the same time. In fact, replicons can initiate at different times during S phase. Replicons containing genes that are actively expressed in a given cells tend to replicate earlier in S phase than do replicons containing nonexpressed genes. This is an example of tissue-specific variation in replication timing. The time during S phase at which a particular origin will fire is determined early in G1, at the time that chromatin domains are repositioned in the nucleus following mitosis and before the

preinitiation complex forms. Events important to the regulation of initiation at replication origins occur at various times during G1, but the full range of proteins and activities carrying out these events is still a matter of study.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Replication in Yeast

Stages of DNA synthesis

The synthesis of any macromolecule proceeds in three stages: **initiation**, **elongation** and **termination**. This is true for DNA replication as well. During initiation, DNA synthesis begins at a specific site, called an **origin of replication**. The circular *E. coli* chromosome has a single origin, called *oriC*. Many bacteria have circular chromosomes with single origins of replication. However, other chromosomes, especially those in eukaryotes, can have multiple origins. During elongation, nucleotides are added to the growing DNA strand as the replication fork moves along the chromosome. Termination are the final steps that occur when all or an appropriate portion (replicon, see below) of the chromosome has been replicated.

The primary control of replication is exerted during initiation. This is economical, of course, since little benefit would come from initiating replication that will never be completed. As will be covered later in this chapter, an examination of the DNA structures, proteins and enzymes needed for initiation show that it is highly regulated. Initiation is an active process, requiring the accumulation of ATP-bound DNA binding proteins at a specific site prior to the start of replication. Both the activity of the initiator proteins and the state of covalent modification of the DNA at the origin are part of the control process.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

CHAPTER OVERVIEW

7: Mutation and Repair of DNA

Most biological molecules have a limited lifetime. Many proteins, lipids and RNAs are degraded when they are no longer needed or damaged, and smaller molecules such as sugars are metabolized to compounds to make or store energy. In contrast, DNA is the most stable biological molecule known, befitting its role in storage of genetic information. The DNA is passed from one generation to another, and it is degraded only when cells die. However, it can change, i.e. it is mutable. **Mutations**, or changes in the nucleotide sequence, can result from errors during DNA replication, from covalent changes in structure because of reaction with chemical or physical agents in the environment, or from transposition. Most of the sequence alterations are **repaired** in cells. Some of the major avenues for changing DNA sequences and repairing those mutations will be discussed in this chapter.

Topic hierarchy

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

7.0: Prelude to Mutations

Sequence alteration in the genomic DNA is the fuel driving the course of evolution. Without such mutations, no changes would occur in populations of species to allow them to adapt to changes in the environment. Mutations in the DNA of germline cells fall into three categories with respect to their impact on evolution. Most have no effect on phenotype; these include sequence changes in the large portion of the genome that neither codes for protein, or is involved in gene regulation or any other process. Some of these **neutral** mutations will become prevalent in a population of organisms (or **fixed**) over long periods of time by stochastic processes. Other mutations do have a phenotype, one that is advantageous to the individuals carrying it. These mutations are fixed in populations rapidly (i.e. they are subject to **positive selection**). Other mutations have a detrimental phenotype, and these are cleared from the population quickly. They are subject to **negative** or **purifying selection**.

A red Darwin hybrid tulip "Apeldoorn" with a mutation that resulted in half of a petal being yellow. (CC BY-SA 3.0; LepoRello).

Whether a mutation is neutral, disadvantageous or useful is determined by where it is in the genome, what the type of change is, and the particulars of the environmental forces operating on the locus. For our purposes, it is important to realize that sequence changes are a natural part of DNA metabolism. However, the amount and types of mutations that accumulate in a genome are determined by the types and concentrations of mutagens to which a cell or organism is exposed, the efficiency of relevant repair processes, and the effect on phenotype in the organism.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

7.1: Mutations and Mutagens

Mutations commonly are *substitutions*, in which a single nucleotide is changed into a different nucleotide. Other mutations result in the loss (*deletion*) or addition (*insertion*) of one or more nucleotides. These insertions or deletions can range from one to tens of thousands of nucleotides. Often an insertion or deletion is inferred from comparison of two homologous sequences, and it may be impossible to ascertain from the data given whether the presence of a segment in one sequence but not another resulted from an insertion of a deletion (in this case, it can be referred to as an "*indel*"). One mechanism for large insertions is the transposition of a sequence from one place in a genome to another (described in Chapter 9).

Types of muations

Nucleotide substitutions are one of two classes. In a **transition**, a purine nucleotide is replaced with a purine nucleotide, or a pyrimidine nucleotide is replaced with a pyrimidine nucleotide. In other words, the base in the new nucleotide is in the same chemical class as that of the original nucleotide. In a **transversion**, the chemical class of the base changes, i.e. a purine nucleotide is replaced with a pyrimidine nucleotide, or a pyrimidine nucleotide is replaced with a purine nucleotide.

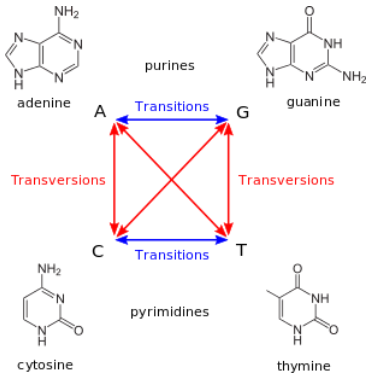


Figure 7.1.1: Diagram of the types of substitutions: transitions and transversions. (CC BY-SA 3.0; Petulda).

Comparison of the sequences of homologous genes between species reveals a pronounced preference for transitions over transversions (about 10-fold), indicating that transitions occur much more frequently than transversions.

Errors in Replication

Despite effective proofreading functions in many DNA polymerases, occasionally the wrong nucleotide is incorporated. It is estimated that *E. coli*DNA polymerase III holoenzyme (with a fully functional proofreading activity) uses the wrong nucleotide during elongation about 1 in 10^8 times. It is more likely for an incorrect pyrimidine nucleotide to be incorporated opposite a purine nucleotide in the template strand, and for a purine nucleotide to be incorporated opposite a pyrimidine nucleotide. Thus these misincorporations resulting in a transition substitution are more common. However, incorporation of a pyrimidine nucleotide opposite another pyrimidine nucleotide, or a purine nucleotide opposite another purine nucleotide, can occur, albeit at progressively lower frequencies. These rarer misincorporations lead to transversions.

Exercise 7.1.1

If a dCTP is incorporated into a growing DNA strand opposite an A in the template strand, what mutation will result? Is it a transition or a transversion?

Exercise 7.1.2

If a dCTP is incorporated into a growing DNA strand opposite a T in the template strand, what mutation will result? Is it a transition or a transversion?

A change in the isomeric form of a purine or pyrimidine base in a nucleotide can result in a mutation. The base-pairing rules are based on the hydrogen-bonding capacity of nucleotides with their bases in the *keto*tautomer. A nucleotide whose base is in the *enol*tautomer can pair with the "wrong" base in another nucleotide. For example, a T in the rare *enol*isomer will pair with a *keto* G (Figure 7.1.2), and an *enol*G will pair with a *keto*T.

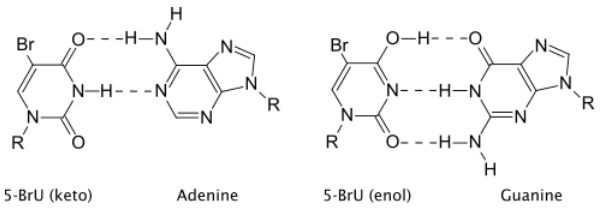


Figure 7.1.2: Illustration of the nucleoside enol 5-bromodeoxyuridine (or 5-BrdU, an analog of thymidine) paired with the nucleoside keto deoxyguanidine. 5-BrdU shifts into the enoltautomer more readily than thymidine does. (CC BY-SA 3.0; Histidine).

The *enol*tautomers of the normal deoxynucleotides guanidylate and thymidylate are rare, meaning that a single molecule is in the *keto*form most of the time, or within a population of molecules, most of them are in the *keto*form. However, certain nucleoside and base analogs adopt these alternative isomers more readily. For instance 5-bromo-deoxyuridine (or 5-BrdU) is an analog of deoxythymidine (dT) that is in the *enol*tautomer more frequently than dT is (although most of the time it is in the *keto*tautomer).

Thus the frequency of misincorporation can be increased by growth in the presence of base and nucleoside analogs. For example, growth in the presence of 5-BrdU results in an increase in the incorporation of G opposite a T in the DNA, as illustrated in Figure 7.1.3. After cells take up the nucleoside 5-BrdU, it is converted to 5-BrdUTP by nucleotide salvage enzymes that add phosphates to its 5' end. During replication, 5-BrdUTP (in the *keto*tautomer) will incorporate opposite an A in DNA. The 5-BrdU can shift into the *enol*form while in DNA, so that when it serves as a template during the next round of replication (arrow 1 in the diagram below), it will direct incorporation of a G in the complementary strand. This G will in turn direct incorporation of a C in the top strand in the next round of replication (arrow 2). This leaves a C:G base pair where there was a T:A base pair in the parental DNA. Once the pyrimidine shifts back to the favored *keto*tautomer, it can direct incorporation of an A, to give the second product in the diagram below (with a BrU-A base pair).

Exercise 7.1.3

Where are the hydrogen bonds in a base pair between *enol* –guanidine and *keto*-thymidine in DNA?

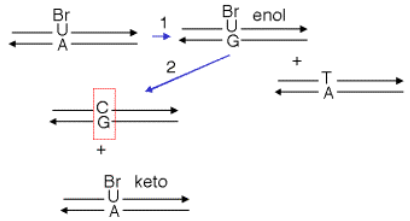


Figure 7.1.3: Replication of a misincorporated nucleotide (or nucleotide analog) will leave a mutation.

Likewise, misincorporation of A and C can occur when they are in the rare *imino*tautomers rather than the favored *amino*tautomers. In particular, *imino*C will pair with *amino*A, and *imino*A will pair with *amino*C (Figure 7.1.4).

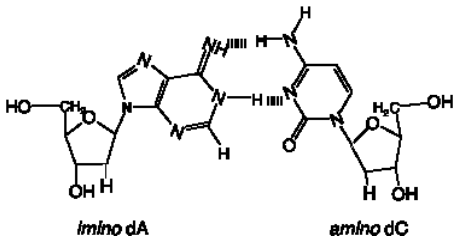


Figure 7.1.4. An A in the rare imino tautomer will pair with amino C. This can cause an A:T to G:C transition.

Misincorporation during replication is the major pathway for introducing *transversions* into DNA. Normally, DNA is a series of purine:pyrimidine base pairs, but in order to have a transversion, a pyrimidine has to be paired with another pyrimidine, or a purine with a purine. The DNA has to undergo local structural changes to accommodate these unusual base pairs. One way this can happen for a purine-purine base pair is for one of the purine nucleotides to shift from the preferred *anti*conformation to the *syn*conformation. Atoms on the "back side" of the purine nucleotide in the *syn*-isomer can form hydrogen bonds with atoms in the rare tautomer of the purine nucleotide, still in the preferred *anti*conformation. For example, an A nucleotide in the *syn*-, *amino*-

isomer can pair with an A nucleotide in the *anti*-, *imino*- form (Figure 7.1.5). Thus the transversion required a shift in the tautomeric form of the base in one nucleotide as well as a change in the base-sugar conformation (*anti*to *syn*) of the other nucleotide.

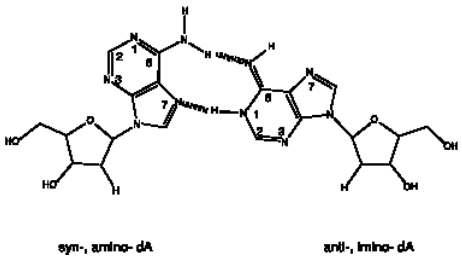


Figure 7.1.5: A base pair between a *syn*-, *amino*- isomer of A and the *anti*-, *imino*- form of A.

Exercise 7.1.4

Why does the shift of a purine nucleotide from *anti* to *syn*help allow a purine:purine base pair? Is this needed for a pyrimidine:pyrimidine base pair?

Errors in replication are not limited to substitutions. **Slippage errors** during replication will add or delete nucleotides. A DNA polymerase can insert additional nucleotides, more commonly when tandem short repeats are the template (e.g. repeating CA dinucleotides). Sometimes the template strand can loop out and form a secondary structure that the DNA polymerase does not read. In this case, a deletion in the nascent strand will result.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

7.2: Reaction with Mutagens

Many mutations do not result from errors in replication. Chemical reagents can oxidize and alkylate the bases in DNA, sometimes changing their base-pairing properties. Radiation can also damage DNA. Examples of these mutagenic reactions will be discussed in this section.

Chemical Modification by Oxidation

When the amino bases, adenine and cytosine, are oxidized, they also lose an amino group. Thus the amine is replaced by a keto group in the product of this oxidative deamination reaction. For instance, oxidation of cytosine produces uracil, which base pairs with adenine (shown for deoxycytidine in Figure 7.6). Likewise, oxidation of adenine yields hypoxanthine, which base pairs with cytosine (Figure 7.7.A). Thus the products of these chemical reactions will be mutations in the DNA, if not repaired. Oxidation of guanine yields xanthine (Figure 7.7.B). In DNA, xanthine will pair with cytosine, as does the original guanine, so this particular alteration is not mutagenic.

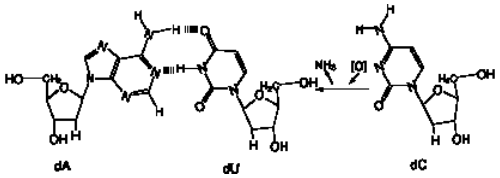


Figure 7.6. Oxidative deamination of deoxycytidine yields deoxyuridine. The deoxyuridine in DNA would direct pairing with dA after replication.

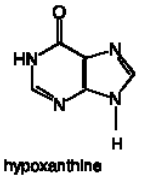


Figure 7.7.A.Structure of hypoxanthine, the product of oxidation deamination of adenine.

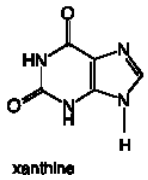


Figure 7.7.B.Structure of xanthine, the product of oxidative deamination of guanine.

Exercise

Both hypoxanthine and xanthine can base pair with cytosine in DNA. Why is this?

Oxidation of C to U occurs spontaneously at a high rate. The frequency is such that 1 in 1000 Cs in the human genome would become Us during a lifetime, if they were not repaired. As will be discussed later, repair mechanisms have evolved to replace a U in DNA with a T.

Methylation of C prior to its oxidative deamination will effectively mask it from the repair processes to remove U's from DNA. This has a substantial impact on the genomes of organisms that methylate C. In many eukaryotes, including vertebrates and plants (but not yeast or *Drosophila*), the principal DNA methyl transferase recognizes the dinucleotide CpG in DNA as the substrate, forming 5-methyl-CpG (Figure 7.8). When 5-methyl cytosine undergoes oxidative deamination, the result is 5-methyl uracil, which is the same as thymine. The surveillance system that recognizes U's in DNA does nothing to the T, since it is a normal component of DNA. Hence the oxidation of 5-methyl CpG to TpG, followed by a round of replication, results in a C:G to T:A transition at former CpG sites (Figure 7.8). This spontaneous deamination is quite frequent; indeed, C to T transitions at CpG dinucleotides are the most common mutations in humans. Since this transition is not repaired, over time the number of CpG dinucleotides is greatly diminished in the genomes of vertebrates and plants.

Me

--CG-- --CG-- [O] --TG-- + NH3 --**T**G--

|||||| ® |||||| ® ||o||| ® |||||| +wt

--GC-- --GC-- --GC-- --AC--

Methyl- Replicate

transferase mutation

Figure 7.8. Methylation of CpG dinucleotides followed by oxidative deamination results in TpG dinucleotides.

Some regions of plant and vertebrate genomes do not show the usual depletion of CpG dinucleotides. Instead, the frequency of CpG approaches that of GpC or the frequency expected from the individual frequency of G and C in the genome. One working definition of these **CpG islands**is that they are segments of genomic DNA at least 100 bp long with a CpG to GpC ratio of at least 0.6. These islands can be even longer and have a CpG/GpC > 0.75. They are distinctive regions of these genomes and are often found in promoters and other regulatory regions of genes. Examination of several of these CpG islands has shown that they are not methylated in any tissue, unlike most of the other CpGs in the genome. Current areas of research include investigating how the CpG islands escape methylation and their role in regulation of gene expression.

Exercise

If a CpG island were to be methylated in the germ line, what would be consequences be over many generations?

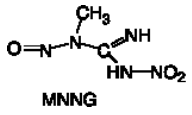
The rate of oxidation of bases in DNA can be increased by treating with appropriate reagents, such as nitrous acid (HNO_2). Thus treatment with nitrous acid will increase the oxidation of C to U, and hence lead to C:G to T:A transitions in DNA. It will also increase the oxidation of adenine to hypoxanthine, leading to A:T to G:C transitions in DNA.

Chemical Modification by Alkylation

Many mutagens are **alkylating agents**. This means that they will add an alkyl group, such as methyl or ethyl, to a base in DNA. Examples of commonly used alkylating agents in laboratory work are N-methyl-nitrosoguanidine and N-methyl-N'-nitro-nitrosoguanidine (MNNG, Figure 7.9.A.). The chemical warfare agents sulfur mustard and nitrogen mustard are also alkylating agents.

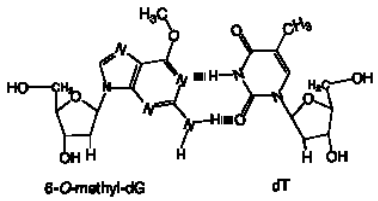
N-methyl-nitrosoguanidine and MNNG transfer a methyl group to guanine (e.g. to the O^6 position) and other bases (e.g. forming 3-methyladenine from adenine). The additional methyl (or other alkyl group) causes a distortion in the helix. The distorted helix can alter the base pairing properties. For instance, O6-methylguanine will sometimes base pair with thymine (Figure 7.9.B.).

A. N-methyl-N'-nitro-N-nitrosoguanidine (MNNG)



MNNG

B. 6-O-methyl-G will pair with T



6-O-methyl-dG dT

Figure 7.9.A. Structure of MNNG and the base pair between O6-methyl G and T

The order of reactivity of nucleophilic centers in purines follows roughly this series:

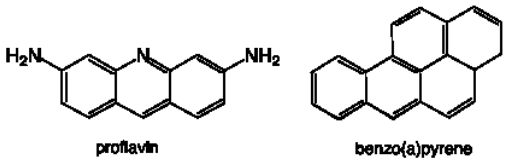
N7-G >> N3-A > N1-A @ N3-G @ O6-G.

A common laboratory reagent for purines in DNA is dimethylsulfate, or DMS. The products of this reaction are primarily N^7-guanine, but N^3-adenine is also detectable. This reaction is used to identify protein-binding sites in DNA, since interaction with a protein can cause decreased reactivity to DMS of guanines within the binding site but enhanced reactivity adjacent to the site. Methylation to form N7-methyl-guanine does not cause miscoding in the DNA, since this modified purine still pairs with C.

Chemicals that Cause Deletions

Some compounds cause a loss of nucleotides from DNA. If these deletions occur in a protein-coding region of the genomic DNA, and are not an integral multiple of 3, they result in a frameshift mutation. These are commonly more severe loss-of-function mutations than are simple substitutions. Frameshift mutagens such as proflavin or ethidium bromide have flat, polycyclic ring structures (Figure 7.10.A.). They may bind to and **intercalate** within the DNA, i.e. they can insert between stacked base pairs. If a segment of the template DNA is the looped out, DNA polymerase can replicate past it, thereby generating a deletion. Intercalating agents can stabilize secondary structures in the loop, thereby increasing the chance that this segment stays in the loop and is not copied during replication (Figure 7.10.B.) Thus growth of cells in the presence of such intercalating agents increase the probability of generating a deletion.

A.



B.

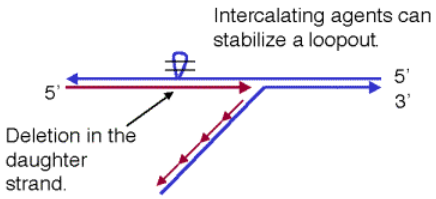


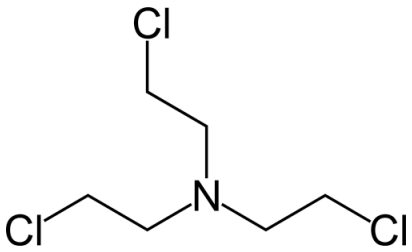
Figure 7.10. Two intercalating agents (A) and their ability to stabilize loops in the template, leading to deletions in the nascent DNA strand (B). Benz(a)pyrenes are present in soot.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Nitrogen Mustard

Nitrogen mustards were produced in the 1920s and 1930s as potential chemical warfare weapons and are similar to sulfur mustard.



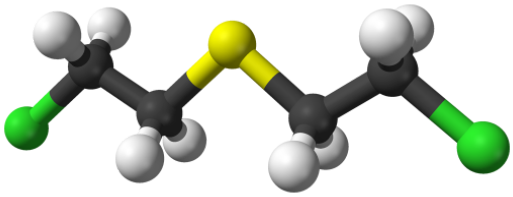
Skeletal formula of tris(2-chloroethyl)amine. Image used wtih permission (Public Domain; Benjah-bmm27).

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Sulfur Mustard

Sulfur mustard is a type of chemical warfare agent. These kinds of agents cause blistering of the skin and mucous membranes on contact. They are called vesicants or blistering agents. Sulfur mustard is also known as “mustard gas or mustard agent,” or by the military designations H, HD, and HT. Sulfur mustard sometimes smells like garlic, onions, or mustard and sometimes has no odor. It can be a vapor (the gaseous form of a liquid), an oily-textured liquid, or a solid.



Ball-and-stick model of the sulfur mustard molecule. (Public Domain; Ben Mills).

Reactions

Sulfur mustard is the organic compound with formula $(\mathrm{Cl{-}CH_2CH_2)_2S}$. In the Depretz method, sulfur mustard is synthesized by treating sulfur dichloride with ethylene:

$$\mathrm{SCl_2 + 2\,C_2H_4 \rightarrow (Cl{-}CH_2CH_2)_2S} \tag{1}$$

In the Levinstein process, sulfur monochloride is used instead:

$$\mathrm{8\,S_2Cl_2 + 16\,C_2H_4 \rightarrow 8\,(Cl{-}CH_2CH_2)_2S + S_8} \tag{2}$$

In the Meyer method, thiodiglycol is produced from chloroethanol and potassium sulfide and chlorinated with phosphorus trichloride:

$$\mathrm{3\,(HO{-}CH_2CH_2)_2S + 2\,PCl_3 \rightarrow 3\,(Cl{-}CH_2CH_2)_2S + 2\,P(OH)_3} \tag{3}$$

In the Meyer-Clarke method, concentrated hydrochloric acid (HCl) instead of PCl_3 is used as the chlorinating agent:

$$\mathrm{(HO{-}CH_2CH_2)_2S + 2\,HCl \rightarrow (Cl{-}CH_2CH_2)_2S + 2\,H_2O} \tag{4}$$

Thionyl chloride and phosgene have also been used as chlorinating agents. It is a viscous liquid at normal temperatures. The pure compound has a melting point of 14 °C (57 °F) and decomposes before boiling at 218 °C (424.4 °F).

References

1. Stewart, Charles D. (2006). *Weapons of mass casualties and terrorism response handbook*. Boston: Jones and Bartlett. p. 47. ISBN 0-7637-2425-4.
2. "Chemical Weapons Production and Storage". Federation of American Scientists.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

7.3: Ionizing Radiation

High energy radiation, such as X-rays, γ-rays, and β particles (or electrons) are powerful mutagens. Since they can change the number of electrons on an atom, converting a compound to an ionized form, they are referred to as **ionizing radiation**. They can cause a number of chemical changes in DNA, including directly break phosphodiester backbone of DNA, leading to deletions. Ionizing radiation can also break open the imidazole ring of purines. Subsequent removal of the damaged purine from DNA by a glycosylase generates an apurinic site.

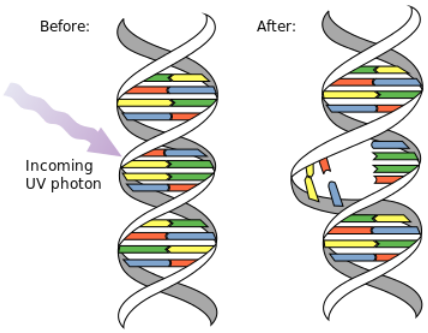


Figure 7.3.1: *Formation of thymine dimer lesion in DNA. The photon causes two consecutive bases on one strand to bind together, destroying the normal base-pairing double-strand structure in that area. Ultraviolet (UV) photons harm the DNA molecules of living organisms in different ways. In one common damage event, adjacent bases bond with each other, instead of across the "ladder." This makes a bulge, and the distorted DNA molecule does not function properly. (Public Domain; Master Uegly).*

Ultraviolet Radiation

Ultraviolet radiation with a wavelength of 260 nm will form *pyrimidine dimers* between adjacent pyrimidines in the DNA. The dimers can be one of two types (Figure 7.11). The major product is a cytobutane-containing thymine dimer (between C5 and C6 of adjacent T's). The other product has a covalent bond between position 6 on one pyrimidine and position 4 on the adjacent pyrimidine, hence it is called the "6-4" photoproduct.

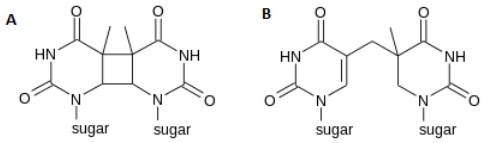


Figure 7.3.2: *Pyrimidine dimers formed by UV radiation, illustrated for adjacent thymidylates on one strand of the DNA. (A) Formation of a covalent bond between the C atoms at position 5 of each pyrimidine and between the C atoms at position 6 of each pyrimidine makes a cyclobutane ring connecting the two pyrimidines. The bases are stacked over each other, held in place by the cyclobutane ring. The C-C bonds between the pyrimidines are exaggerated in this drawing so that the pyrimidine ring is visible. (B) Another photoproduct is made by forming a bond between the C atom at position 6 of one pyrimidine and position 4 of the adjacent pyrimidine, with loss of the O previously attached at position 4. (Public Domain; Master Uegly).*

The pyrimidine dimers cause a distortion in the DNA double helix. This distortion blocks replication and transcription.

Exercise 7.3.1

What is the physical basis for this distortion in the DNA double helix?

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

7.4: Repair Mechanisms

The second part of this chapter examines the major classes of DNA repair processes. These are:

- reversal of damage,
- nucleotide excision repair,
- base excision repair,
- mismatch repair,
- recombinational repair, and
- error-prone repair.

Many of these processes were first studies in bacteria such as *E. coli*, however only a few are limited to this species. For instance, nucleotide excision repair and base excision repair are found in virtually all organisms, and they have been well characterized in bacteria, yeast, and mammals. Like DNA replication itself, repair of damage and misincorporation is a very old process.

Reversal of damage

Some kinds of covalent alteration to bases in DNA can be directly reversed. This occurs by specific enzyme systems recognizing the altered base and breaking bonds to remove the adduct or change the base back to its normal structure.

Photoreactivation is a light-dependent process used by bacteria to reverse pyrimidine dimers formed by UV radiation. The enzyme photolyase binds to a pyrimidine dimer and catalyzes a second photochemical reaction (this time using visible light) that breaks the cyclobutane ring and reforms the two adjacent thymidylates in DNA. Note that this is not formally the reverse of the reaction that formed the pyrimidine dimers, since energy from visible light is used to break the bonds between the pyrimidines, and no UV radiation is released. However, the result is that the DNA structure has been returned to its state prior to damage by UV. The photolyase enzyme has two subunits, which are encoded by the *phrA* and *phrB*genes in *E. coli*.

A second example of the reversal of damage is the **removal of methyl groups**. For instance, the enzyme *O6*-methylguanine methyltransferase, encoded by the *ada*gene in *E. coli*, recognizes *O6*-methylguanine in duplex DNA. It then removes the methyl group, transferring it to an amino acid of the enzyme. The methylated enzyme is no longer active, hence this has been referred to as a suicide mechanism for the enzyme.

Excision Repair

The most common means of repairing damage or a mismatch is to cut it out of the duplex DNA and recopy the remaining complementary strand of DNA, as outlined in Figure 7.12. Three different types of excision repair have been characterized: nucleotide excision repair, base excision repair, and mismatch repair. All utilize a **cut, copy, and paste** mechanism. In the **cutting**stage, an enzyme or complex removes a damaged base or a string of nucleotides from the DNA. For the **copying**, a DNA polymerase (DNA polymerase I in *E. coli*) will copy the template to replace the excised, damaged strand. The DNA polymerase can initiate synthesis from 3' OH at the single-strand break (nick) or gap in the DNA remaining at the site of damage after excision. Finally, in the **pasting**stage, DNA ligase seals the remaining nick to give an intact, repaired DNA.

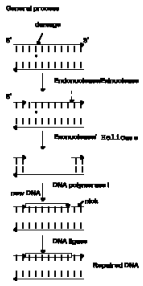


Figure 7.12. A general scheme for excision repair, illustrating the cut (steps 1 and 2), copy (step 3) and paste (step 4) mechanism.

Nucleotide Excision Repair (NER)

In *nucleotide excision repair*, damaged bases are cut out within a string of nucleotides, and replaced with DNA as directed by the undamaged template strand. This repair system is used to remove pyrimidine dimers formed by UV radiation as well as nucleotides modified by bulky chemical adducts. The common feature of damage that is repaired by nucleotide excision is that the modified nucleotides cause a significant distortion in the DNA helix. NER occurs in almost all organisms examined.

Some of the best-characterized enzymes catalyzing this process are the UvrABC excinuclease and the UvrD helicase in *E. coli*. The genes encoding this repair function were discovered as mutants that are highly sensitive to UV damage, indicating that the mutants are defective in UV repair. As illustrated in Figure 7.13, wild type *E. coli* cells are killed only at higher doses of UV radiation. Mutant strains can be identified that are substantially more sensitive to UV radiation; these are defective in the functions needed for UV-resistance, abbreviated *uvr*. By collecting large numbers of such mutants and testing them for their ability to restore resistance to UV radiation in combination, complementation groups were identified. Four of the complementation groups, or genes, encode proteins that play major rules in NER; they are *uvrA*, *uvrB*, *uvrC*and *uvrD*.

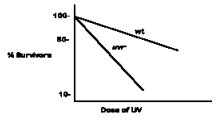


Figure 7.13. Survival curve of bacteria exposed to UV radiation. Cultures of bacteria are exposed to increasing doses of UV radiation, plotted along the horizontal axis. Samples of each irradiated culture are then plated and the number of surviving colonies are counted (plotted as a logarithmic function on the vertical axis). Mutant strains that are more sensitive to UV damage are defective in the genes that confer UV-resistance, i.e. they are defective in *uvr* functions.

The enzymes encoded by the *uvr*genes have been studied in detail. The polypeptide products of the *uvrA*, *uvrB*, and *uvrC*genes are subunits of a multisubunit enzyme called the **UvrABC excinuclease**. UvrA is the protein encoded by *uvrA*, UvrB is encoded by *uvrB*, and so on. The UvrABC complex recognizes damage-induced structural distortions in the DNA, such as pyrimidine dimers. It then cleaves on both sides of the damage. Then UvrD (also called helicase II), the product of the *uvrD*gene, unwinds the DNA, releasing the damaged segment. Thus for this system, the UvrABC and UvrD proteins carry out a series of steps in the cutting phase of excision repair. This leaves a gapped substrate for copying by DNA polymerase and pasting by DNA ligase.

The UvrABC proteins form a dynamic complex that recognizes damage and makes endonucleolytic cuts on both sides. The two cuts around the damage allow the single-stranded segment containing the damage to be excised by the helicase activity of UvrD. Thus the UvrABC dynamic complex and the UvrBC complex can be called **excinucleases**. After the damaged segment has been excised, a gap of 12 to 13 nucleotides remains in the DNA. This can be filled in by DNA polymerase and the remaining nick sealed by DNA ligase. Since the undamaged template directs the synthesis by DNA polymerase, the resulting duplex DNA is no longer damaged.

In more detail, the process goes as follows (Figure 7.14). $UvrA_2$ (a dimer) and Uvr B recognize the damaged site as a (UvrA)2UvrB complex. $UvrA_2$ then dissociates, in a step that requires ATP hydrolysis. This is an autocatalytic reaction, since it is catalyzed by UvrA, which is itself an ATPase. After UvrA has dissociated, UvrB (at the damaged site) forms a complex with UvrC. The UvrBC complex is the active nuclease. It makes the incisions on each side of the damage, in another step that requires ATP. The phosphodiester backbone is cleaved 8 nucleotides to the 5' side of the damage and 4-5 nucleotides on the 3' side. Finally, the UvrD helicase then unwinds DNA so the damaged segment is removed. The damaged DNA segment dissociates attached to the UvrBC complex. Like all helicase reactions, the unwinding requires ATP hydrolysis to disrupt the base pairs. Thus ATP hydrolysis is required at three steps of this series of reactions.

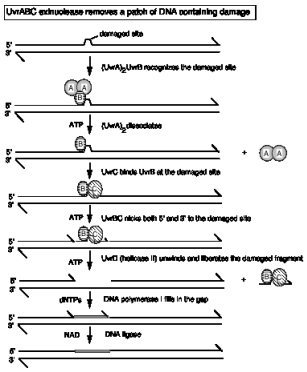


Figure 7.14.The Uvr(A)BC excinuclease of E. coli recognizes AP sites, thymine dimers, and other structural distortions and makes nicks on both sides of the damaged region. The 12-13 nucleotide-long fragment is released together with the excinuclease by helicase II action.

Exercise 7.9

How does an excinuclease differ from an exonuclease and an endonuclease?

Nucleotide excision repair is very active in mammalian cells, as well as cells from may other organisms. The DNA of a normal skin cell exposed to sunlight would accumulate thousands of dimers per day if this repair process did not remove them! One human genetic disease, called xeroderma pigmentosum (XP), is a skin disease caused by defect in enzymes that remove UV lesions. Fibroblasts isolated from individual XP patients are markedly sensitive to UV radiation when grown in culture, similar to the phenotype shown by *E. coliuvr*mutants. These XP cell lines can be fused in culture and tested for the ability to restore resistance to UV damage. XP cells lines that do so fall into different complementation groups. Several complementation groups, or genes, have been defined in this way. Considerable progress has been made recently in identifying the proteins encoded by each XP gene (Table 7.2). Note the tight analogy to bacterial functions needed for NER. Similar functions are also found in yeast (Table 7.2). Additional proteins utilized in eukaryotic NER include hHR23B (which forms a complex with the DNA-damage sensor XPC), ERCCI (which forms a complex with the XPF to catalyze incision 5' to the site of damage), the several other subunits of TFIIH (see Chapter 10) and the single-strand binding protein RPA.

Table 7.2: *Genes affected in XP patients, and encoded proteins*

Human Gene	Protein Function	Homologous to *S. cerevisiae*
XPA	Binds damaged DNA	Rad14
XPB	3' to 5' helicase, component of TFIIH	Rad25
XPC	DNA-damage sensor (in complex with hHR23B)	Rad4
XPD	5' to 3' helicase, component of TFIIH	Rad3
XPE	Binds damaged DNA	
XPF	Works with ERRC1 to cut DNA on 5' side of damage	Rad1
XPG	Cuts DNA on 3' side of damage	Rad2

NER occurs in two modes in many organisms, including bacteria, yeast and mammals. One is the global repair that acts throughout the genome, and the second is a specialized activity is that is coupled to transcription. Most of the XP gene products listed in Table 2 function in both modes of NER in mammalian cells. However, XPC (acting in a complex with another protein called hHR23B) is a DNA-damage sensor that is specific for global genome NER. In transcription coupled NER, the elongating RNA polymerase stalls at a lesion on the template strand; perhaps this is the damage recognition activity for this mode of NER. One of the basal transcription factors that associates with RNA polymerase II, TFIIH (see Chapter 10), also plays a role in both types of NER. A rare genetic disorder in humans, Cockayne syndrome (CS), is associated with a defect specific to transcription coupled repair. Two complementation groups have been identified, *CSA*and *CSB*. Determination of the nature and activity of the proteins encoded by them will provide additional insight into the efficient repair of transcribed DNA strands. The phenotype of CS patients is pleiotropic, showing both photosensitivity and severe neurological and other developmental disorders, including premature aging. These symptoms are more severe than those seen for XP patients with no detectable NER, indicating that transcription-coupled repair or the CS proteins have functions in addition to those for NER.

Other genetic diseases also result from a deficiency in a DNA repair function, such as Bloom's syndrome and Fanconi's anemia. These are intensive areas of current research. A good resource for updated information on these and other inherited diseases, as well as human genes in general, is the Online Mendelian Inheritance in Man, or OMIM, accessible at http://www.ncbi.nlm.nih.gov.

Ataxia telangiectasia, or AT, illustrates the effect of alterations in a protein not directly involved in repair, but perhaps signaling that is necessary for proper repair of DNA. AT is a recessive, rare genetic disease marked by uneven gait (ataxia), dilation of blood vessels (telangiectasia) in the eyes and face, cerebellar degeneration, progressive mental retardation, immune deficiencies, premature aging and about a 100-fold increase in susceptibility to cancers. That latter phenotype is driving much of the interest in this locus, since heterozygotes, which comprise about 1% of the population, also have an increased risk of cancer, and may account for as much as 9% of breast cancers in the United States. The gene that is mutated in AT (hence called "ATM") was isolated in 1995 and localized to chromosome 11q22-23.

The ATM gene does not appear to encode a protein that participates directly in DNA repair (unlike the genes that cause XP upon mutation). Rather, AT is caused by a defect in a cellular signaling pathway. Based on homologies to other proteins, the ATM gene product may be involved in the regulation of telomere length and cell cycle progression. The C-terminal domain is homologous to phosphatidylinositol-3-kinase (which is also a Ser/Thr protein kinase) - hence the connection to signaling pathways. The ATM protein also has regions of homology to DNA-dependent protein kinases, which require breaks, nicks or gaps to bind DNA (via subunit Ku); binding to DNA is required for the protein kinase activity. This suggests that ATM protein could be involved in targeting the repair machinery to such damage.

Base Excision Repair

Base excision repair differs from nucleotide excision repair in the types substrates recognized and in the initial cleavage event. Unlike NER, the base excision machinery recognizes damaged bases that do not cause a significant distortion to the DNA helix, such as the products of oxidizing agents. For example, base excision can remove uridines from DNA, even though a G:U base pair does not distort the DNA. Base excision repair is versatile, and this process also can remove some damaged bases that do distort the DNA, such as methylated purines. In general, the initial recognition is a specific damaged base, not a helical distortion in the DNA. A second major difference is that the initial cleavage is directed at the glycosidic bond connecting the purine or pyrimidine base to a deoxyribose in DNA. This contrasts with the initial cleavage of a phosphodiester bond in NER.

Cells contain a large number of specific **glycosylases** that recognize damaged or inappropriate bases, such as uracil, from the DNA. The glycosylase removes the damaged or inappropriate base by catalyzing cleavage of the N-glycosidic bond that attaches the base to the sugar-phosphate backbone. For instance, uracil-N-glycosylase, the product of the *ung* gene, recognizes uracil in DNA and cuts the N-glycosidic bond between the base and deoxyribose (Figure 7.15). Other glycosylases recognize and cleave damaged bases. For instance methylpurine glycosylase removes methylated G and A from DNA. The result of the activity of these glycosylases is an apurinic/apyrimidinic site, or AP site (Figure 7.15). At an AP site, the DNA is still an intact duplex, i.e. there are no breaks in the phosphodiester backbone, but one base is gone.

Next, an **AP endonuclease** nicks the DNA just 5' to the AP site, thereby providing a primer for DNA polymerase. In *E. coli*, the 5' to 3' exonuclease function of DNA polymerase I removes the damaged region, and fills in with correct DNA (using the 5' to 3' polymerase, directed by the sequence of the undamaged complementary strand).

Additional mechanisms have evolved for keeping U's out of DNA. *E. coli*also has a dUTPase, encoded by the *dut*gene, which catalyzes the hydrolysis of dUTP to dUMP. The product dUMP is the substrate for thymidylate synthetase, which catalyzes conversion of dUMP to dTMP. This keeps the concentration of dUTP in the cell low, reducing the chance that it will be used in DNA synthesis. Thus the combined action of the products of the *dut+ ung*genes helps prevent the accumulation of U's in DNA.

Exercise

In base excision repair, which enzymes are specific for a particular kind of damage and which are used for all repair by this pathway?

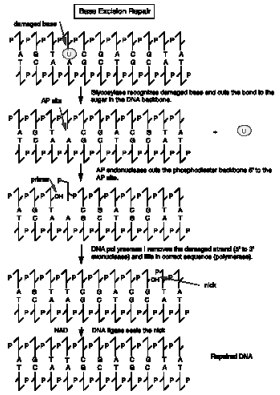


Figure 7.15. Base excision repair is initiated by a glycosylase that recognizes and removes chemically damaged or inappropriate bases in DNA. The glycosylase cleaves the glycosidic bond between the base and the sugar, leaving an apurinic/apyrimidinic site. The AP endonuclease can then nick the phosphodiester backbone 5' to the AP site. When DNA polymerase I binds the free primer end at the nick, its 5'-3' exonuclease activity cuts a few nucleotides ahead of the missing base, and its polymerization activity fills the entire gap of several nucleotides.

Mismatch Repair

The third type of excision repair we will consider is **mismatch repair**, which is used to repair errors that occur during DNA synthesis. Proofreading during replication is good but not perfect. Even with a functional e subunit, DNA polymerase III allows the wrong nucleotide to be incorporated about once in every 108 bp synthesized in *E. coli*. However, the measured mutation rate in bacteria is as low as one mistake per 1010 or 1011 bp. The enzymes that catalyze **mismatch repair**are responsible for this final degree of accuracy. They recognize misincorporated nucleotides, excise them and replace them with the correct nucleotides. In contrast to nucleotide excision repair, mismatch repair does not operate on bulky adducts or major distortions to the DNA helix. Most of the mismatches are substitutes within a chemical class, e.g. a C incorporated instead of a T. This causes only a subtle helical distortions in the DNA, and the misincorporated nucleotide is a normal component of DNA. The ability of a cell to recognize a mismatch reflects the exquisite specificity of **MutS**, which can distinguish normal base pairs from those resulting from misincorporation. Of course, the repair machinery needs to know which of the nucleotides at a mismatch pair is the correct one and which was misincorporated. It does this by determining which strand was more recently synthesized, and repairing the mismatch on the nascent strand.

In *E. coli*, the methylation of A in a GATC motif provides a covalent marker for the parental strand, thus methylation of DNA is used to discriminate parental from progeny strands. Recall that the ***dam* methylase** catalyzes the transfer of a methyl group to the A of the pseudopalindromic sequence GATC in duplex DNA. Methylation is delayed for several minutes after replication. IN this interval before methylation of the new DNA strand, the mismatch repair system can find mismatches and direct its repair activity to nucleotides on the unmethylated, newly replicated strand. Thus replication errors are removed preferentially.

The enzyme complex MutH-MutL-MutS , or MutHLS, catalyzes mismatch repair in *E. coli*. The genes that encode these enzymes, *mutH*, *mutL*and *mutS*, were discovered because strains carrying mutations in them have a high frequency of new mutations. This is called a **mutator phenotype**, and hence the name *mut*was given to these genes. Not all mutator genes are involved in mismatch repair; e.g., mutations in the gene encoding the proofreading enzyme of DNA polymerase III also have a mutator phenotype. This gene was independently discovered in screens for defects in DNA replication (*dnaQ*) and mutator genes (*mutD*). Three complementation groups within the set of mutator alleles have been implicated primarily in mismatch repair; these are *mutH*, *mutL*and *mutS*.

MutS will recognize seven of the eight possible mismatched base pairs (except for C:C) and bind at that site in the duplex DNA (Figure 7.16). **MutH**and **MutL** (with ATP bound) then join the complex, which then moves along the DNA in either direction until it finds a hemimethylated GATC motif, which can be as far a few thousand base pairs away. Until this point, the nuclease function of MutH has been dormant, but it is activated in the presence of ATP at a hemimethylated GATC. It cleaves the unmethylated DNA strand, leaving a nick 5' to the G on the strand containing the unmethylated GATC (i.e. the new DNA strand). The same strand is nicked on the other side of the mismatch. Enzymes involved in other processes of repair and replication catalyze the remaining steps. The segment of single-stranded DNA containing the incorrect nucleotide is to be excised by UvrD, also known as helicase II and MutU. SSB and exonuclease I are also involved in the excision. As the excision process forms the gap, it is filled in by the concerted action of DNA polymerase III (Figure 7.16.).

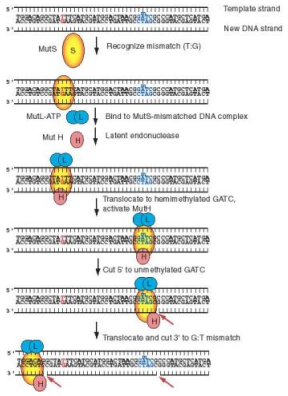


Figure 7.16 (part 1). Mismatch Repair by MutHLS: recognition of mismatch (shown in red), identifying the new DNA strand (using the hemimethylated GATC shown in blue) and cutting to encompass the unmethylated GATC and the misincorporated nucleotide (red G).

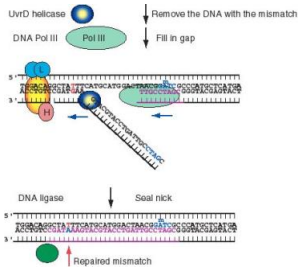


Figure 7.16 (part 2). Mismatch Repair: excision of the DNA with the misincorporated nucleotide bu Uvr D (aided by exonuclease I and SSB), gap filling by DNA polymerase III and ligation.

Mismatch repair is highly conserved, and investigation of this process in mice and humans is providing new clues about mutations that cause cancer.Homologs to the *E. coli* genes *mutL*and *mutS*have been identified in many other species, including mammals. The key breakthrough came from analysis of mutations that cause one of the most common hereditary cancers, *hereditary nonpolyposis colon cancer*(HNPCC). Some of the genes that, when mutated, cause this disease encode proteins whose amino acid sequences are significantly similar to those of two of the *E. coli*mismatch repair enzymes. The human genes are called *hMLH1*(for human *mutL*homolog 1), *hMSH1*, and *hMSH2*(for human *mutS* homolog 1 and 2, respectively). Subsequent work has shown that these enzymes in humans are involved in mismatch repair. Presumably the increased frequency of mutation in cells deficient in mismatch repair leads to the accumulation of mutations in proto-oncogenes, resulting in dysregulation of the cell cycle and loss of normal control over the rate of cell division.

Exercise

The human homologs to bacterial enzymes involved in mismatch repair are also implicated in homologous functions. Given the human homologs discussed above, which enzymatic functions found in bacterial mismatch repair are also found in humans? What functions are missing, and hence are likely carried out by an enzyme not homologous to those used in bacterial mismatch repair?

Recombination Repair (Retrieval system)

In the three types of excision repair, the damaged or misincorporated nucleotides are cut out of DNA, and the remaining strand of DNA is used for synthesis of the correct DNA sequence. However, this complementary strand is not always available. Sometimes DNA polymerase has to synthesize past a lesion, such as a pyrimidine dimer or an AP site. One way it can do this is to stop on one side of the lesion and then resume synthesis about 1000 nucleotides further down. This leaves a gap in the strand opposite the lesion (Figure 7.17).

The information needed at the gap is retrieved from the normal daughter molecule by bringing in a single strand of DNA, using RecA-mediated recombination (see Chapter VIII). This fills the gap opposite the dimer, and the dimer can now be replaced by excision repair (Figure 7.17). The resulting gap in the (previously) normal daughter can be filled in by DNA polymerase, using the good template.

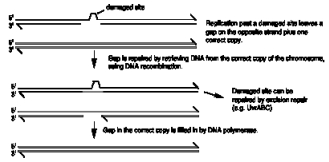


Figure 7.17. Recombination repair, a system for retrieval of information

Translesion Synthesis

As just described, DNA polymerase can skip past a lesion on the template strand, leaving behind a gap. It has another option when such a lesion is encountered, which is to synthesis DNA in a non-template directed manner. This is called **translesion synthesis**, bypass synthesis, or error-prone repair. This is the last resort for DNA repair, e.g. when repair has not occurred prior to replication. In translesion replication, the DNA polymerase shifts from template directed synthesis to catalyzing the incorporation of random nucleotides. These random nucleotides are usually mutations (i.e. in three out of four times), hence this process is also designated error-prone repair.

Translesion synthesis uses the products of the *umuC*and *umuD*genes. These genes are named for the UV nonmutable phenotype of mutants defective in these genes

Exercise

Question 7.11. Why do mutations in genes required for translesion synthesis (error prone repair) lead to a *non*mutable phenotype?

UmuD forms a homodimer that also complexes with UmuC. When the concentration of single-stranded DNA and RecA are increased (by DNA damage, see next section), RecA stimulates an autoprotease activity in $UmuD_2$ to form $UmuD'_2$. This cleaved form is now active in translesional synthesis. UmuC itself is a DNA polymerase. A multisubunit complex containing UmuC, the activated $UmuD'_2$ and the a subunit of DNA polymerase III catalyze translesional synthesis. Homologs of the UmuC polymerase are found in yeast (RAD30) and humans (XP-V).

The SOSresponse

A coordinated battery of responses to DNA damage in *E. coli*is referred to as the SOS response. This name is derived from the maritime distress call, “SOS” for "Save Our Ship". Accumulating damage to DNA, e.g. from high doses of radiation that break the DNA backbone, will generate single-stranded regions in DNA. The increasing amounts of single-stranded DNA induce SOS functions, which stimulate both the recombination repair and the translesional synthesis just discussed.

Key proteins in the SOS response are **RecA** and **LexA**. RecA binds to single stranded regions in DNA, which activates new functions in the protein. One of these is a capacity to further activate a latent proteolytic activity found in several proteins, including the LexA repressor, the **UmuD**protein and the repressor encoded by bacteriophage lambda (Figure 7.18). RecA activated by binding to single-stranded DNA is not itself a protease, but rather it serves as a co-protease, activating the latent proteolytic function in LexA, UmuD and some other proteins.

In the absence of appreciable DNA damage, the LexA protein represses many operons, including several genes needed for DNA repair: *recA, lexA, uvrA, uvrB, and umuC*.When the activated RecA stimulates its proteolytic activity, it cleaves itself (and other proteins), leading to coordinate induction of the SOS regulated operons (Figure 7.18).

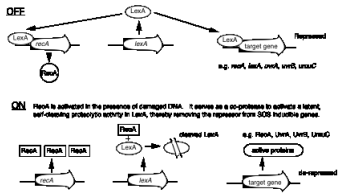


Figure 7.18. RecA and LexA control the SOS response.

Restriction/Modification systems

The DNA repair systems discussed above operate by surveillance of the genome for damage or misincorporation and then bring in enzymatic machines to repair the defects. Other systems of surveillance in bacterial genomes are **restriction/modification systems**. These look for foreign DNA that has invaded the cell, and then destroy it. In effect, this is another means of protecting the genome from the damage that could result from the integration of foreign DNA.

These systems for safeguarding the bacterial cell from invasion by foreign DNA use a combination of covalent modification and restriction by an endonuclease. Each species of bacteria modifies its DNA by **methylation** at specific sites (Figure 7.19). This protects the DNA from cleavage by the corresponding **restriction endonuclease**. However, any foreign DNA (e.g. from an infecting bacteriophage or from a different species of bacteria) will not be methylated at that site, and the restriction endonuclease will cleave there. The result is that invading DNA will be cut up and inactivated, while not damaging the host DNA.

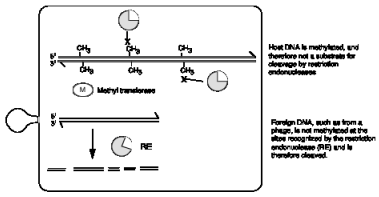


Figure 7.19. Restriction/modification systems in bacteria.

Any DNA that escapes the restriction endonuclease will be a substrate for the methylase. Once methylated, the bacterium now treats it like its own DNA, i.e. does not cleave it. This process can be controlled genetically and biochemically to aid in recombinant DNA work. Generally, the restriction endonuclease is encoded at the *r*locus and the methyl transferase is encoded at the *m* locus. Thus passing a plasmid DNA through an *r*-*m*+strain (defective in restriction but competent for modification) will make it resistant to restriction by strains with a wildtype *r*+gene. For some restriction/modification systems, both the endonuclease and the methyl transferase are available commercially. In these cases, one can modify the foreign DNA (e.g. from humans) prior to ligating into cloning vectors to protect it from cleavage by the restriction endonucleases it may encounter after transformation into bacteria.

For the type II restriction/modification systems, the methylation and restriction occurs at the same, pseudopalindromic site. These are the most common systems, with a different sequence specificity for each bacterial species. This has provided the large variety of restriction endonucleases that are so commonly used in molecular biology.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

7.E : Mutation and Repair of DNA (Exercises)

Q7.12

If the top strand of the segment of DNA GGTCGTT were targeted for reaction with nitrous acid, and then it underwent two rounds of replication, what are the likely products?

Q7.13

Are the following statements about nucleotide excision repair in *E. coli* true or false?

a. UvrA and UvrB recognize structural distortions resulting from damage in the DNA helix.
b. In a complex with UvrB, UvrC cleaves the damaged strand on each side of the lesion.
c. The helicase UvrD unwinds the DNA, thereby dissociating the damaged patch.

Q7.14

Are the following statements about mismatch repair in *E. coli* true or false?

a. MutS will recognize a mismatch.
b. MutL, in a complex with ATP, will bind to the MutS (bound to the mismatched region) and activate MutH.
c. MutH will cleave 5' to the G of the nearest methylated GATC motif (GmeATC).
d. The mismatch repair system can discriminate between old versus newly synthesized strands of DNA.

For the **next 6** problems, consider the following DNA sequence, from the first exon of the *HRAS* gene. A transversion of G to T at position 24 confers anchorage independence and tumorigenicity to NIH 3T3 cells (fibroblasts). This mutation is one step in tumorigenic transformation of bladder cells, and it likely plays a role in other cancers.

10 20 30

5' TAAGCTGGTG GTGGTGGGCG CCGGCGGTGT

3' ATTCGACCAC CACCACCCGC GGCCGCCACA

Q7.15

hat would the sequence be if the G at position 14 (top strand) were alkylated at the O6 position by MNNG and then went through 2 rounds of replication?

Q7.16

hat would the sequence be if the C at position 24 (bottom strand) were oxidized by HNO_2 and then went through 2 rounds of replication?

Q7.17

What would happen if this sequence were irradiated with UV at a wavelength of 260 nm?

Q7.18

If you were in charge of maintaining this DNA sequence, and you had the enzymatic tools known in *E. coli*, how would you repair the damage from question 7.15? Consider what would happen if the damage were corrected before or after replication.

Q7.19

How could

a. the oxidative damage in problem 7.16 or
b. the UV products in problem 7.17 be repaired?

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

7.S: Mutation and Repair of DNA (Summary)

Summary: Causes of Transitions and Transversions

Table 7.1 lists several causes of mutations in DNA, including mutagens as well as mutator strains in bacteria. Note that some of these mutations lead to mispairing (substitutions), others lead to distortions of the helix, and some lead to both. Transitions can be generated both by damage to the DNA and by misincorporation during replication. Transversions occur primarily by misincorporation during replication. The frequency of such errors is greatly increased in mutator strains, e.g. lacking a proofreading function in the replicative DNA polymerase. Also, after a bacterial cell has sustained sufficient damage to induce the SOS response, the DNA polymerase shifts into a an error-prone mode of replication. This can also be a source of mutant alleles.

Table. 7.1. Summary of effects of various agents that alter DNA sequences (mutagens and mutator genes)

Agent (mutagen, etc.)	Example	Result
Nucleotide analogs	BrdUTP	transitions, e.g. A:T to G:C
Oxidizing agents	nitrous acid	transitions, e.g. C:G to T:A
Alkylating agents	nitrosoguanidine	transitions, e.g. G:C to A:T
Frameshift mutagens	Benz(a)pyrene	deletions (short)
Ionizing radiation	X-rays, g-rays	breaks and deletions (large)
UV	UV, 260 nm	Y-dimers, block replication
Misincorporation:		
Altered DNA Pol III	*mutD=dnaQ*; e subunit of DNA PolIII	transitions, transversions and frameshifts in mutant strains
Error-prone repair	Need UmuC, UmuD, DNA PolIII	transitions and transversions in wild-type during SOS
Other mutator genes	*mutM*, *mutT*, *mutY*	transversions in the mutant strains

Additional Readings

- Friedberg, E. C., Walker, G. C., and Siede, W. (1995) **DNA repair and mutagenesis**, ASM Press, Washington, D.C.
- Kornberg, A. and Baker, T. (1992) **DNA Replication**, 2nd Edition, W. H. Freeman and Company, New York.
- Zakian, V. (1995) *ATM*-related genes: What do they tell us about functions of the human gene? **Cell** 82: 685-687.
- Kolodner, R. (1996) Biochemistry and genetics of eukaryotic mismatch repair. **Genes & Development**10:1433-1442.
- Sutton MD, Smith BT, Godoy VG, Walker GC. (2000) The SOS response: recent insights into umuDC-dependent mutagenesis and DNA damage tolerance.**Annu Rev Genet**34:479-497.
- De Laat, W. L., Jaspers, N. C. J. and Hoeijmakers, J. H. J. (1999) Molecular mechanism of nucleotide excision repair. **Genes & Development**13: 768-785. This review focuses on nucleotide excision repair in mammals.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

CHAPTER OVERVIEW

8: Recombination of DNA

The chapter on mutation and repair of DNA dealt mainly with small changes in DNA sequence, usually single base pairs, resulting from errors in replication or damage to DNA. The DNA sequence of a chromosome can change in large segments as well, by the processes of recombination and transposition. **Recombination** is the production of new DNA molecule(s) from two parental DNA molecules or different segments of the same DNA molecule; this will be the topic of this chapter. **Transposition** is a highly specialized form of recombination in which a segment of DNA moves from one location to another, either on the same chromosome or a different chromosome; this will be discussed in the next chapter.

Suggested readings

- Holliday, R. (1964) A mechanism for gene conversion in fungi. Genetics Research 5: 282-304.
- Orr-Weaver, T. L., Szostak, J. W. and Rothstein, R. J. (1981) Yeast transformation: a model system for the study of recombination. Proc. Natl. Acad. Sci. USA 78: 6354-6358.
- Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J. and Stahl, F. W. (1983) The double-strand-break repair model for recombination. Cell 33: 25-35.
- Stahl, F. W. (1994) The Holliday junction on its thirtieth anniversary. Genetics 138: 241-246.
- Kowalczykowski, S.C., Dixon, D. A., Eggleston, A. K., Lauder, S. D. and Rehrauer, W. M. (1994) Microbiological Reviews 58:401-465.
- Eggleston, A. K. and West, S. C. (1996) Exchanging partners: recombination in E. coli. Treand in Genetics 12: 20-25.
- Edelmann, W. and Kucherlapati, R. (1996) Role of recombination enzymes in mammalian cell survival. Proc. Natl. Acad. Sci. USA 93: 6225-6227.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

8.1: Types and Examples of Recombination

At least four types of naturally occurring recombination have been identified in living organisms (Figure 8.1).

1. **General or homologous recombination** occurs between DNA molecules of very similar sequence, such as homologous chromosomes in diploid organisms. General recombination can occur throughout the genome of diploid organisms, using one or a small number of common enzymatic pathways. This chapter will be concerned almost entirely with general recombination.
2. **Illegitimate or nonhomologous** recombination occurs in regions where no large-scale sequence similarity is apparent, e.g. translocations between different chromosomes or deletions that remove several genes along a chromosome. However, when the DNA sequence at the breakpoints for these events is analyzed, short regions of sequence similarity are found in some cases. For instance, recombination between two similar genes that are several million bp apart can lead to deletion of the intervening genes in somatic cells.
3. **Site-specific recombination** occurs between particular short sequences (about 12 to 24 bp) present on otherwise dissimilar parental molecules. Site-specific recombination requires a special enzymatic machinery, basically one enzyme or enzyme system for each particular site. Good examples are the systems for integration of some bacteriophage, such as l, into a bacterial chromosome and the rearrangement of immunoglobulin genes in vertebrate animals.
4. The third type is **replicative recombination**, which generates a new copy of a segment of DNA. Many transposable elements use a process of replicative recombination to generate a new copy of the transposable element at a new location.

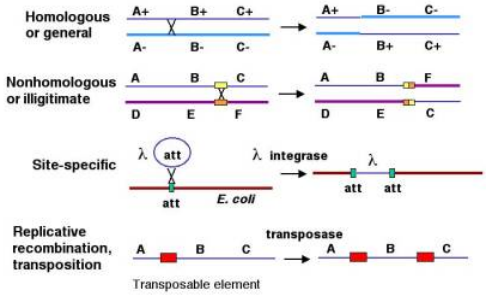


Figure 8.1. Types of natural recombination.Each line represents a chromosome or segment of a chromosome; thus a single line represents both strands of duplex DNA. For **homologous or general recombination**, each homologous chromosome is shown as a different shade of blue and a distinctive thickness, with different alleles for each of the three genes on each. Recombination between genes A and B leads to a reciprocal exchange of genetic information, changing the arrangement of alleles on the chromosomes. For **nonhomologous (or illegitimate) recombination**, two different chromosomes (denoted by the different colors and different genes) recombine, moving, e.g. gene C so that it is now on the same chromosome as genes D and E. Although the sequences of the two chromosomes differ for most of their lengths, the segments at the sites of recombination may be related, denoted by the yellow and orange rectangles. **Site-specific recombination** leads to the combination of two different DNA molecules, illustrated here for a bacteriophage l integrating into the E. colichromosome. This reaction is catalyzed by a specific enzyme that recognizes a short sequence present in both the phage DNA and the target site in the bacterial chromosome, called att. **Replicative recombination** is seen for some transposable elements, shown as red rectangles, again using a specific enzyme, in this case encoded by the transposable element.

Recombinant DNA technology uses two other types of recombination. The **directed cutting and rejoining of different DNA molecules *in vitro*** using restriction endonucleases and DNA ligases is well-known, as covered in Chapter 2. Once made, these recombinant DNA molecules are then introduced into a host organism, often a bacterium. If the recombinant DNA is a plasmid, phage or other molecule capable of replicating in the host, it will stay extrachromosomal. However, one can introduce the recombinant DNA molecule into a host in which it cannot replicate, such as a plant, an animal cell in culture, or a fertilized mouse egg. In order for the host to be stably transformed, the introduced DNA has to be taken up into a host chromosome. In bacteria and yeast, this can occur by homologous recombination at a reasonably high frequency. However, this does not occur in plant or animal cells. In contrast, at a low frequency, some of these introduced DNA molecules are incorporated into random locations in the chromosomes of the host cell. Thus **random recombination** into chromosomes can make stably transfected cells and transgenic plants and animals. The mechanism of this recombination during transformation or transfection is not well understood, although it is commonly used in the laboratory.

General recombination is an integral part of the complex process of **meiosis** in sexually reproducing organisms. It results in a **crossing over** between pairs of genes along a chromosome, which are revealed in appropriate matings (Chapter 1). The **chiasmata** that link homologous chromosomes during meiosis are the likely sites of the crossovers that result in recombination. General recombination also occurs in nonsexual organisms when two copies of a chromosome or chromosomal segment are present. We

have encountered this as recombination during F-factor mediated conjugal transfer of parts of chromosomes in *E. coli* (Chapter 1). Recombination between two phage during a mixed infection of bacteria is another example. Also, the retrieval system for post-replicative repair (Chapter 7) involves general recombination.

The mechanism of recombination has been intensively studied in bacteria and fungi, and some of the enzymes involved have been well characterized. However, a full picture of the mechanism, or mechanisms, of recombination has yet to be achieved. We will discuss the general properties of recombination, cover two models of recombination, and discuss some of the properties of key enzymes in the pathways of recombination.

Reciprocal and Nonreciprocal Recombination

General recombination can appear to result in either an equal or an unequal exchange of genetic information. Equal exchange is referred to as **reciprocal recombination**, as illustrated in Figure 8.1. In this example, two homologous chromosomes are distinguished by having wild type alleles on one chromosome (A+, B+ and C+) and mutant alleles on the other (A-, B- and C-). Homologous recombination between genes A and B exchanges the segment of one chromosome containing the wild type alleles of genes B and C (B+ and C+) for the segment containing the mutant alleles (B- and C-) on the homologous chromosome. This could be explained by breaking and rejoining of the two homologous chromosomes during meiosis; however, we will see later that the enzymatic mechanism is more complex than simple cutting and ligation. The DNA that is removed from the top (thin dark blue) chromosome is joined with the bottom (thick light blue) chromosome, and the DNA removed from the bottom chromosome is added to the top chromosome. This process resulting in new DNA molecules that carry genetic information derived from both parental DNA molecules is called **reciprocal recombination**. The number of alleles for each gene remains the same in the products of this recombination, only their arrangement has changed.

General recombination can also result in a one-way transfer of genetic information, resulting in an allele of a gene on one chromosome being changed to the allele on the homologous chromosome. This is called **gene conversion**. As illustrated in Figure 8.2, recombination between two homologous chromosomes A+B+C+ and A-B-C- can result in a new arrangement, A-B+C-, without a change in the parental A+B+C+. In this case, the allele of gene B on the bottom chromosome has changed from B- to B+ without a reciprocal change on the other chromosome. Thus, in contrast to reciprocal recombination, the number of types of alleles for gene B has changed in the products of this recombination; now there is only one (B+). This is an example of interchromosomal gene conversion, i.e. between homologous chromosomes. Similar copies of genes can be on the same chromosome, and these can undergo gene conversion as well. Cases of intrachromosomal gene conversion have been documented for the gamma-globin genes of humans. The occurrence of gene conversion during general recombination is one indication that the enzymatic mechanism is not a simple cutting and pasting.

Figure 8.2.Gene conversion changing allele B- on the bottom (thick, light blue) chromosome to B+. Note that the arrangement of alleles on the top (thin, dark blue) chromosome has not changed.

Exercise 8.1

Why would you not interpret the A-B+C- chromosome as resulting from two reciprocal crossovers, one on each side of gene B?

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

8.2: Detecting Recombination

Mendel's Second Law described the random assortment of alleles of pairs of genes. However, certain pairs of genes show deviations from this random assortment, leading to the conclusion that those genes are linked on a chromosome. The linkage is not always complete, meaning that nonparental genotypes are seen in a proportion of the progeny. This is explained by crossing over between the gene pairs during meiosis in the parents.

Let's think about the general recombination shown in Figure 8.1 in this context. The two chromosomes outlined in the figure are in a heterozygous parent, with the wild type alleles for genes A and B (A+ and B+) are on one chromosome and the mutant alleles (A- and B-) are on the homologous chromosome (we can ignore gene C for this discussion.) Homologous recombination during meiosis can generate the new chromosomes shown, now with A+ and B- on one chromosome and A- and B+ on the other. However, this crossover will not occur between genes A and B on all chromosomes undergoing meiosis in this parent. Although recombination is an essential part of meiosis (see next section), the sites of recombination on a particular chromosome varies from cell to cell. In fact, the probability that a crossover will occur between two genes is a measure of the genetic distance between them (reviewed in Chapter 1). The recombinant chromosomes resulting from a crossover are revealed in a mating between the heterozygous parent (A+B+/A-B-) and a homozygous recessive individual (A-B-/A-B-). Most of the germ cells contributed by the heterozygous parent will have one of the parental chromosomes A+B+ or A-B-, but those germ cells resulting from the crossover between genes A and B will have the recombinant chromosomes (either A+B- or A-B+). The homozygous recessive parent will contribute only A-B- chromosomes. Thus in the progeny, one sees mainly offspring whose phenotype is determined by one of the chromosomes in the heterozygous parent, either wild type A and B (genotype of A+B+/A-B-) or mutant A and B (genotype A-B-/A-B-). However, some of the progeny will show a wild type A and a mutant B phenotype, or vice versa. These carry the chromosomes resulting from the crossover (genotype of A+B-/A-B- or A-B+/A-B-). The frequency with which one sees progeny with nonparental phenotypes is related to their distance apart on the chromosome; this measure is referred to as a genetic distance or a recombination distance.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

8.3: Meiotic Recombination

A diploid organism has two copies of each chromosome. If it has four chromosomes, there are two pairs, A and A' and B and B', not four different chromosomes A, B, C and D. One copy of each chromosome came from its father (e.g. A and B) and one copy of each came from its mother (e.g. A' and B'). Meiosis is the process of reductive division whereby a diploid organism generates haploid germ cells (in this case, with two chromosomes), and each germ cell has a single copy of **each** chromosome. In this example, meiosis does not generate germ cells with A and A' or B and B', rather it produces cells with A and B, or A and B', or A' and B, or A' and B'. The homologous chromosomes, each consisting of two sister chromatids, are paired during the first phase of meiosis, e.g., A with A' and B with B' (Figure 8.3; see also Figs. 1.3 and 1.4). Then the homologous chromosomes are moved to separate cells at the end of the first phase, insuring that the two homologs do not stay together during reductive division in the second phase of meiosis. Thus each germ cell receives the haploid complement of the genetic material, i.e. one copy of each chromosome. The combination of two haploid sets of chromosomes during fertilization restores the diploid state, and the cycle can resume. Failure to distribute one copy of each chromosome to each germ cell has severe consequences. Absence of one copy of a chromosome in an otherwise diploid zygote is likely fatal. Having an extra copy of a chromosome (trisomy) also causes problems. In humans, trisomy for chromosomes 15 or 18 results in perinatal death and trisomy 21 leads to developmental defects known as Down's syndrome.

Exercise 8.2

If this diploid organism with chromosomes A, A', B and B' underwent meiosis **without** homologous pairing and separation of the homologs to different cells, what fraction of the resulting haploid cells would have an A-type chromosome (A or A') and a B-type chromosome (B or B')?

The ability of homologous chromosomes to be paired during the first phase of meiosis is fundamental to the success of this process, which maintains a correct haploid set of chromosomes in the germ cell. **Recombination is an integral part of the pairing of homologous chromosomes**. It occurs between non-sister chromatids during the pachytene stage of meiosis I (the first stage of meiosis) and possibly before, when the homologous chromosomes are aligned in zygotene (Figure 8.3). The crossovers of recombination are visible in the diplotene phase. During this phase, the homologous chromosomes partially separate, but they are still held together at joints called **chiasmata**; these are likely the actual crossovers between chromatids of homologous chromosomes. The chiasmata are progressively broken as meiosis I is completed, corresponding to resolution of the recombination intermediates. During anaphase and telophase of meiosis I, each homologous chromosome moves to a different cell, i.e. A and A' in different cells, B and B' in different cells in our example. Thus recombinations occur in every meiosis, resulting in at least one exchange between pairs of homologous chromosomes per meiosis.

Recent genetic evidence demonstrates that recombination is required for homologous pairing of chromosomes during meiosis. Genetic screens have revealed mutants of yeast and *Drosophila* that block pairing of homologous chromosomes. These are also defective in recombination. Likewise, mutants defective in some aspects of recombination are also defective in pairing. Indeed, the process of synapsis (or pairing) between homologous chromosomes in zygotene, crossing over between homologs in pachytene, and resolution of the crossovers in the latter phases of meiosis I (diakinesis, metaphase I, and anaphase I) correspond to the synapsis, formation of a recombinant joint and resolution that mark the progression of recombination, as will be explained below.

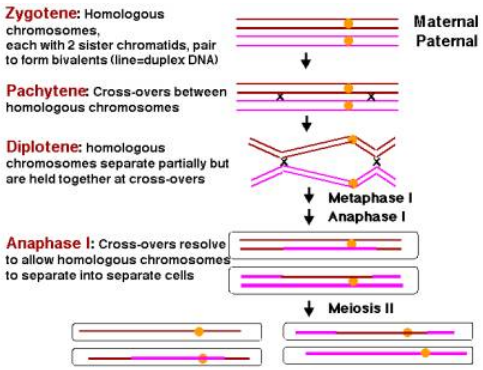


Figure 8.3.Homologous pairing and recombination during the first stage of meiosis (meiosis I). After DNA synthesis has been completed, two copies of each homologous chromosome are still connected at centromeres (yellow circles). This diagram starts with replicated chromosomes, referred to as the four-strand stage in the literature on meiosis and recombination. In this usage, each "strand" is a **chromatid** and is a duplex DNA molecule. In this diagram, each duplex DNA molecule is shown as a single line, brown for the two sister chromatids of chromosome derived from the mother (maternal) and pink for the sister chromatids from the paternal chromosome. Only one homologous pair is shown, but usually there are many more, e.g. 4 pairs of chromosomes in Drosophilaand 23 pairs in humans. During the meiosis I, the homologous chromosomes align and then separate. At the zygotene stage, the two homologous chromosomes, each with two sister chromatids, pair along their length in a process called synapsis. The resulting group of four chromatids is called a tetrad or bivalent. During pachytene, recombination occurs between a maternal and a paternal chromatid, forming crossovers between the homologous chromosomes. The two homologous chromosomes separate along much of their length at diplotene, but they continue to be held together at localized chiasmata, which appear as X-shaped structures in micrographs. These physical links are thought to be the positions of crossing over. During metaphase and anaphase of the first meiotic division, the crossovers are gradually broken (with those at the ends resolved last) and the two homologous chromosomes (each still with two chromatids joined at a centromere) are moved into separate cells. During the second meiotic division (meiosis II), the centromere of each chromosome separates, allowing the two chromatids to move to separate cells, thus finishing the reductive division and making four haploid germ cells.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

8.4: Advantages of Genetic Recombination

Not only is recombination needed for homologous pairing during meiosis, but recombination has at least two additional benefits for sexual species. It makes new combinations of alleles along chromosomes, and it restricts the effects of mutations largely to the region around a gene, not the whole chromosome.

Since each chromosome undergoes at least one recombination event during meiosis, new combinations of alleles are generated. The arrangement of alleles inherited from each parent are not preserved, but rather the new germ cells carry chromosomes with new combinations of alleles of the genes (Figure 8.4). This remixing of combinations of alleles is a rich source of diversity in a population.

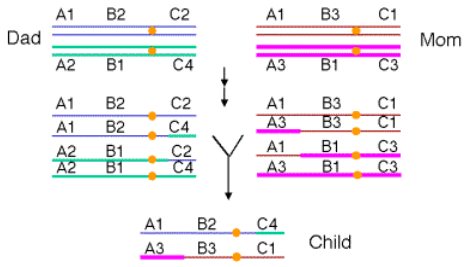


Figure 8.4. Recombination during meiosis generates new combinations of alleles in the offspring. One homologous pair of chromosomes is illustrated, starting at the “four-strand” stage. Each line is a duplex DNA molecule in a chromatid. The two chromosomes in the father (inherited from the paternal grandparents) are blue and green; the homologous chromosomes in the mother (inherited from the maternal grandparents) are brown and pink. All chromosomes have genes A, B and C; different numbers refer to different alleles. In this illustration, a crossover on the short arm of the chromosome during development of the male germ cells links allele 4 of gene C with alleles 1 of gene A and allele 2 of gene B, as well as the reciprocal arrangement. A crossover on the long arm of the chromosome is illustrated for development of the female germ cell, making the new combination A3, B3 and C1. A child can have the new chromosomes A1B2C4 and A3B3C1. Note that neither of these combinations was in the father or mother.

Over time, recombination will separate alleles at one locus from alleles at a linked locus. A chromosome through generations is not fixed, but rather it is "fluid," having many different combinations of alleles. This allows nonfunctional (less functional) alleles to be cleared from a population. If recombination did not occur, then one deleterious mutant allele would cause an entire chromosome to be eliminated from the population. However, with recombination, the mutant allele can be separated from the other genes on that chromosome. Then negative selection can remove defective alleles of a gene from a population while affecting the frequency of alleles only of genes in tight linkage to the mutant gene. Conversely, the rare beneficial alleles of genes can be tested in a population without being irreversibly linked to any potentially deleterious mutant alleles of nearby genes. This keeps the effective target size for mutation close to that of a gene, not the whole chromosome.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

8.5: Evidence for Heteroduplexes from Recombination in Fungi

The mechanism by which recombination occurs has been studied primarily in fungi, such as the budding yeast *Saccharomyces cerevisiae* and the filamentous fungus *Ascomycetes*, and in bacteria. The fungi undergo meiosis, and hence some aspects of their recombination systems may be more similar to that of plants and animals than is that of bacteria. However, the enzymatic functions discovered by genetic and biochemical studies of recombination in bacteria are also proving to have counterparts in eukaryotic organisms as well. We will refer to studies mainly in fungi for the models of recombination, and to studies mainly in bacteria for the enzymatic pathways.

Many important insights into the mechanism of recombination have come from studies in fungi. One fundamental observation is that recombination proceeds by the formation of a region of heteroduplex, i.e. the recombination products have a region with one strand from one chromosome and the complementary strand from the other chromosome. Thus recombination is not a simple cut and paste operation, unlike the joining of two different molecules by recombinant DNA technology. The two recombining molecules are joined and form a hybrid, or heteroduplex, over part of their lengths.

The anatomy and physiology of the filamentous fungus Ascomycetes allows one to observe this heteroduplex formed during recombination. A cell undergoing meiosis starts with a 4n complement of chromosomes (i.e. twice the diploid number) and undergoes two rounds of cell division to form four haploid cells. In fungi these haploid germ cells are spores, and they are found together in an ascus. They can be separated by dissection and plated individually to examine the phenotype of the four products of meiosis. This is called **tetrad analysis**.

The fungus *Ascomycetes* goes one step further. After meiosis is completed, the germ cells undergo one further round of replication and mitosis. This separates each individual polynucleotide chain (or “strand” in the sense used in nucleic acid biochemistry) of each DNA duplex in the meiotic products into a separate spore. The eight spores in the ascus reflect the genetic composition of each of the eight polynucleotide chains in the four homologous chromosomes. (The two sister chromatids in each homologous chromosome become two chromosomes after meiosis, and each chromosome is a duplex of two polynucleotide chains.)

The order of the eight spores in the ascus of Ascomycetes reflects the descent of the spores from the homologous chromosomes. As shown in Figure 8.5, a heterozygote with a “blue” allele on one homologous chromosome and a “red” allele on the other will normally produce four “blue” spores and four “red” spores. The four spores with the same phenotype were derived from one homologous chromosome and are adjacent to each other in the ascus. This is called a 4:4 parental ratio, i.e. with respect to the phenotypes of the parent of the heterozygote.

The evidence for heteroduplex formation comes from deviations from the normal 4:4 ratio. Sometimes a **3:5 parental ratio** is seen for a particular genetic marker. This shows that one polynucleotide chain of one allele has been lost (giving 4-1=3 spores with the corresponding phenotype in the ascus) and replaced by the polynucleotide chain of the other allele (giving 4+1=5 spores with the corresponding phenotype). As illustrated in Figure 8.5, this is 3 blue spores and 5 red spores. The segment of the chromosome containing this gene was a heteroduplex with one chain from each of two alleles. The round of replication and mitosis that follows meiosis in this fungus allows the two chains to be separated into two alleles that generated a different phenotype in a plating assay. Thus this 3:5 ratio results from post-meiotic segregation of the two chains of the different alleles. In this fungus, a region of heteroduplex can be directly observed by a plating assay.

The region of heteroduplex is associated with a recombination between the chromosomes. Other genes flank the region of heteroduplex shown in Figure 8.5. In many cases, the arrangement of alleles of these flanking genes has changed from that on the parental chromosomes, reflecting a recombination. For instance, let the region of heteroduplex be in a gene B, flanked by gene A in the left and gene C on the right. Each gene has a blue allele and a red allele, making the parental chromosomes AbBbCb and ArBrCr. If one monitored the phenotypes of determined by genes A and C (in addition to B) in the third and fourth spores (derived from the chromosome with the heteroduplex), they would see the phenotypes for the nonparental chromosomes AbBbCr and AbBrCr. This change in the flanking markers (genes A and C) reflects a recombination. Thus the heteroduplex can be found between markers that have undergone recombination.

Other markers can show a 2:6 parental ratio. This means that one of the alleles (formerly blue in fig. 8.5) has been changed to the other allele (now red), in a process called gene conversion. This can occur between flanking markers that have been switched because of recombination. Thus like the heteroduplex, the region of gene conversion is associated with recombination. Models for recombination need to incorporate both phenomenon into their proposed mechanism.

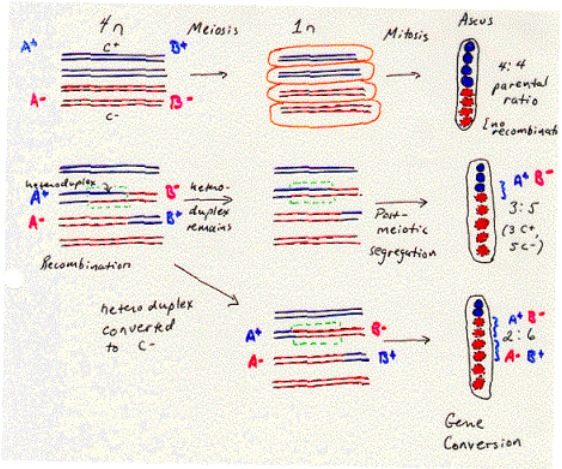


Figure 8.5. Spores formed during meiosis in Ascomycetes reflect the genetic composition of the parental DNA chains. The four homologous chromosomes in the 4n state are shown as duplex DNA molecules, with one line for each DNA chain. Two sister chromatids are blue and two sister chromatids are red, reflecting their ability to be distinguished in a plating assay for particular genes along the chromosome. Meiosis places each of the four homologous chromosomes into a different cell, and in this species, it is followed by replication and mitosis so that each of the eight spores (circles in the elongated ellipse representing the ascus) has the genetic composition of each of the eight DNA chains in the four chromosomes that result from meiosis (two complementary chains per chromosome). A region of heteroduplex can be seen as a 3:5 parental ratio after post-meiotic segregation. A region of gene conversion can be seen as a 2:6 parental ratio.

Exercise 8.3.

Imagine that you are studying a fungus that generates an ascus with 8 spores like *Ascomycetes*, in which the products of meiosis complete an additional round of replication and mitosis. You generate a heterozyous strain by mating a parent that was homozyous for the markers leu+, SmR, ade8+ and another that was leu-, SmS, ade8-. Previous studies had shown that all three markers are linked in the order given. Each of these pairs of alleles can be distinguished in a plating assay. The allele leu+ confers leucine auxotrophy whereas leu- confers leucine prototrophy. The allele SmR confers resistance to spectinomycin whereas SmS is sensitive to this antibiotic. Colonies of fungi with the ade8+ allele give a red color in under appropriate conditions in a plate, but those with the ade8- are white. Analysis of the individual spores from an ascus gave the following phenotypes results. The spores are numbered in the order they were in the ascus. What are the corresponding genotypes of the chromosome in each spore? How do you interpret these results with respect to recombination?

Spore	leucine	Spectinomycin	Color in ade test
1	prototroph	resistant	red
2	prototroph	resistant	red
3	prototroph	resistant	white
4	prototroph	sensitive	white
5	auxotroph	sensitive	red
6	auxotroph	sensitive	red
7	auxotroph	sensitive	white
8	auxotroph	sensitive	white

8.6: Holliday Model for General Recombination - Single Strand Invasion

In 1964, Robin Holliday proposed a model that accounted for heteroduplex formation and gene conversion during recombination. Although it has been supplanted by the double-strand break model (at least for recombination in yeast and higher organisms), it is a useful place to start. It illustrates the critical steps of pairing of homologous duplexes, formation of a heteroduplex, formation of the recombination joint, branch migration and resolution.

The steps in the Holliday Model are illustrated in Figure 8.6.

1. Two homologous chromosomes, each composed of duplex DNA, are **paired** with similar sequences adjacent to each other.
2. An endonuclease **nicks** at corresponding regions of homologous strands of the paired duplexes. This is shown for the strands with the arrow to the right in the figure.
3. The nicked ends dissociate from their complementary strands and each **single strand invades the other duplex**. This occurs in a reciprocal manner to produce a **heteroduplex region**derived from one strand from each parental duplex.
4. DNA ligase **seals the nicks**. The result is a stable **joint molecule**, in which one strand of each parental duplex crosses over into the other duplex.This X-shaped joint is called a **Holliday intermediate** or **Chi structure**.
5. Branch migration then expands the region of heteroduplex. The stable joint can move along the paired duplexes, feeding in more of each invading strand and extending the region of heteroduplex.
6. The recombination intermediate is then **resolved** by nicking a strand in each duplex and ligation.

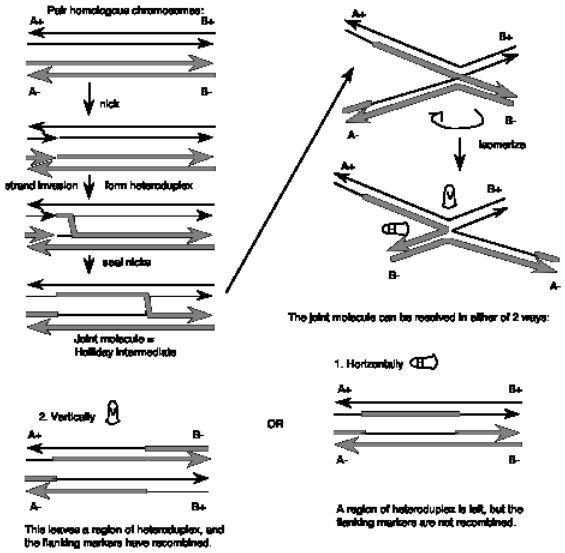


Figure 8.6. Holliday model for general recombination: Single strand invasion. Each of the polynucleotide chains (or strands of the duplex) is shown with a particular orientation, indicated by the arrows. The chromosomes with thick chains and thin chains are homologous. The chains closest to each other in this diagram of the homologous chromosomes are shown in the same orientation. (In contrast to many of the figures in this book, the top strand of each duplex is not necessarily oriented 5' to 3' left to right.) The Holliday model does not specify a particular end (5' or 3') for the invading single strand, but for ease in following the events, the ends are given an orientation in the figure.

Resolution can occur in either of two ways, only one of which results in an exchange of flanking markers after recombination. The two modes of resolution can be visualized by rotating the duplexes so that no strands cross over each other in the illustration (Figure 8.6). In the "horizontal" mode of resolution, the nicks are made in the same DNA strands that were originally nicked in the parental duplexes. After ligation of the two ends, this produces two duplex molecules with a patch of heteroduplex, but no recombination of flanking regions. In contrast, for the "vertical" mode of resolution, the nicks are made in the other strands, i.e. those not nicked in the original parental duplexes. Ligation of these two ends also leaves a patch of heteroduplex, but additionally causes **recombination of flanking regions**. Note that "horizontal" and "vertical" are just convenient designations for the two modes based on the two-dimensional drawings that we can make. The important distinction in terms of genetic outcome is whether the resolution steps target the strands initially cleaved or the other strand.

The steps in this model of general recombination can be viewed in a dynamic form by visiting a web site maintained by geneticists at the University of Wisconsin (URL is www.wisc.edu/genetics/Holliday/index.html). This shows the steps in the Holliday model in a movie, illustrating the actions much more vividly than static diagrams.

The recombinant joint proposed by Holliday has been visualized in electron micrographs of recombining DNA duplexes (Figure 8.7A). It has the proposed X shape. Although this joint is drawn with some distance between the duplexes in illustrations, in fact the two duplexes are juxtaposed, and only a very few base pairs are broken in the Holliday intermediate (Figure 8.7B). The structure is symmetrical , and it is likely that the choice between “horizontal” and “vertical” resolution is a random event by the resolving nuclease. It chooses two strands, but it cannot tell which were initially cleaved and which were not.

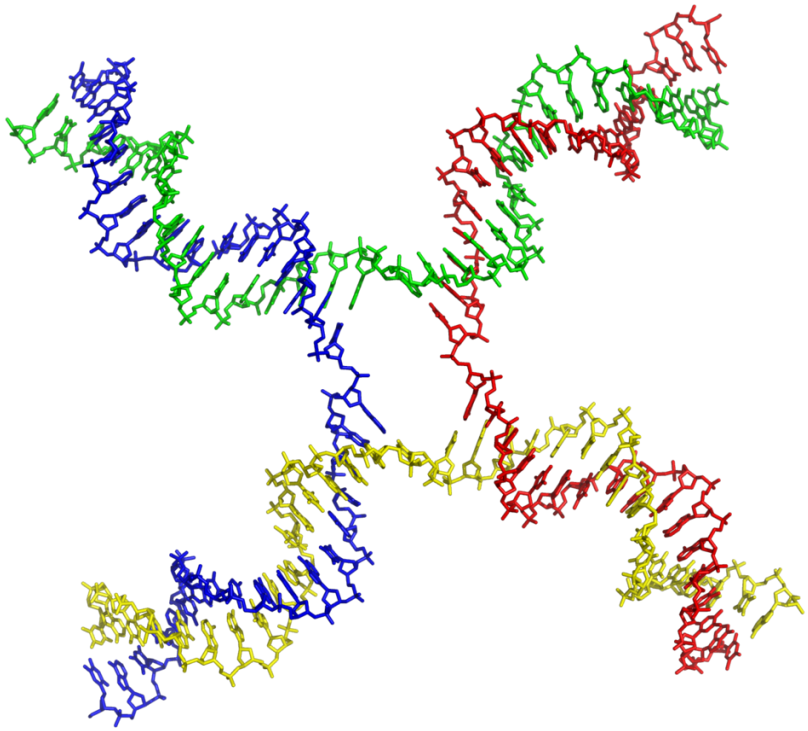


Figure 8.7: Holliday junction from X-ray crystallography of a RuvA-Holliday junction complex (from Hargreaves et al. (1998) Nature Structural Biology 5: 441-4460. For this view, the RuvA protein tetramer was removed and only the phosphodiester backbones of the two duplexes (four strands) are shown. Note the kinks in the DNA in the center of the structure. These correspond to about three nucleotides in each strand that are not paired as in B form DNA. (CC BY-SA 3.0; Zephyri).

Studies of recombination between chromosomes with limited homology have shown that the minimum length of the region required to establish the connection between the recombining duplexes is about 75 bp. If the homology region is shorter than this, the rate of recombination is substantially reduced.

The patch of heteroduplex can be replicated (Figure 8.8) or repaired to generate a gene conversion event. As shown in Figure 8.8, replication through the products of horizontal resolution (from Figure 8.6) will generate a duplex from each strand of the heteroduplex. If we consider the parental chromosomes to be A+C+B+ and A-C-B-, and the heteroduplex to be in gene C, the products of replication can have a the parental C+ converted to a C- but still flanked by A+ and B+ or C- converted to C+ but still flanked by A- and B-. In either case, gene C has changed to a new allele without affecting the flanking markers.

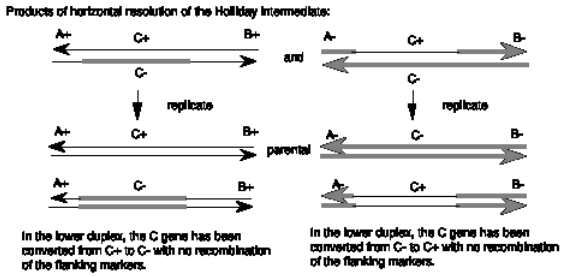


Figure 8.8. Gene conversion can occur by replication through the heteroduplex region.

Although the original Holliday model accounted for many important aspects of recombination (all that were known at the time), some additional information requires changes to the model. For instance, the Holliday model treats both duplexes equally; both are the invader and the target of the strand invasion. Also, no new DNA synthesis is required in the Holliday model. However, subsequent work showed that one of the duplex molecules is the used preferentially as the donor of genetic information. Hence additional models, such as one from Meselson and Radding, incorporated new DNA synthesis at the site of the nick to make and degradation of a strand of the other duplex to generate asymmetry into the two duplexes, with one the donor the other the recipient of DNA. These ideas and others have been incorporated into a new model of recombination involving double strand breaks in the DNAs.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

8.7: Double-strand-break model for Recombination

Several lines of evidence, primarily from studies of recombination in yeast, required changes to the reciprocal exchange of DNA chains initiated at single-strand nicks. As just mentioned, one DNA duplex tended to be the donor of information and the other the recipient, in contrast to the equal exchange predicted by the original Holliday model. Also, in yeast, recombination could be initiated by double-strand breaks. For instance, both DNA strands are cleaved (by the HO endonuclease) to initiate recombination between the *MAT*and *HML*(*R*) loci in mating type switching in yeast. Using plasmids transformed into yeast, it was shown that a double-strand gap in the "aggressor" duplex could be used to initiate recombination, and the gap was repaired during the recombination (this experiment is explored in problem 8.___). In this case, the gap in one duplex was filled by DNA donated from the other substrate. All this evidence was incorporated into a major new model for recombination from Jack Szostak and colleagues in 1983. It is called the **double-strand-break model**. New features in this model (contrasting with the Holliday model) are initiation at double-strand breaks, nuclease digestion of the aggressor duplex, new synthesis and gap repair. However, the fundamental Holliday junction, branch migration and resolution are retained, albeit with somewhat greater complexity because of the additional numbers of Holliday junctions. Although many aspects of the recombination mechanism differ

The steps in the double-strand-break model up to the formation of the joint molecules are diagrammed in Figure 8.9.

1. An endonuclease cleaves both strands of one of the homologous DNA duplexes, shown as thin blue lines in Figure 8.9. This is the **aggressor duplex**, since it initiates the recombination. It is also the **recipient**of genetic information, as will be apparent as we go through the model.
2. The cut is enlarged by an exonuclease to generate a gap with 3' single-stranded termini on the strands.
3. One of the free 3' ends invades a homologous region on the other duplex (shown as thick red lines), called the **donor** duplex. The formation of heteroduplex also generates a **D-loop**(a displacement loop), in which one strand of the donor duplex is displaced.
4. The D-loop is extended as a result of **repair synthesis** primed by the invading 3' end. The D-loop eventually gets large enough to cover the entire gap on the aggressor duplex, i.e. the one initially cleaved by the endonuclease. The newly synthesized DNA uses the DNA from the invaded DNA duplex (thick red line) as the template, so the new DNA has the sequence specified by the invaded DNA.
5. When the displaced strand from the donor (red) extends as far as the other side of the gap on the recipient (thin blue), it will anneal with the other 3' single stranded end at that end of the gap. The displaced strand has now filled the gap on the aggressor duplex, donating its sequence to the duplex that was initially cleaved. **Repair synthesis**catalyzed by DNA polymerase converts the donor D-loop to duplex DNA. During steps 4 and 5, the duplex that was initially invaded serves as the **donor**duplex; i.e. it provides genetic information during this phase of repair synthesis. Conversely, the aggressor duplex is the recipient of genetic information. Note that the single strand invasion models predict the opposite, where the initial invading strand is the donor of the genetic information.
6. DNA ligase will seal the nicks, one on the left side of the diagram in Figure 8.9 and the other on the right side. Although the latter is between a strand on the bottom duplex and a strand on the top duplex, it is equivalent to the ligation in the first nick (the apparent physical separation is an artifact of the drawing). In both cases, sealing the nick forms a Holliday junction.

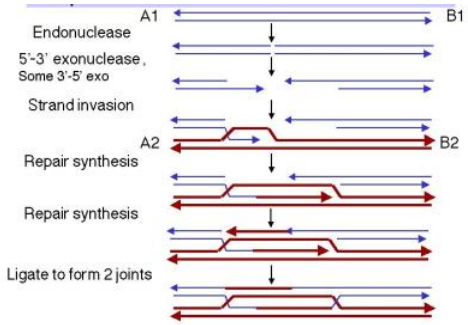


Figure 8.9. Steps in the double-strand-break model for recombination, from initiation to formation of the recombinant joints.

At this point, the recombination intermediate has **two recombinant joints** (Holliday junctions). The original gap in the aggressor duplex has been filled with DNA donated by the invaded duplex. The filled gap is now **flanked by heteroduplex**. The heteroduplexes are arranged **asymmetrically**, with one to the left of the filled gap on the aggressor duplex and one to the right of the filled gap on the donor duplex. Branch migration can extend the regions of heteroduplex from each Holliday junction.

The recombination intermediate can now be resolved. The presence of two recombination joints adds some complexity, but the process is essentially the same as discussed for the Holliday model. Each joint can be resolved horizontally or vertically. The key factor is whether the joints are resolved in the same mode or sense (both horizontally or both vertically) or in different modes.

If both joints are resolved the same sense (Figure 8.10), the original duplexes will be released, each with a region of altered genetic information that is a "footprint" of the exchange event. That region of altered information is the original gap, plus or minus the regions covered by branch migration. For instance, if both joints are resolved by cutting the originally cleaved strands ("horizontally" in our diagram of the Holliday model), then you have no crossover at either joint. If both joints are resolved by cleaving the strands not cut originally ("vertically" in our diagram of the Holliday model), then you have a crossover at both joints. This closely spaced double crossover will produce no recombination of flanking markers.

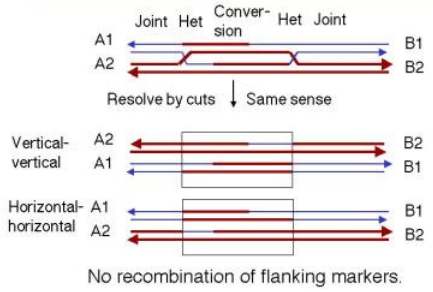


Figure 8.10. Resolution of intermediates in the double-strand-break model by cutting the recombinant joints in the same mode or sense. The box outlines the region between the two resolved junctions.

In contrast, if each joint is resolved in opposite directions (Figure 8.11), then there will be recombination between flanking markers. That is, one joint will not give a crossover and the other one will.

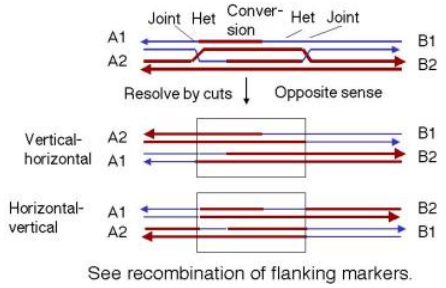


Figure 8.11.Resolution of intermediates in the double-strand-break model by cutting the recombinant joints in the opposite mode or sense.

Several features distinguish the double-strand-break model from the single-strand nick model initially proposed by Holliday. In the double-strand-break model, the region corresponding to the original gap now has the sequence of the donor duplex in both molecules. This is flanked by heteroduplexes at each end, one on each duplex. Hence the arrangement of heteroduplex is **asymmetric**; i.e. there is a different heteroduplex in each duplex molecule. Part of one duplex molecule has been converted to the sequence of the other (the recipient, initiating duplex has been converted to the sequence of the donor). In the single strand invasion model, each DNA duplex has heteroduplex material covering the region from the initial site of exchange to the migrating branch, i.e. the heteroduplexes are symmetric. In variations of the model (Meselson-Radding) in which some DNA is degraded and re-synthesized, the initiating chromosome is the donor of the genetic information.

These models also have many important features in common. Steps that are common to all the models include the generation of a single strand of DNA at an end, a search for homology, strand invasion or strand exchange to form a joint molecule, branch migration, and resolution. Enzymes catalyzing each of these steps have been isolated and characterized. This is the topic of the rest of this chapter.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

8.8: Enzymes required for recombination in E. coli

The initial steps in finding enzymes that carry out recombination were genetic screens for mutants of *E. coli* that are defective in recombination. Assays were developed to test for recombination, and mutants that showed a decrease in recombination frequency were isolated. These were assigned to complementation groups called *recA*, *recB*, *recC*, *recD*, and so forth. Roughly 20 different genes (different *rec* complementation groups) have been identified in *E. coli*. Each gene encodes an enzyme or enzyme subunit required for recombination.

Many of these genes have been cloned and their encoded products characterized in terms of a variety of enzymatic functions. However, we still do not have a clear picture of how all these enzymes work together to carry out recombination, nor has recombination has been reconstituted *in vitro* from purified components. Further complicating matters is the presence of multiple pathways for recombination. Much work remains to be done to completely understand recombination at a biochemical level. Despite this, the array of recombination enzymes gives us at least a partial view of the mechanisms of recombination. Also, the enzymes characterized in *E. coli* have homologs and counterparts in other species. Some aspects of the recombination machinery appear to be conserved across a wide phylogenetic range.

The major enzymatic steps are outlined in Figure 8.12. Three different pathways have been characterized that differ in the steps used to generate the invading single strand of DNA. All three pathways use RecA for homologous pairing and strand exchange, RuvA and RuvB for branch migration, and RuvC and DNA ligase for resolution. These steps and enzymes will be considered individually in the following sections.

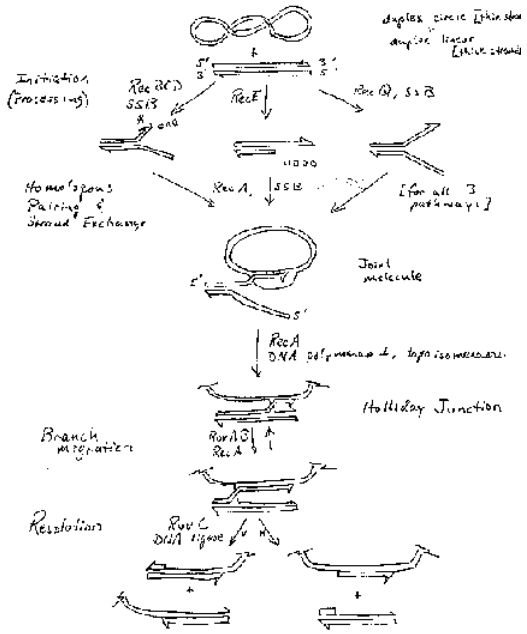


Figure 8.12. Enzymatic Steps in Recombination. Three pathways for recombination are shown, starting with a covalently closed, supercoiled circle (with each strand of the duplex shown as a thin line) and a linear duplex (with each strand shown as a thick white line) as the substrates. The three pathways differ in the enzymes used for initiation, but subsequent steps use enzymes common to all three. Adapted from Kowalczykowski, et al. (1994) Microbiological Reviews, 58:401-465.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

8.9: Generation of Single Strands

One of the major pathways for generating 3’ single-stranded termini uses the **RecBCD enzyme**, also known as exonuclease V (Figure 8.13). The three subunits of this enzyme are encoded by the genes *recB*, *recC*, and *recD*. Each model for recombination requires a single-strand with with a free end for strand invasion, and this enzyme does so, but with several unexpected features.

RecBCD has multiple functions, and it can switch activities. It is a **helicase** (in the presence of SSB), an **ATPase** and a **nuclease**. The nuclease can be a 3’ to 5’ exonuclease, and endonuclease or a 5’ to 3’ exonuclease, at different steps of the process.

The **helicase** activity of the RecBCD enzyme initiates unwinding only on DNA containing a free duplex end. It binds to the duplex end, using the energy of ATP hydrolysis to travel along the duplex, unwinding the DNA. The enzyme complex tracks along the top strand faster than it does on the bottom strand, so single-stranded loops emerge, getting progressively larger as it moves down the duplex. These loops can be visualized in electron micrographs. RecBCD is also a **3' to 5' exonuclease** during this phase, removing the end of one of the unwound strands (Figure 8.13).

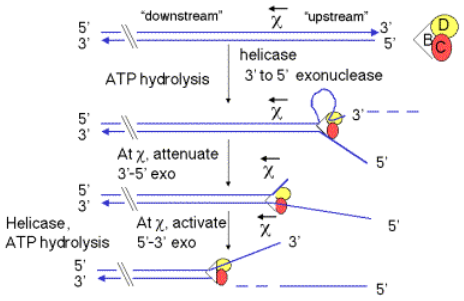


Figure 8.13. Generation of a 3'-single-stranded terminus by RecBCD enzyme

The activities of the RecBCD enzyme change at particular sequences in the DNA called **chi sites**(for the Greek letter c). The sequence of a chi site is 5' GCTGGTGG; this occurs about once every 4 kb on the *E. coli*genome. Genetic experiments show that RecBCD **promotes recombination** most frequently at chi sites. These sites were first discovered as mutations in bacteriophage l that led to increased recombination at those sites. These mutations altered the l sequence at the site of the mutation to become a chi site (GCTGGTGG).

When the RecBCD enzyme encounters a chi site, it will leave an **extruded single strand** close to this site (4 to 6 nucleotides 3' to it). A chi site serves as a signal to RecBCD to **shift the polarity of its exonuclease function**. Before reaching the chi site, RecBCD acts primarily as a 3’ to 5’ exonuclease, e.g. working on the top strand in Figure 8.13. At the chi site, the 3’ to 5’ exonuclease function is suppressed, and afterthe chi site, RecBCD converts to a 5’ to 3’ exonuclease, now working on the other strand (e.g. the bottom strand in Figure 8.13). Presumably, the strand that will be the substrate for the 5’ to 3’ exonuclease is nicked in concert with this conversion in polarity of the exonuclease. This process leaves the chi site at the 3’ end of a single stranded DNA. This is the substrate to which RecA can bind to initiate strand exchange (see below).

Some tests of the models for recombination have examined whether chi sites serve preferentially as either donors or recipients of the DNA during recombination. However, both results have been obtained, which makes it difficult to tie this activity precisely into either model for recombination. The genetic evidence is clear, however, that it is needed for one major pathway of recombination.

Exercise 8.4

What are the predictions of the Holliday model and the double-strand-break model for whether chi sites would be used as donors or recipients of genetic information during recombination?

An alternative pathway for generating single-strand ends for recombination uses the enzyme **RecE**, also known as exonuclease VIII. This pathway is revealed in *recBCD-* mutants. RecE is a 5’ to 3’ exonuclease that digests double-stranded linear DNA, thereby generating single-stranded 3' tails. RecE is encoded on a cryptic plasmid in *E. coli*. It is similar to the *red*exonuclease encoded by bacteriophage l.

A third pathway used the **RecQ**helicase, which is also a DNA-dependent ATPase. This pathway is revealed in *recBCD- recE-* mutants. The result of its helicase activity, in the presence of SSB, is the formation of a DNA molecule with single-stranded 3' tails, which can be used for strand invasion.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

8.10: Synapsis and Invasion of Single Strands

The pairing of the two recombining DNA molecules (**synapsis**) and **invasion of a single strand** from the initiating duplex into the other duplex are both catalyzed by the multi-functional protein **RecA**. This invasion of the duplex DNA by a single stranded DNA results in the replacement of one of the strands of the original duplex with the invading strand, and the replaced strand is displaced from the duplex. Hence this reaction can also be called **strand assimilation** or **strand exchange**. RecA has many activities, including stimulating the protease function of LexA and UmuD (see Chapter 7), binding to and coating single-stranded DNA, stimulating homologous pairing between single-stranded and duplex DNA, assimilating single-stranded DNA into a duplex, and catalyzing the hydrolysis of ATP in the presence of DNA (i.e. it is a DNA-dependent ATPase). It is required in all 3 pathways for recombination. For instance, the DNA molecule with a single-stranded 3' end generated by the RecBCD enzyme can be assimilated into a homologous region of another duplex, catalyzed by RecA and requiring the hydrolysis of ATP (Figure 8.14).

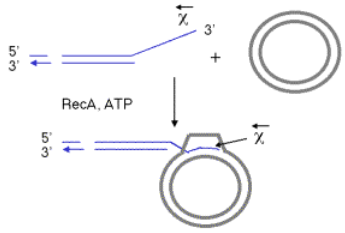


Figure 8.14.The single strand of DNA with a free 3' end, generated by the RecBCD enzyme, can invade a homologous duplex DNA molecule in a reaction promoted by RecA. The chi site is close to the 3' end of the single strand. The invading DNA molecule is shown with a thin, blue line for each strand. The target molecule is a duplex circle, shown as a thick gray line for each strand. ATP is required for this reaction, and it is hydrolyzed to ADP and phosphate.

The process of single-strand assimilation occurs in three steps, as illustrated in Figure 8.15. First, RecA polymerizes onto single-stranded DNA in the presence of ATP to form the **presynaptic filament**. The single strand of DNA lies within a deep groove of the RecA protein, and many RecA-ATP molecules coat the single-stranded DNA. One molecule of the RecA protein covers 3 to 5 nucleotides of single-stranded DNA. The nucleotides are extended axially so they are about 5 Angstroms apart in the single-stranded DNA, about 1.5 times longer than in the absence of RecA-ATP.

Next, the presynaptic filament aligns with homologous regions in the duplex DNA. A substantial length of the three strands are held together by a polumer of RecA-ATP molecules. The aligned duplex and single strand forms a **paranemic joint**, meaning that the single strand is not intertwined with the double strand at this point. The duplex DNA, like the single-stranded DNA, is extended to about 1.5 times longer than in normal B form DNA (18.6 bp per turn). This extension is thought to be important in homologous pairing.

Finally, the strands are exchanged from to form a **plectonemic joint**. In this stage, the invading single strand is now intertwined with the complementary strand in the duplex, and one strand of the invaded duplex is now displaced. In *E. coli*, exchange occurs in a 5' to 3' direction relative to the single strand and requires ATP hydrolysis. In contrast, the yeast homolog, Rad51, causes the single-strand to invade with the opposite polarity, i.e. 3' to 5'. Thus the direction of this polarity is not a universally conserved feature of recombination mechanisms.

The product of strand assimilation is a heteroduplex in which one strand of the duplex was the original single-stranded DNA. The other strand of the original duplex is displaced.

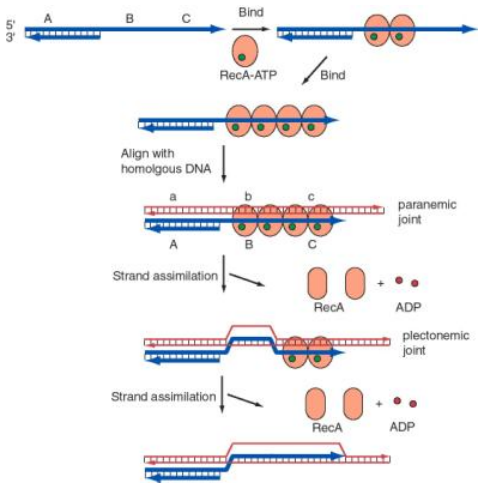


Figure 8.15. Role of RecA in assimilation of single-stranded DNA. A DNA molecule with a single-stranded 3’ end is shown with a thick blue line for each strand. A, B, and C denote particular DNA sequences. A homologous duplex is shown with thin red lines for each strand, with a, b, and c homologous to A, B and C, respectively. RecA is an orange-brown oval. It has a different conformation (shape) when ATP (green circle) is bound. The ATPase activity of RecA generates ADP (red circle) and an altered conformation of RecA, which dissociates as the single strand is assimilated. The single strand enters the duplex with a 5’ to 3’ polarity (relative to the orientation of the invading single strand).

Many details of the activity of RecA have been revealed by *in vitro*assays for single strand assimilation, or strand exchange. The DNA substrates for strand exchange catalyzed by RecA must meet three requirements. There must be a region of single stranded DNA on which RecA can bind and polymerize, the two molecules undergoing strand exchange must have a region of homology, and there must be a afree end within the region of homology. The latter requirement can be overcome by providing a topoisomerase.

One such assay is the conversion of a single-stranded circular DNA to a duplex circle (Figure 8.16). The substrates for this reaction are a circular single-stranded DNA and a homologous linear duplex. These are mixed together in the presence of RecA and ATP. Many RecA-ATP molecules coat the single-stranded circle to form the nucleoprotein presynaptic filament, as discussed above. During synapsis, annealing is initiated with the 3' end of the strand complementary to the single-stranded circle. Thus the single strand invades with 5' to 3' polarity (with reference to its own polarity). Strand displacement, driven by ATP hydrolysis to dissociate the RecA, results in the formation of a nicked circle (one strand of which was the original single-stranded circle) and a linear single strand of DNA.

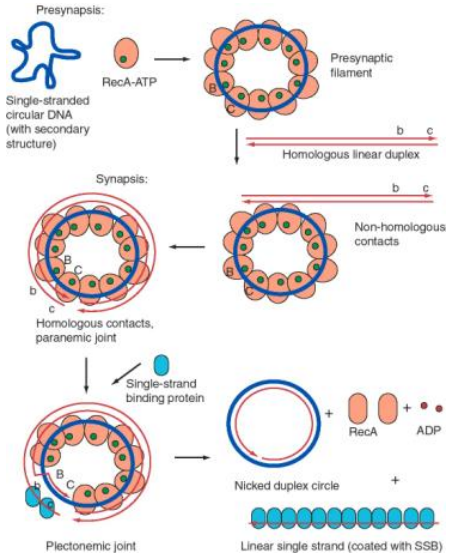


Figure 8.16. An in vitro assay for single-strand assimilation catalyzed by RecA plus ATP. Strand exchange between an invading single-stranded circle (thick blue line) and a linear duplex DNA (thin red lines), mediated by RecA plus ATP, results in a nicked duplex circle and a single-stranded linear DNA coated with single-stranded binding protein, or SSB. Regions B and C are homologous to regions b and c, respectively; they are shown as markers but the entire DNA in both molecules is homologous. SSB helps to stimulate this reaction by helping RecA overcome secondary structure in the single-stranded DNA.

Exercise

Try to relate this *in vitro* assay to the steps in the double-strand-break model for recombination. What step(s) in the model does this mimic? What else is needed for to get to the recombinant joints (Holliday junctions)?

The structure of *E. coli*RecA bound by ADP, both monomer and polymer, have been solved by X-ray crystallography. As shown in Figure 8.17, the central domain has the binding site for ATP and ADP, and is presumably the site of binding of the single-stranded and double-stranded DNA. The domains extending away from the central region are involved in polymerization of RecA proteins and in interactions between the presynaptic fibers.

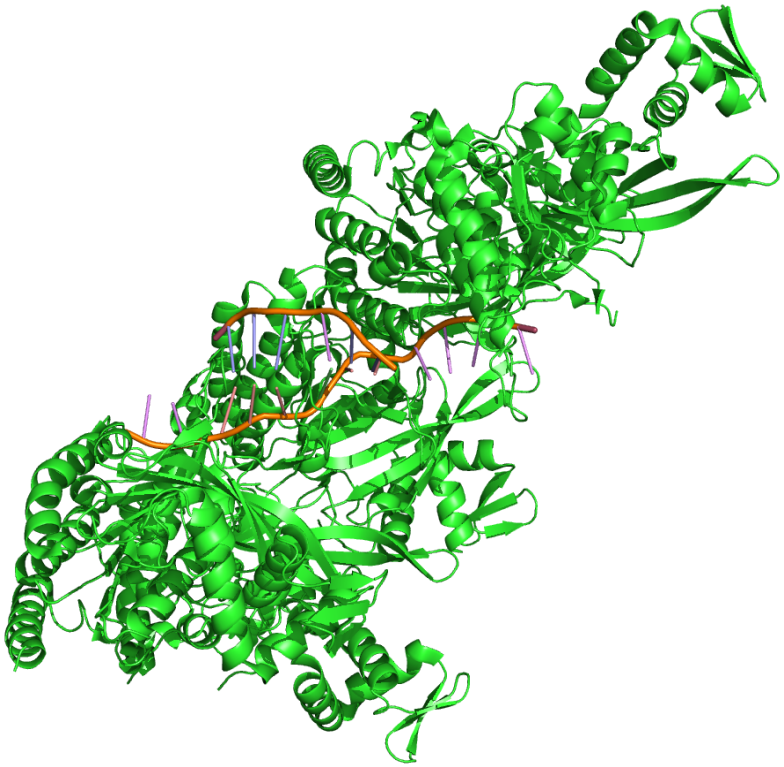


Figure 8.17. A static view of the three-dimensional structure of RecA, as determined by R. M. Story and T. A. Steitz (1992) "Structure of the recA protein-ADP complex" Nature 355: 374-376. Alpha helices are shown as green cylinders with the peptide backbone wrapped around them. Beta-sheets are yellow-brown arrows, and other regions of the peptide backbone are blue. The ADP is shown as a wire diagram, with C atoms gray, N atoms white, O atoms red and P atoms orange. Atomic coordinates were obtained from the MMDB server at NCBI and rendered in CN3D. A screen shot of one view is shown. Files for virtual 3-D viewing are available at the course web site. (CC BY-SA 3.0; Emw).

Proteins homologous to the *E. coli* RecA are found in yeast (Rad51 and Dmc1) and in mice (Rad51). Given the universality of recombination, it is likely that homologs will be found in virtually all species. Mutations in the *E. coli* recA gene reduce conjugational recombination by as much as 10,000 fold, so it is clear that RecA plays a central role in recombination. However, null mutations in *recA* are **not** lethal, nor are null mutations in the yeast homologs *RAD51* and *DMC1*. In contrast, mice homozygous for a knockout mutation in the *Rad51* gene die very early in development, at the 4-cell stage. This indicates that in mice, this RecA homolog is playing a novel role in replication or repair, presumably in addition to its role in recombination.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

8.11: Branch Migration

The movement of a Holliday junction to generate additional heteroduplex requires two proteins. One is the **RuvA** tetramer, which recognizes the structure of the Holliday junction. A rendering of the structure derived from X-ray crystallographic analysis of the RuvA-Holliday junction crystals is shown in Figure 8.18.

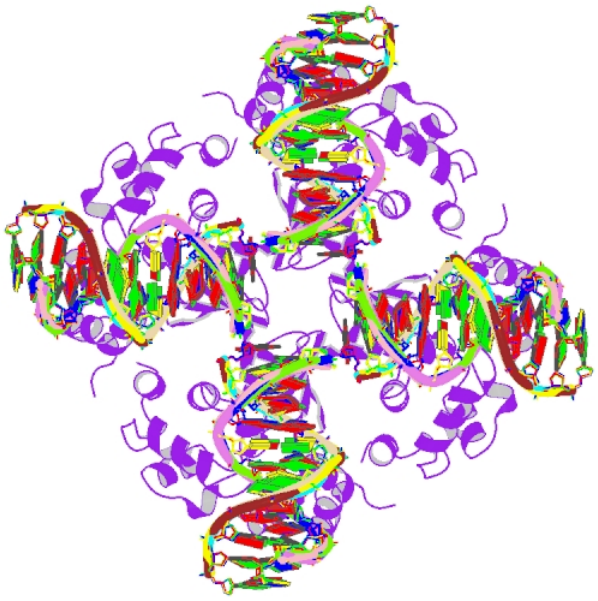


Figure 8.18: Three-dimensional structure of the RuvA tetramer complexed with a Holliday junction [from Hargreaves et al. (1998) Nature Structural Biology 5: 441-4460]. For the RuvA protein, alpha helices are green cylinders, beta sheets are brown arrows and loops are blue. The four strands of the two duplexes in the Holliday junction are red lines. The atomic coordinates were downloaded from the Molecular Structure database at NCBI, rendered in Cn3D v.3.0, and a pict file obtained as a screen shot. The kin file for viewing the virtual 3-D image on your own computer is accessible at the course web site.

RuvB is an ATPase. It forms hexameric rings that provide the motor for branch migration. As illustrated in Figure 8.19, RuvA tetramers recognize the Holliday junction, and RuvB uses the energy of ATP hydrolysis to unwind the parental duplexes and form heteroduplexes between them.

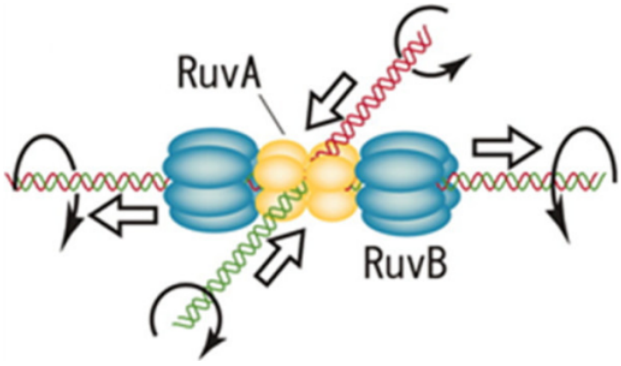


Figure 8.19: Branch migration of RuvA-RuvB in solution. The four monomers of RuvA combine around a central pen to accommodate the square planar configuration of the Holliday junction in which the four DNA duplex arms attach to grooves on the concave surface of RuvA. Through ATP hydrolysis, the two hexameric RuvB rings encircle and translocate the dsDNA arms. Curved arrows indicate rotation of DNA while the thick arrows indicate translocation of dsDNA through the junction. DNA rotation during Holliday junction branch occurs at a V(max) of 1.6 revolutions per second, or 8.3 bp per second. Adopted from reference Eggleston, A. K. and West, S. C. (1996) Trends in Genetics 12: 20-25. (Public Domain).

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

8.12: Resolution

RuvC is the endonuclease that cleaves the Holliday junctions (Figure 8.20). It forms dimers that bind to the Holliday junction; recent data indicate an interaction among RuvA, RuvB and RuvC as a complex at the Holliday junction. The structure of the RuvA-Holliday junction complex (Figure 8.18) suggests that the open structure of the junction stabilized by the binding of RuvA may expose a surface that is recognized by Ruv C for cleavage. RuvC cleaves symmetrically, in two strands with the same nearly identical sequences, thereby producing ligatable products.

The preferred site of cleavage by RuvC is 5' WTT'S, where W = A or T and S = G or C, and ' is the site of cleavage. RuvC can cut strands for either horizontal or vertical resolution. Strand choice is influenced by the sequence preference and also by the presence of RecA protein, which favors vertical cleavage (i.e. to cause recombination of flanking markers).

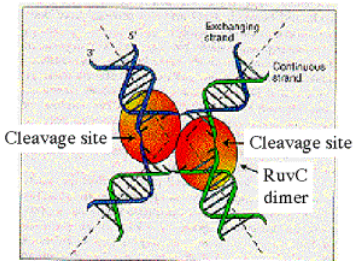


Figure 8.20. Resolution requires cleavage by RuvC dimers. Adapted from Eggleston, A. K. and West, S. C. (1996) Trends in Genetics 12: 20-25.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

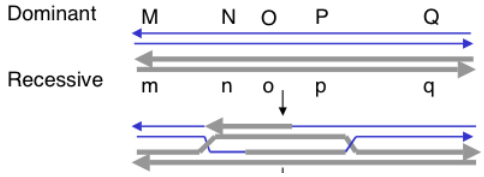
8.E: Recombination of DNA (Exercises)

Question 8.6. According to the Holliday model for genetic recombination, what factor determines the length of the heteroduplex in the recombination intermediate?

Question 8.7. Holliday junctions can be resolved in two different ways. What are the consequences of the strand choice used in resolution?

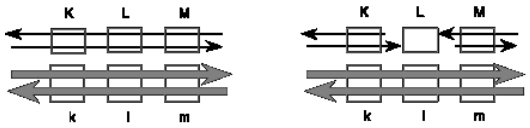
Question 8.8. Why do models for recombination include the generation of heteroduplexes in the products?

Question 8.9. Consider two DNA duplexes that undergo recombination by the double-strand break mechanism. The parental duplex indicated by thin lines has dominant alleles for genes M, N, O, P, and Q, and the parental duplex shown in thick lines has recessive alleles, indicated by the lower case letters. The recombination intermediate with two Holliday structures is also shown.



- a) What duplexes result from resolution of the left Holliday junction vertically and the right junction horizontally?
- b) After the vertical-horizontal resolution, what will the genotype be of the recombination products with respect to the flanking markers M and Q? In answering, use a slash to separate the designation for the 2 chromosomes, each of which is indicated by a line (i.e. the parental arrangement is M___Q / m___q).
- c) If the products of the vertical-horizontal resolution were separated by meiosis, and then replicated by mitosis to generate 8 spores in an ordered array (as in the *Ascomycete*fungi), what would be the phenotype of the spores with respect to alleles of gene O? Assume that the sister chromatids of these chromosomes did not undergo recombination in this region (i.e. one parental duplex from each homologous chromosome remains from the 4n stage).

For the **next 3** problems, consider two DNA duplexes that undergo recombination by the double-strand break mechanism. The parental duplex denoted by thin black lines has dominant alleles (capital letters) for genes (or loci) K, L, and M, and the parental duplex denoted by thick gray lines has recessive alleles, indicated by k, l, m. The genes are shown as boxes with gray outlines. In the diagram on the right, the double strand break has been made in the L gene in the black duplex and expanded by the action of exonucleases.

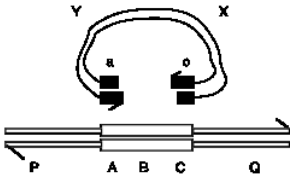


Question 8.10. When recombination proceeds by the double-strand break mechanism, what is the structure of the intermediate with Holliday junctions, prior to branch migration? Please draw the structure, and distinguish between the DNA chains from the parental duplexes.

Question 8.11. If the recombination intermediates are resolved to generate a chromosome with the dominant K allele of the K gene and the recessive m allele of the M gene on the same chromosome (K___m), which allele (dominant L or recessive l) will be be at the L, or middle, gene?

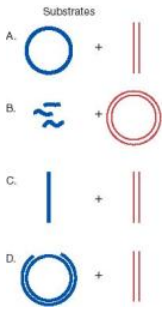
Question 8.12. If the left Holliday junction slid leftward by branch migration all the way through the K gene (K allele on the black duplex, k allele on the gray duplex), what will the structure of the product be, prior to resolution?

Question 8.13. According to the original Holliday model and the double-strand break model for recombination, what are the predicted outcomes of recombination between a linear duplex chromosome and a (formerly) circular duplex carrying a gap in the region of homology? The homology is denoted by the boxes labeled ABC on the linear duplex and ac on the gapped circle. The regions flanking the homology (P and Q versus X and Y) are not homologous.



The results of an experiment like this are reported in Orr-Weaver, T. L., Szostak, J. W. and Rothstein, R. J. (1981) Yeast transformation: a model system for the study of recombination. Proc. Natl. Acad. Sci. USA 78: 6354-6358. These data were instrumental in formulating the double-strand-break model for recombination.

Question 8.14.A variety of *in vitro* assays have been developed for strand exchange catalyzed by RecA. For each of the substrates shown below, what are the expected products when incubated with RecA and ATP (and SSB to facilitate removal of secondary structures from single-stranded DNA)? In practice, the reactions proceed in stages and one can see intermediates, but answer in terms of the final products after the reaction has gone to completion.



In each case, the molecule with at least partical single stranded region is shown with thick blue strands, and the duplex that will be invaded is shown with thin red lines. The DNA substrates are as follows.

- A. Single-stranded circle and duplex linear. The two substrates are the same length and are homologous throughout.
- B. Single-stranded short linear fragments and duplex circle. The short fragments are homologous to the circle.
- C. Single-stranded linear and duplex linear. The two substrates are the same length and are homologous throughout.
- D. Gapped circle and duplex linear. The intact strand of the circle is the same length as the linear and is homologous throughout. The gapped strand of the circle is complementary to the intact strand, of course, but is just shorter.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

CHAPTER OVERVIEW

9. Transposition of DNA

The final method of changing the DNA in a genome that we will consider is **transposition**, which is the movement of DNA from one location to another. Segments of DNA with this ability to move are called **transposable elements**. Transposable elements were formerly thought to be found only in a few species, but now they are recognized as components of the genomes of virtually all species.

References

- Shapiro, J. A. (editor) (1983) Mobile Genetic Elements (Academic Press, Inc., New York).
- Fedoroff, N. and Botstein, D. (1992) The Dynamic Genome: Barbara McClintock's Ideas in the Century of Genetics (Cold Spring Harbor Press, Plainview, NY).
- McClintock, B. (1952) Chromosome organization and genic expression. Cold Spring Harbor Symposium on Quantitative Biology 16: 13-47.
- Fedoroff, N., Wessler, S., and Shure, M. (1983) Isolation of the transposable maize controlling elements Ac and Ds. Cell 35:235-242.Boeke, J.D., Garfinkel, D.J., Styles, C.A., and Fink G.R. (1985) Ty elements transpose through an RNA intermediate. Cell 40:491-500
- Kazazian, H.H. Jr, Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., and Antonarakis, S.E. (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature 332:164-166.
- International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921. The material from pages 879-889 covers human repeats and transposable elements.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

9.1: Transposable Elements (Transposons)

Transposable elements (both active and inactive) occupy approximately half the human genome and a substantially greater fraction of some plant genomes! These movable elements are ubiquitous in the biosphere, and are highly successful in propagating themselves. We now realize that some transposable elements are also viruses, for instance, some retroviruses can integrate into a host genome to form endogenous retroviruses. Indeed, some viruses may be derived from natural transposable elements and vice versa. Since viruses move between individuals, at least some transposable elements can move between genomes (between individuals) as well as within an individual's genome. Given their prevalence in genomes, the function (if any) of transposable elements has been much discussed but is little understood. It is not even clear whether transposable elements should be considered an integral part of a species' genome, or if they are successful parasites. They do have important effects on genes and their phenotypes, and they are the subject of intense investigation.

Transposition is related to replication, recombination and repair. The process of moving from one place to another involves a type of recombination, insertions of transposable elements can cause mutations, and some transpositions are replicative, generating a new copy while leaving the old copy intact. However, this ability to move is a unique property of transposable elements, and warrants treatment by itself.

Properties and effects of transposable elements

The defining property of **transposable elements** is their **mobility**; i.e. they are genetic elements that can move from one position to another in the genome. Beyond the common property of mobility, transposable elements show considerable diversity. Some move by DNA intermediates, and others move by RNA intermediates. Much of the mechanism of transposition is distinctive for these two classes, but all transposable elements effectively insert at staggered breaks in chromosomes. Some transposable elements move in a **replicative** manner, whereas others are **nonreplicative**, i.e. they move without making a copy of themselves.

Transposable elements are major forces in the evolution and rearrangement of genomes (Figure 9.1). Some transposition events **inactivate** genes, since the coding potential or expression of a gene is disrupted by insertion of the transposable element. A classic example is the *r* allele (*rugosus*) of the gene encoding a starch branching enzyme in peas is nonfunctional due to the insertion of a transposable element. This allele causes the wrinkled pea phenotype in homozygotes originally studied by Mendel. In other cases, transposition can **activate** nearby genes by bringing an enhancer of transcription (within the transposable element) close enough to a gene to stimulate its expression. If the target gene is not usually expressed in a certain cell type, this activation can lead to pathology, such as activation of a proto-oncogene causing a cell to become cancerous. In other cases, **no obvious phenotype** results from the transposition. A particular type of transposable element can activate, inactivate or have no effect on nearby genes, depending on exactly where it inserts, it's orientation and other factors.

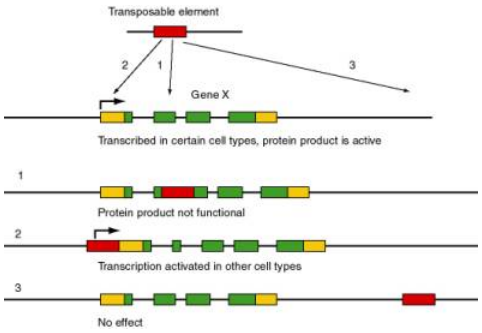


Figure 9.1. Possible effects of movement of a transposable element in the function and expression of the target gene. The transposable element is shown as a red rectangle, and the target gene (X) is composed of multiple exons. Protein coding regions of exons are green and untranslated regions are gold. The angled arrow indicates the start site for transcription.

Transposable elements can cause deletions or inversions of DNA. When transposition generates two copies of the same sequence in the same orientation, recombination can delete the DNA between them. If the two copies are in the opposite orientations, recombination will invert the DNA between them.

As part of the mechanism of transposition, additional DNA sequences can be mobilized. DNA located between two copies of a transposable element can be moved together with them when they move. In this manner, transposition can move DNA sequences that are not normally part of a transposable element to new locations. Indeed, "host" sequences can be acquired by viruses and propagated by infection of other individuals. This may be a natural means for evolving new strains of viruses. One of the most

striking examples is the acquisition and modification of a proto-oncogene, such as cellular *c-src*, by a retrovirus to generate a modified, transforming form of the gene, called *v-src*. These and related observations provided insights into the progression of events that turn a normal cell into a cancerous one. They also point to the continual acquisition (and possibly deletion) of information from host genomes as a natural part of the evolution of viruses.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

9.2: Are Transposons Parasites or Symbionts?

Do the transposable elements confer some selective advantage on the "host"? Or are they merely parasitic or "selfish," existing only to increase the number of copies of the element? This critical issue is a continuing controversy. As just mentioned, certain results of transposition can be detrimental, leading to a loss of function or changes in regulation of the genes at the site of integration after movement. Also, we are starting to appreciate the intimate connection between viruses and transposable elements. Thus one can view many transposable elements as parasites on the genome. The number of transposable elements can expand rapidly in a genome. For instance, it appears that transposable elements making up a majority of the genome of maize are not abundant in the wild parent, teosinte. Thus this massive expansion has occurred since the domestication of corn, roughly within the past 10,000 years.

However, other studies indicate that the presence of transposable elements is beneficial to an organism. Two strains of bacteria, one with a normal number of transposable elements and the other with many fewer, can be grown in competitive conditions. The strain with the higher number of transposable elements has a growth advantage under these conditions. Various proposals have been made as to the nature of that advantage. One intriguing possibility is that the mechanism of transposition affords an opportunity to seal chromosome breaks. Other possible benefits have not been excluded. Thus the relationship between transposable elements and their hosts may be as much symbiotic as parasitic. Resolving these issues is an interesting challenge for future research.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

9.5: Transposition occurs by Insertion into Staggered Breaks

Staggered Break in a chromosome

A common property of virtually all transposable elements is that they move by inserting into a staggered break in a chromosome, i.e. one strand is slightly longer than the other at the break (Figure 9.9). The first indication of this was the observation that the same short DNA sequence is found on each side of a transposable element. The sequence within these **flanking direct repeats (FDRs)** is distinctive for each copy of the transposable element, but the size of the FDR is characteristic of a particular family of transposable elements. Some families of transposable elements have FDRs as short as 4 bp and other families have FDRs as long as 12 bp. However, within a particular family, the sequence of the FDR will differ between individual copies. These FDRs are hallmarks of transposable elements.

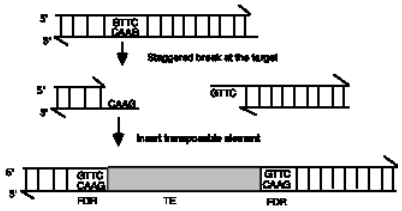


Figure 9.9. Flanking direct repeats are generated by insertions at staggered breaks.

Since the FDRs are distinctive for each copy, they are not part of the transposable element themselves. Some families of transposable elements do have repeated sequences at their flanks that are identical for all members of the family, but these are integral parts of the transposable element. The variation in sequence of the FDRs indicates that they are generated from the target sites for the transposition events. If the transposable element inserted into a break in the chromosome that left a short overhang (one strand longer than the other), and this overhang were filled in by DNA polymerase as part of the transposition, then the sequence of that overhang would be duplicated on each side of the new copy. Such a break with an overhang is called a staggered break. The size of the staggered break would determine the size of the FDR.

Mechanistic studies of the enzymes used for transposition have shown that such staggered breaks are made at the target site prior to integration and are repaired as part of the process of transposition (see below). The staggered breaks are used in transposition both by DNA intermediates and by RNA intermediates.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

9.6: Classes of Transposable Elements

The two major classes of transposable elements are defined by the intermediates in the transposition process. One class moves by DNA intermediates, using transposases and DNA polymerases to catalyze transposition. The other class moves by RNA intermediates, using RNA polymerase, endonucleases and reverse transcriptase to catalyze the process. Both classes are abundant in many species, but some groups of organisms have a preponderance of one or the other. For instance, bacteria have mainly the DNA intermediate class of transposable elements, whereas the predominant transposable elements in mammalian genomes move by RNA intermediates.

- Transposable elements that move via DNA intermediates
- Transposable elements that move via RNA intermediates

Transposable elements that move via DNA intermediates

Among the most thoroughly characterized transposable elements are those that move by DNA intermediates. In bacteria, these are either short insertion sequences or longer transposons.

An **insertion sequences**, or **IS**, is a short DNA sequence that moves from one location to another. They were first recognized by the mutations they cause by inserting into bacterial genes. Different insertion sequences range in size from about 800 bp to 2000 bp. The DNA sequence of an IS has inverted repeats (about 10 to 40 bp) at its termini (Figure 9.10A.). Note that this is different from the FDRs, which are duplications of the target site. The inverted repeats are part of the IS element itself. The sequences of the inverted repeats at each end of the IS are very similar but not necessarily identical. Each family of insertion sequence in a species is named IS followed by a number, e.g. IS1, IS10, etc.

An insertion sequence encodes a **transposase** enzyme that catalyzes the transposition. The amount of transposase is well regulated and is the primary determinant of the rate of transposition. **Transposons** are larger transposable elements, ranging in size from 2500 to 21,000 bp. They usually encode a **drug resistance gene or other marker** besides the functions required for transposition (Figure 9.10.B.). One type of transposon, called a **composite transposon**, has an IS element at each end (Figure 9.10.C.). One or both IS elements may be functional; these encode the transposition function for this class of transposons. The IS elements flank the drug resistance gene (or other selectable marker). It is likely that the composite transposon evolved when two IS elements inserted on both sides of a gene. The IS elements at the end could either move by themselves or they can recognize the ends of the closely spaced IS elements and move them together with the DNA between them. If the DNA between the IS elements confers a selective advantage when transposed, then it will become fixed in a population.

Exercise 9.3

What are the predictions of this model for formation of a composite transposon for the situation in which a transposon in a small circular replicon, such as a plasmid?

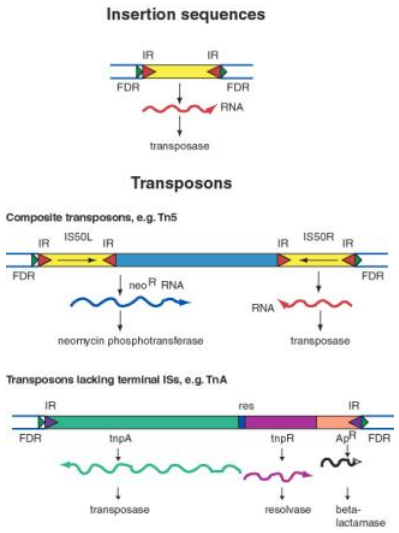


Figure 9.10. General structure of insertion sequences and transposons. Flanking direct repeats (FDRs) are shown as green triangles, inverted repeats (IRs) are red or purple triangles, insertion sequences (ISs) are yellow boxes with red triangles at the end, and other genes are boxes of different colors. The boxes and triangles include both strands of duplex DNA. DNA outside the FDRs is shown as one thick blue line for each strand. Tn5 has an IS50 element on each side, in an inverted orientation. Transcripts are shown as curly lines with an arrowhead pointing in the direction of transcription. The neo^R gene for Tn5 is composed partly of the leftward IS (ISL) and partly of other sequences (included in the blue box). The transposase for Tn5 is encoded in the rightward IS (ISR).

The TnA family of transposons has been intensively studied for the mechanism of transposition. Members of the TnA family have terminal inverted repeats, but lack terminal IS elements (Figure 9.10). The *tnpA*gene of the TnA transposon encodes a transposase, and the *tnpR*gene encodes a resolvase. TnA also has a selectable marker, *Ap*R, which encodes a beta-lactamase and makes the bacteria resistance to ampicillin.

Transposable elements that move via DNA intermediates are not limited to bacteria, but rather they are found in many species. The P elements and *copia*family of repeats are examples of such transposable elements in *Drosophila*, as are *mariner*elements in mammals and the controlling elements in plants. Indeed, the general structure of **controlling elements in maize** is similar to that of **bacterial transposons**. In particular, they end in inverted repeats and encode a transposase. As illustrated in Figure 9.11, the DNA sequences at the ends of an *Ac* element are very similar to those of a *D*selement. However, internal regions, which normally encode the transposase, have been deleted. This is why *D*selements cannot transpose by themselves, but rather they require the presence of the intact transposon, *Ac*, in the cell to provide the transposase. Since transposase works in *trans*, the *Ac*element can be anywhere in the genome, but it can act on *D*selements at a variety of sites. Note that *Ac* is an **autonomous transposon** because it provides its own transposase and it has the inverted repeats needed to act as the substrate for transposase.

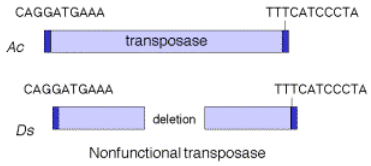


Figure 9.11. Structure of Ac and Ds controlling elements in maize is similar to that of an intact (Ac) or defective (Ds) transposon.

Mechanism of DNA-mediated transposition

Some families of transposable elements that move via a DNA intermediate do so in a replicative manner. In this case, transposition generates a new copy of the transposable element at the target site, while leaving a copy behind at the original site. A cointegrate structure is formed by fusion of the donor and recipient replicons, which is then resolved (Figure 9.12). Other families use a nonreplicative mechanism. In this case, the original copy excises from the original site and move to a new target site, leaving the original site vacant.

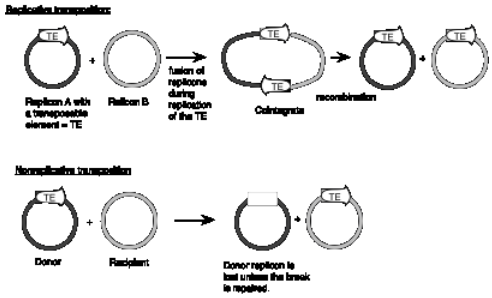


Figure 9.12: Contrasts between replicative and nonreplicative transposition. The transposable element (TE) is shown as an open arrow. The thick line for each replicon represents double stranded DNA; the different shadings represent different sequences.

Studies of bacterial transposons have shown that replicative transposition and some types of nonreplicative transposition proceed through a strand-transfer intermediate (also known as a crossover structure), in which both the donor and recipient replicons are attached to the transposable element (Figure 9.13). For replicative transposition, DNA synthesis through the strand-transfer intermediate produces a transposable element at both the donor and target sites, forming the cointegrate intermediate. This is subsequently resolved to separate the replicons. DNA synthesis does not occur at the crossover structure in nonreplicative transposition, thus leaving a copy only at the new target site. In an alternative pathway for nonreplicative transposition, the transposon is excised by two double strand breaks, and is joined to the recipient at a staggered break (illustrated at the bottom of Figure 9.12).

In more detail, there are two steps in common for replicative and nonreplicative transposition, generating the strand-transfer intermediate (Figure 9.13).

1. The transposase encoded by a transposable element makes four nicks initially. Two nicks are made at the target site, one in each strand, to generate a staggered break with 5' extensions (3' recessed). The other two nicks flank the transposon; one nick is made in one DNA strand at one end of the transposon, and the other nick is made in the other DNA strand at the other end. Since the transposon has inverted repeats at each end, these two nicks that flank the transposon are cleavages in the same sequence. Thus the transposase has a sequence-specific nicking activity. For instance, the transposase from TnA binds to a sequence of about 25 bp located within the 38 bp of inverted terminal repeat (Figure 9.10). It nicks a single strand at each end of the transposon, as well as the target site (Figure 9.13). Note that although the target and transposon are shown apart in the two-dimensional drawing in Figure 9.13, they are juxtaposed during transposition.
2. At each end of the transposon, the 3' end of one strand of the transposon is joined to the 5' extension of one strand at the target site. This ligation is also catalyzed by transposase. ATP stimulates the reaction but it can occur in the absence of ATP if the substrate is supercoiled. Ligation of the ends of the transposon to the target site generates a strand-transfer intermediate, in which the donor and recipient replicons are now joined by the transposon.

After formation of the strand-transfer intermediate, two different pathways can be followed. For replicative transposition, the 3' ends of each strand of the staggered break (originally at the target site) serve as primers for repair synthesis (Figure 9.13). Replication followed by ligation leads to the formation of the cointegrate structure, which can then be resolved into the separate replicons, each with a copy of the transposon. The resolvase encoded by transposon TnA catalyzes the resolution of the cointegrate structure. The site for resolution (res) is located between the divergently transcribed genes for tnpA and tnpR (Figure 9.10). TnA resolvase also negatively regulates expression of both tnpA and tnpR (itself).

For nonreplicative transposition, the strand-transfer intermediate is released by nicking at the ends of the transposon not initially nicked. Repair synthesis is limited to the gap at the flanking direct repeats, and hence only one copy of the transposon is left. This copy is ligated to the new target site, leaving a vacant site in the donor molecule.

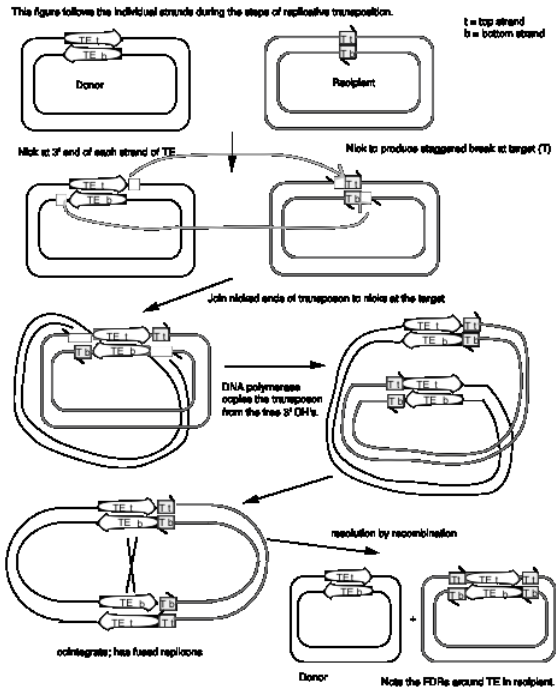


Figure 9.13. Mechanism of transposition via a strand-transfer intermediate.

The enzyme transposase can recognize specific DNA sequences, cleave two duplex DNA molecules in four places, and ligate strands from the donor to the recipient. This enzyme has a remarkable ability to generate and manipulate the ends of DNA. A three-dimensional structure for the Tn5 transposase in complex with the ends of the Tn5 DNA has been solved by Rayment and colleagues. One static view of this protein DNA complex is in Figure 9.14.A. The transposase is a dimer, and each double-stranded DNA molecule (donor and target) is bound by both protein subunits. This orients the transposon ends into the active sites, as shown in the figure. Also, an image with just the DNA (Figure 9.14.B.) shows considerable distortion of the DNA helix at the ends. This recently determined structure is a good starting point to better understand the mechanism for strand cleavage and transfer.

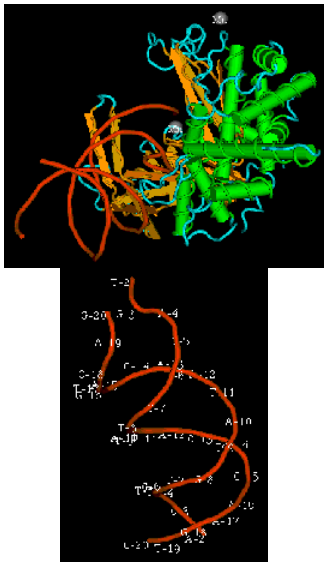


Figure 9.14. Three-dimensional structure of the Tn5 transposase in complex with Tn5 transposon DNA. A. The dimer of the Tn5 transposase is shown bound to a fragment of duplex DNA from the end of the transposon. Alpha helices are green cylinders, beta sheets are yellow-brown, flat arrows and protein loops are blue wires. The DNA is a duplex of two red wires, one for each strand. B. The DNA is shown without the protein and with the nucleotides labeled. The end of the DNA at the top of this panel is oriented into the active site in the middle of the protein in panel A. The structure was determined by Davies DR, Goryshin IY, Reznikoff WS, Rayment I. (2000) “Three-dimensional structure of the Tn5 synaptic complex transposition intermediate.” Science 289:77-85. These images was obtained by downloading the atomic coordinates from the Molecular Modeling Database at NCBI, viewing them with CN3D 3.0 and saving static views as screen shots. The file for observing a virtual three-dimensional image is available at the course website.

Transposable elements that move via RNA intermediates

Transposable DNA sequences that move by an RNA intermediate are called **retrotransposons**. They are very common in eukaryotic organisms, but some examples have also been found in bacteria. Some retrotransposons have long terminal repeats (LTRs) that regulate expression (Figure 9.15). The LTRs were initially discovered in retroviruses. They have now been seen in some but not all retrotransposons. They have a strong promoter and enhancer, as well as signals for forming the 3' end of mRNAs after transcription. The presence of the LTR is distinctive for this family, and members are referred to as LTR-containing retrotransposons. Examples include the yeast *Ty-1*family and retroviral proviruses in vertebrates. Retroviral proviruses encode a reverse transcriptase and an endonuclease, as well as other proteins, some of which are needed for viral assembly and structure.

Others retrotransposons are in the large and diverse class of non-LTR retrotransposons (Figure 9.15). One of the most prevalent examples is the family of long interspersed repetitive elements, or LINEs. It was initially found in mammals but has now been found in a broad range of phyla, including fungi. The first and most common LINE family in mammals is the LINE1 family, also called L1. An older family, but discovered later, is called LINE2. Full-length LINEs are about 7000 bp long, and there are about 10,000 copies in humans. Many other copies are truncated from the 5' ends. Like retroviral proviruses, the full-length L1 encodes a reverse transcriptase and an endonuclease, as well as other proteins. However, the promoter is not an LTR. Other abundant non-LTR retrotransposons, initially discovered in mammals, are short interspersed repetitive elements, or SINEs. These are about 300 bp long. *Alu*repeats, with over a million copies, comprise the predominant class of SINEs in humans. Non-LTR retrotransposons besides LINEs are found in many other species, such as *jockey* repeats in *Drosophila*.

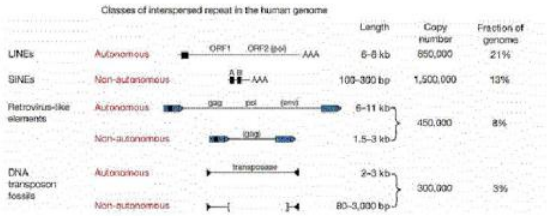


Figure 9.15. Four classes of transposable elements make up the vast majority of human repetitive DNA. From the Nature paper "Initial sequencing and analysis of the human genome," by the International Human Genome Consortium.

Extensive studies in of genomic DNA sequences have allowed the reconstruction of the history of transposable elements in humans and other mammals. The major approach has been to classify the various types of repeats (themselves transposable elements), align the sequences and determine how different the members of a family are from each other. Since the vast majority of the repeats are no longer active in transposition, and have no other obvious function, they will accumulate mutations rapidly, at the neutral rate. Thus the sequence of more recently transposing members are more similar to the source sequence than are the members that transposed earlier. The results of this analysis show that the different families of repeats have propagated in distinct waves through evolution (Figure 9.16). The LINE2 elements were abundant prior to the mammalian divergence, roughly 100 million years ago. Both LINE1 and *Alu* repeats have propagated more recently in humans. It is likely that the LINE1 elements, which encode a nuclease and a reverse transcriptase, provide functions needed for the transposition and expansion of *Alu* repeats. LINE1 elements have expanded in all orders of mammals, but each order has a distinctive SINE, all of which are derived from a gene transcribed by RNA polymerase III. This has led to the idea that LINE1 elements provide functions that other different transcription units use for transposition.

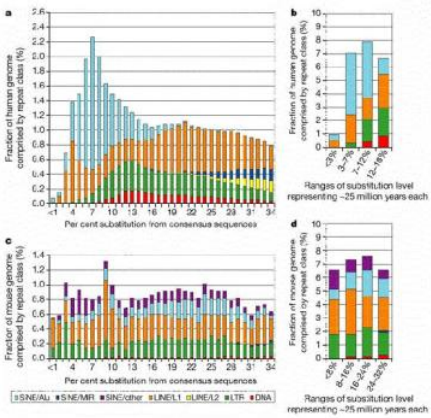


Figure 9.16. Age distribution of repeats in human and mouse. The LINE2 and MIR repeats propagated before the mammalian radiation, about 100 million years ago, but Alu repeats are formed by recent transpositions in primates (light blue portion of the bar graphs in **a**and **b**). The LINE1 and LTR repeats are transposing with about the same frequency as they have historically in the mouse lineage (panels **c** and **d)**, but few repeats are still transposing in human (panels **a** and **b**). From the Nature paper "Initial sequencing and analysis of the human genome," by the International Human Genome Consortium.

Mechanism of retrotransposition

Although the mechanism of retrotransposition is not completely understood, it is clear that at least two enzymatic activities are utilized. One is an integrase, which is an endonuclease that cleaves at the site of integration to generate a staggered break (Figure 9.17). The other is RNA-dependent DNA polymerase, also called reverse transcriptase. These activities are encoded in some autonomous retrotransposons, including both LTR-retrotransposons such as retroviral proviruses and non-LTR-retrotransposons such as LINE1 elements.

The RNA transcript of the transposable element interacts with the site of cleavage at the DNA target site. One strand of DNA at the cleaved integration site serves as the primer for reverse transcriptase. This DNA polymerase then copies the RNA into DNA. That cDNA copy of the retrotransposon must be converted to a double stranded product and inserted at a staggered break at the target site. The enzymes required for joining the reverse transcript (first strand of the new copy) to the other end of the staggered break and for second strand synthesis have not yet been established. Perhaps some cellular DNA repair functions are used.

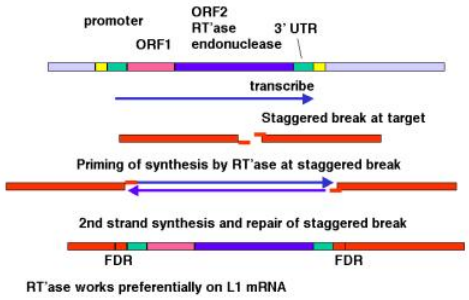


Figure 9.17. Transposition via an RNA-intermediate in retrotransposons. LINE1, or L1 repeats are shown as an example.

The model shown in Figure 9.17 is consistent with any RNA serving as the template for synthesis of the cDNA from the staggered break. However, LINE1 mRNA is clearly used much more often than other RNAs. The basis for the preference of the retrotransposition machinery for LINE1 mRNA is still being studied. Perhaps the endonuclease and reverse transcriptase stay associated with the mRNA that encodes them after translation has been completed, so that they act in cis with respect to the LINE1 mRNA. Other repeats that have expanded recently, such as Alu repeats in humans, may share sequence determinants with LINE1 mRNA for this cis preference.

Clear evidence that retrotransposons can move via an RNA intermediate came from studies of the yeast Ty-1 elements by Gerald Fink and his colleagues. They placed a particular Ty-1 element, called TyH3 under control of a GAL promoter, so that its transcription (and transposition) could be induced by adding galactose to the media. They also marked TyH3 with an intron. After inducing transcription of TyH3, additional copies were found at new locations in the yeast strain. When these were examined structurally, it was discovered that the intron had been removed. If the RNA transcript is the intermediate in moving the Ty-1

element, it is subject to splicing and the intron can be removed. Hence, these results fit the prediction of an RNA-mediated transposition. They demonstrate that during transposition, the flow of Ty-1 sequence information is from DNA to RNA to DNA.

Exercise 9.4

If yeast Ty-1 moved by the mechanism illustrated for DNA-mediated replicative transposition in Figure 9.13, what would be predicted in the experiment just outlined? Also, would you expect an increase in transposition when transcription is induced?

Additional Consequences of Transposition

Not only can transposable elements interrupt genes or disrupt their regulation, but they can cause additional rearrangements in the genome. Homologous recombination can occur between any two nearly identical sequences. Thus when transposition makes a new copy of a transposable element, the two copies are now potential substrates for recombination. The outcome of recombination depends on the orientation of the two transposable elements relative to each other. Recombination between two transposable elements in the same orientation on the same chromosome leads to a deletion, whereas it results in an inversion if they are in opposite orientations (Figure 9.18).

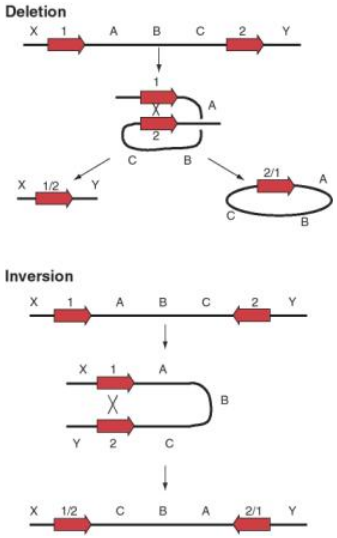


Figure 9.18. Possible outcomes of recombination between two transposable elements.

The preference of the retrotransposition machinery for LINE1 mRNA does not appear to be absolute. Many processed genes have been found in eukaryotic genomes; these are genes that have no introns. In many cases, a homologous gene with introns is seen in the genome, so it appears that these processed genes have lost their introns. It is likely that these were formed when processed mRNA derived from the homologous gene with introns was copied into cDNA and reinserted into the genome. Many, but not all, of these processed genes are pseudogenes, i.e. they have been mutated such that they no longer encode proteins. Other examples of active processed genes have inserted next to promoters and encode functional proteins.

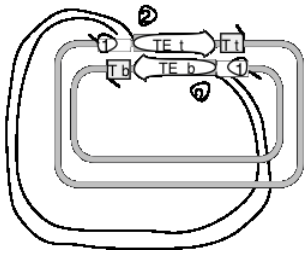
Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

9.E: Transposition of DNA (Exercises)

Question 9.5. Suppose you are studying a gene that is contained within a 5 kb *Eco*RI fragment for the wild type allele. When analyzing mutations in that gene, you found one that converted the 5 kb fragment to an 8 kb *Eco*RI fragment. Further analysis showed that the additional 3 kb of DNA was flanked by direct repeats of 6 bp, that the terminal 30 bp of the additional DNA was identical at each end but in an inverted orientation. Recombinant plasmids carrying the 8 kb *Eco*RI fragment conferred resistance to the antibiotic kanamycin in the host bacteria, whereas neither the parental cloning vector nor a recombinant plasmid carrying the 5 kb *Eco*RI fragment did. What do you conclude is the basis for this mutation? What other enzyme activities might you expect to be encoded in the additional DNA?

Use the following diagram to answer the **next two** questions. Transposase encoded by a transposable element (TE) has nicked on each side of the TE in the donor (black) replicon and made a staggered break in the recipient (gray) replicon, and the ends of the TE have been joined to the target (T) site in the recipient replicon. The strands of the replicons have been designated top (t) or bottom (b). The open triangles with 1 or 2 in them just refer to locations in the figure; they are not part of the structure.



Question 9.6. The action of DNA polymerase plus dNTPs, primed at positions 1, followed by ligase (with ATP or NAD) leads to what product or result? (In this scenario, nothing occurs at positions 2).

Question 9.7. The action of an endonuclease at the positions labeled 2 followed by DNA polymerase and dNTPs to fill in the gaps (from positions 1 to the next 5' ends of DNA fragments), and finally DNA ligase (with ATP or NAD) leads to what product or result?

Question 9.8. Refer to the model for a crossover intermediate in replicative transposition in Fig. 9.13. If the transposon moved to a second site on the same DNA molecule by replicative transposition (not to a different molecule as shown in the Figure), what are the consequences for the DNA between the donor and recipient sites?

Question 9.9. The technique of transposon tagging uses the integration of transposons to mutate a large numbers of genes while leaving a "tag" in the mutated gene to allow subsequent isolation of the gene using molecular probes (such as hybridization probes for the transposon). What is a good candidate for transposon tagging in mammalian cells?

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Additional consequences of transposition

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Welcome to the Biology Library. This Living Library is a principal hub of the LibreTexts project, which is a multi-institutional collaborative venture to develop the next generation of open-access texts to improve postsecondary education at all levels of higher learning. The LibreTexts approach is highly collaborative where an Open Access textbook environment is under constant revision by students, faculty, and outside experts to supplant conventional paper-based books.

Mechanism of DNA-mediated transposition

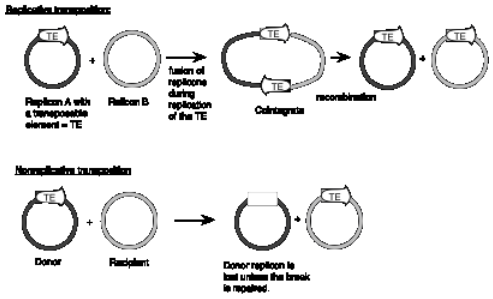


Figure 9.13).

1. The **transposase** encoded by a transposable element makes four nicks initially. Two nicks are made at the target site, one in each strand, to generate a staggered break with 5' extensions (3' recessed). The other two nicks flank the transposon; one nick is made in one DNA strand at one end of the transposon, and the other nick is made in the other DNA strand at the other end. Since the transposon has inverted repeats at each end, these two nicks that flank the transposon are cleavages in the same sequence. Thus the transposase has a sequence-specific nicking activity. For instance, the transposase from TnA binds to a sequence of about 25 bp located within the 38 bp of inverted terminal repeat (Figure 9.10). It nicks a single strand at each end of the transposon, as well as the target site (Figure 9.13). Note that although the target and transposon are shown apart in the two-dimensional drawing in Figure 9.13, they are juxtaposed during transposition.
2. At each end of the transposon, the 3' end of one strand of the transposon is joined to the 5' extension of one strand at the target site. This ligation is also catalyzed by transposase. ATP stimulates the reaction but it can occur in the absence of ATP if the substrate is supercoiled. Ligation of the ends of the transposon to the target site generates a strand-transfer intermediate, in which the donor and recipient replicons are now joined by the transposon.

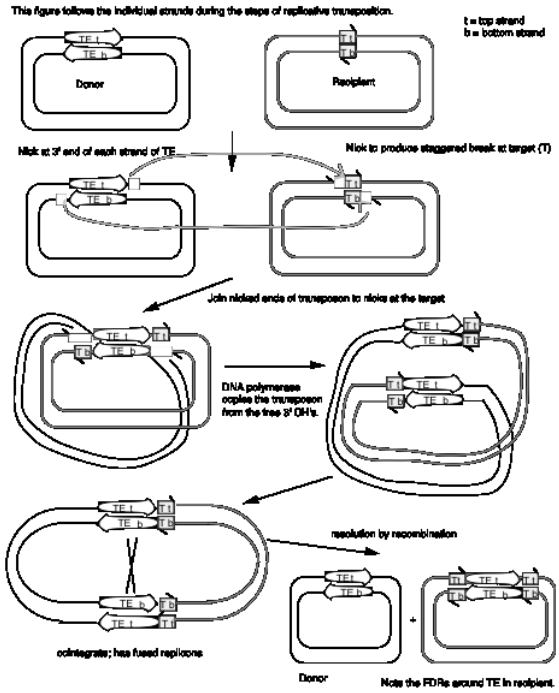


Figure 9.14.Three-dimensional structure of the Tn5 transposase in complex with Tn5 transposon DNA. A. The dimer of the Tn5 transposase is shown bound to a fragment of duplex DNA from the end of the transposon. Alpha helices are green cylinders, beta sheets are yellow-brown, flat arrows and protein loops are blue wires. The DNA is a duplex of two red wires, one for each strand. B. The DNA is shown without the protein and with the nucleotides labeled. The end of the DNA at the top of this panel is oriented into the active site in the middle of the protein in panel A. The structure was determined by Davies DR, Goryshin IY, Reznikoff WS, Rayment I. (2000) “Three-dimensional structure of the Tn5 synaptic complex transposition intermediate.” Science 289:77-85. These images was obtained by downloading the atomic coordinates from the Molecular Modeling Database at NCBI, viewing them with CN3D 3.0 and saving static views as screen shots.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Mechanism of Retrotransposition

Although the mechanism of retrotransposition is not completely understood, it is clear that at least two enzymatic activities are utilized. One is an **integrase**, which is an endonuclease that cleaves at the site of integration to generate a **staggered break** (Figure 9.17). The other is RNA-dependent DNA polymerase, also called **reverse transcriptase**. These activities are encoded in some autonomous retrotransposons, including both LTR-retrotransposons such as retroviral proviruses and non-LTR-retrotransposons such as LINE1 elements.

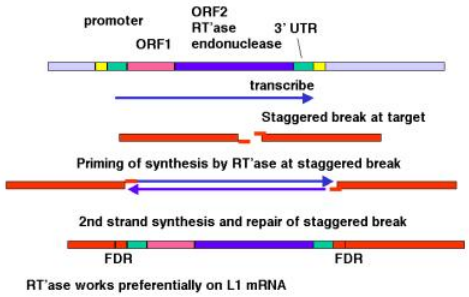


Figure 9.17. Transposition via an RNA-intermediate in retrotransposons.LINE1, or L1 repeats are shown as an example.

The **RNA transcript** of the transposable element interacts with the site of cleavage at the DNA target site. One strand of DNA at the cleaved integration site serves as the primer for reverse transcriptase. This DNA polymerase then copies the RNA into DNA. That cDNA copy of the retrotransposon must be converted to a double stranded product and inserted at a staggered break at the target site. The enzymes required for joining the reverse transcript (first strand of the new copy) to the other end of the staggered break and for second strand synthesis have not yet been established. Perhaps some cellular DNA repair functions are used.

The model shown in Figure 9.17 is consistent with any RNA serving as the template for synthesis of the cDNA from the staggered break. However, LINE1 mRNA is clearly used much more often than other RNAs. The basis for the preference of the retrotransposition machinery for LINE1 mRNA is still being studied. Perhaps the endonuclease and reverse transcriptase stay associated with the mRNA that encodes them after translation has been completed, so that they act in *cis*with respect to the LINE1 mRNA. Other repeats that have expanded recently, such as *Alu*repeats in humans, may share sequence determinants with LINE1 mRNA for this *cis*preference.

Clear evidence that retrotransposons can move via an RNA intermediate came from studies of the yeast *Ty-1*elements by Gerald Fink and his colleagues. They placed a particular *Ty-1*element, called *TyH3*under control of a *GAL*promoter, so that its transcription (and transposition) could be induced by adding galactose to the media. They also marked *TyH3*with an intron. After inducing transcription of *TyH3*, additional copies were found at new locations in the yeast strain. When these were examined structurally, it was discovered that the intron had been removed. If the RNA transcript is the intermediate in moving the *Ty-1*element, it is subject to splicing and the intron can be removed. Hence, these results fit the prediction of an RNA-mediated transposition. They demonstrate that during transposition, the flow of *Ty-1*sequence information is from DNA to RNA to DNA.

Exercise

If yeast *Ty-1* moved by the mechanism illustrated for DNA-mediated replicative transposition in Figure 9.13, what would be predicted in the experiment just outlined? Also, would you expect an increase in transposition when transcription is induced?

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

Unstable Alleles

Insertion of a controlling element can generate an unstable allele of a locus

The insertion of a controlling element can generate an **unstable allele of a locus**, designated ***mutable***. This instability can be seen both in somatic and in germline tissues. The instability can result from reversion of a mutation, due to the excision and transposition of the controlling element. After excision and re-integration, the transposable element can alter the expression of a gene at the new location. This new phenotype indicated that the element was mobile.

An example of the effects of integration and excision of a transposable element can be seen at the *bronze* locus in maize (Figure 9.8). The aleurone is the surface layer of endosperm in a kernel of maize. The wild type has a deep bluish-purple color. This is determined by the *bronze*locus. The *Bz* allele is dominant and confers the bluish-purple color to the aleurone. The *bz*allele is recessive, and gives a bronze color to the aleurone when homozygous. In *Bz*kernels, anthocyanin is produced. *Bz*encodes UDPglucose:flavanoid 3-*O*-glucosyltransferase (UFGT), an enzyme needed for anthocyanin production. The loss-of-function *bz* alleles have no UFGT activity, and the bluish-purple anthocyanins are not produced. Some alleles of *bronze* show an unstable, or mutable, phenotype. In the *bz-m*alleles, clones of cells regain the bluish-purple color characteristic of *Bz* cells. This produces patches of bluish-purple color in the aleurone of kernels (Figure 9.8).

This mutation in the *bz-m*allele is the insertion of the *Ds*(dissociation) transposable element. *Ds*disrupts the function of the UFGT gene to give a bronze color to the seed kernel. In the presence of the *Ac*(activator) element, the *Ds* can excise from the locus, restoring a functional UFGT gene (and a bluish-purple color). This occurs in some but not all cells in the developing seed and is clonally inherited, resulting in the patches of blue on a bronze background for each kernel.

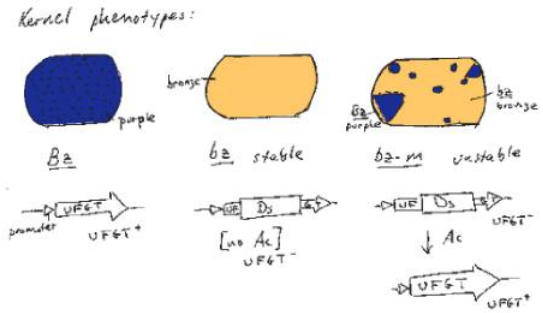


Figure 9.8. Frequent excision of a *Ds* allele generates an unstable, or mutable, phenotype at the bronze locus.

Current methods for observing transposition and transposable elements

Movement of DNA segments can be observed by a variety of modern techniques. In organisms with a short generation time, such as bacteria and yeast, one can simply monitor many generations for the number and positions of a family of repeated DNA elements by blot-hybridization analysis of genomic DNA. Using a *Ty-1*DNA fragment as a probe, about 20 hybridizing bands could be seen at the start of an experiment, meaning that about 20 copies were present in the yeast genome. The size of the restriction fragment containing each element was distinctive, as determined by restriction endonuclease cleavage sites that flanked the different locations of each element. After growing for many generations, some new bands were observed, showing that new *Ty-1*elements had been generated and moved to new locations. These observations led to this family of repeats being christened *Ty-1*, for transposable element, yeast, number 1.

Evidence for transposition in many organisms comes from analysis of new mutations. Transposable elements appear to be the major source of new mutation in *Drosophila*, and they have been shown to cause mutations in bacteria, fungi, plants and animals. One example from humans is a new mutation causing hemophilia. A patient from a family with no prior history was diagnosed with hemophilia, resulting from an absence of factor VIII. By molecular cloning techniques, Kazazian and his colleagues showed that the mutant factor VIII gene had a copy of a LINE1, or L1, repeat inserted. In contrast to most L1 repeats in the human genome, whose sequences have diverged from a predicted source gene, the sequence of this L1 was very close to that predicted for an active L1. Tests showed that the patient's parents did not carry this mutation in their factor VIII genes. Screening a genomic library for L1s that were almost identical to the mutagenic L1 revealed a full-length, active L1 that was the source, on a different chromosome. The appearance of a new L1 in the factor VIII gene, making an allele that was not present in the parents, is a strong

argument for transposition. The further studies identifying a source gene and showing that the source gene is active in transposition make the evidence unequivocal.

Now that it is recognized that most repetitive elements in many species result from transposition events, it is easy to find transposable elements or their progeny. A comprehensive database of repetitive elements in many species is maintained as RepBase (J. Jurka) and the program RepeatMasker (Green and Smit) will widely used to find matches to these repeats. Virtually all these repeats are made by transposition.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

SECTION OVERVIEW

Unit III: The Pathway of Gene Expression

10: Transcription: RNA polymerases

10.E: Transcription: RNA polymerases (Exercises)

11: Transcription: Promoters, terminators and mRNA

11.E: Transcription: Promoters, terminators and mRNA (Exercises)

12: RNA processing

12.0: Overview of RNA Processing
12.1: Cutting and Trimming RNA
12.2: Modifications at the 5' and 3' ends of mRNA
12.3: Multiple Mechanisms are Used for Splicing Different Types of Introns
12.4: Self-splicing by group I introns (pre-rRNA of Tetrahymena)
12.5: RNAs Can Function as Enzymes
12.6: Splicing of introns in pre-mRNAs
12.7: Splicing of group II introns
12.8: Alternative Splicing
12.9: RNA editing
12.E: RNA Processing (Exercises)

13: Genetic code

13.E: Genetic Code (Exercises)

14: Translation (Protein synthesis)

14.E: Translation - Protein synthesis (Exercises)

Thumbnail: A ribosome produces a protein using mRNA as template. (Public Domain; LadyofHats).

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

10: Transcription: RNA polymerases

Recall the Central Dogma of molecular biology: DNA is transcribed into RNA, which is translated into protein. We will cover the material in that order, since that is the direction that information flows.

Introduction

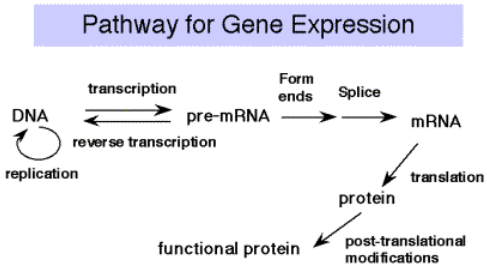


Figure 3.1.2: Sequential addition of ribonucleotides to growing RNA

The liberated pyrophosphate is cleaved in the cell to 2 Pi, an energetically favorable reaction that drives the reaction in the direction of synthesis. In the presence of excess PPi, the reverse reaction of pyrophosphorolysis can occur. Synthesis always proceeds in a 5' to 3' direction (with respect to the growing RNA chain). The template is read in a 3' to 5' direction.

B. E. coli RNA polymerase structure

1. **This one RNA polymerase synthesizes all classes of RNA**

mRNA, rRNA, tRNA

2. **It is composed of four subunits**.

a. **Core and holoenzyme**

a2bb's a2bb' + s

Holoenzyme = a2bb's = core + s = can **initiate**transcription accurately as the proper site, as determined by the promoter

Core = a2bb' = can **elongate** a growing RNA chain

A **promoter** can be defined in two ways.

1. The sequence of DNA required for accurate, specific intiation of transcription
2. The sequence of DNA to which RNA polymerase binds to accurately initiate transcription.

b. Subunits

Subunit	Size	Gene	Function
b'	160 kDa	*rpoC*	b' + b form the catalytic center.
b	155 kDa	*rpoB*	b' + b form the catalytic center.
a	40 kDa	*rpoA*	enzyme assembly; also binds UP sequence in the promoter
s	70 kDa (general)	*rpoD*	confers specificity for promoter; binds to -10 and -35 sites in the promoter

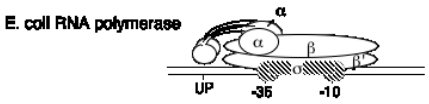
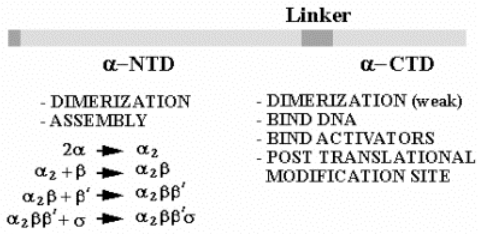


Figure 3.1.7. Role of the **asubunit in assembly and other functions**



C. *E. coli* RNA polymerase mechanism

Mode of action of **sfactors**

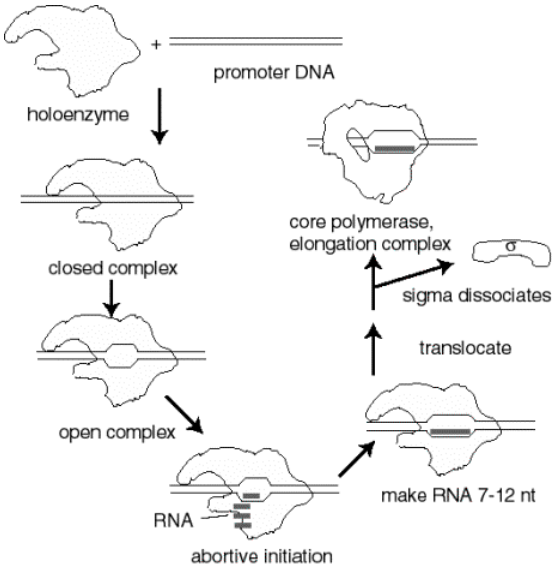
The presence of the s factor causes the RNA polymerase holoenzyme to be selective in choosing the site of initiation. This is accomplished primarily through effects on the dissociation rate of RNA polymerase from DNA.

- a. Core has strong affinity for general DNA sequences. The $t_{1/2}$ for dissociation of the complex of core-DNA is about 60 min. This is useful during the elongation phase, but not during initiation.
- b. Holoenzyme has a reduced affinity for general DNA; it is decreased about 10^4 fold. The $t_{1/2}$ for dissociation of holoenzyme from general DNA is reduced to about 1 sec.
- c. Holoenzyme has a greatly increased affinity for promoter sequences.The $t_{1/2}$ for dissociation of holoenzyme from promoter sequences is of the order of hours.

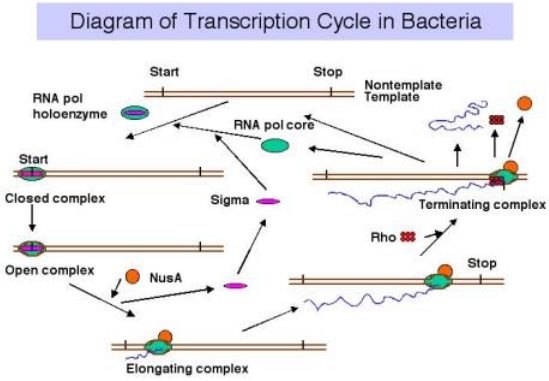
Events at initiation of transcription

- a. RNA polymerase holoenzyme binds to the promoter to form a **closed complex**; at this stage there is no unwinding of DNA.
- b. The polymerase-promoter complex undergoes the closed to open transition, which is a melting or unwinding of about 12 bp.
- c. The initiating nucleotides can bind to the enzyme, as directed by their complementary nucleotides in the DNA template strand, and the enzyme will catalyze formation of a phosphodiester bond between them. This polymerase-DNA-RNA complex is referred to as the ternary complex.
- d. During **abortive initiation**, the polymerase catalyzes synthesis of short transcripts about 6 or so nucleotides long and then releases them.
- e. This phase ends when the nascent RNA of ~6 nucleotides binds to a second RNA binding site on the enzyme; this second site is distinct from the catalytic center. This binding is associated with "resetting" the catalytic center so that the enzyme will now catalyze the synthesis of oligonucleotides 7-12 long.
- f. The enzyme now translocates to an new position on the template. During this process **sigma leaves the complex**. A conformational change in the enzyme associated with sigma leaving the complex lets the "thumb" wrap around the DNA template, locking in processivity. Thus the core enzyme catalyzes RNA synthesis during elongation, which continues until "signals" are encountered which indicate termination.

Figure 3.1.8. Events at initiation



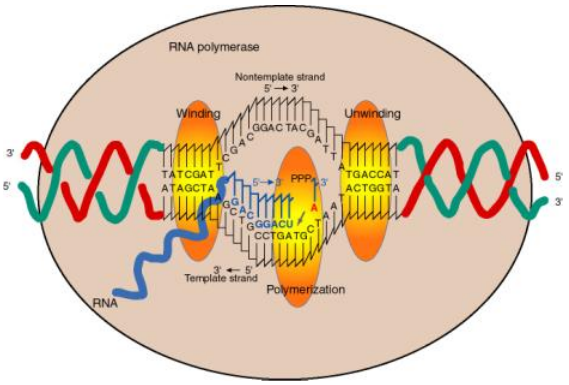
3. Transcription cycle



4. Sites on RNA Polymerase core

a. The enzyme covers about 60 bp of DNA, with a transcription bubble of about 17 bp unwound.

Figure 3.1.10



d The incoming nucleotide (NTP) that will be added to the growing RNA chain binds adjacent to the 3' end of the growing RNA chain, as directed by the template, at the active site for polymerization.

e. The incoming nucleotide is linked to the growing RNA chain by nucleophilic attack of the 3' OH on the a phosphoryl of the NTP, with liberation of pyrophosphate.

f. The reaction progresses (the enzyme moves) about 50 nts per sec. This is much slower than the rate of replication (about 1000 nts per sec).

g. If the template is topologically constrained, the DNA ahead of the RNA polymerase becomes overwound (positive superhelical turns) and the DNA behind the RNA polymerase becomes underwound (negative superhelical turns).

The effect of the unwinding of the DNA template by RNA polymerase is to decrease T by 1 for every 10 bp unwound. Thus DT = -1, and since DL = 0, then DW = +1 for every 10 bp unwound. This effect of the increase in W will be exerted in the DNA ahead of the polymerase.

The effect of rewinding the DNA template by RNA polymerase is just the opposite, of course. T will increase by 1 for every 10 bp rewound. Thus DT= +1, and since DL = 0, then DW = -1 for every 10 bp rewound. This effect of the decrease in W will be exerted in the DNA behind the polymerase, since that is where the rewinding is occurring.

5. Inhibitors: useful reagents and clues to function

a. Rifamycins, e.g. rifampicin: bind the b subunit to block initiation. The drug prevents addition of the 3rd or 4th nucleotide, hence the initiation process cannot be completed.

How do we know the site of rifampicin action is the b subunit? Mutations that confer resistance to rifampicin map to the *rpoB*gene.

b. Streptolydigins: bind to the b subunit to inhibit chain elongation.

These effects of rifamycins and streptolydigins, and the fact that they act on the b subunit, argue that the b subunit is required for nucleotide addition to the growing chain.

c. Heparin, a polyanion, binds to the b' subunit to prevent binding to DNA in vitro

D. Eukaryotic RNA polymerases

1. Eukaryotes have 3 different RNA polymerases in their nuclei.

1. a. Each nuclear RNA polymerase is a large protein with about 8 to 14 subunits. MW is approximately 500,000 for each.
2. b. Each polymerase has a different function:

RNA polymerase	localization	synthesizes	effect of a-amanitin
RNA polymerase I	nucleolus	pre-rRNA	none
RNA polymerase II	nucleoplasm	pre-mRNA some snRNAs	inhibited by low concentrations (0.03 mg/ml)
RNA polymerase III	nucleoplasm	pre-tRNA, other small RNAs some snRNAs	inhibited by high concentrations (100 mg/ml)

2. **Subunit structures**

a. The genes and encoded proteins for the subunits of the yeast RNA polymerases have been isolated and the sequences determined, and some functional analysis has been done.

b. Some of the subunits are homologous to bacterial RNA polymerases: The largest two subunits are homologs of b and b'. The roughly 40 kDa subunit is the homolog of a.

c. Some subunits are common to all three RNA polymerases.

d. Example of yeast RNA polymerase II:

Approximate size (kDa)	subunits per polymerase	role / comment	
220	1	related to b'	catalytic?
130	1	related to b	catalytic?
40	2	related to a	assembly?
35	< 1		

Approximate size (kDa)	subunits per polymerase	role / comment	
30	2	common to all 3	
27	1	common to all 3	
24	< 1		
20	1	common to all 3	
14	2		
10	1		

e. The largest subunit has a **carboxy-terminal domain (CTD)**with an unusual structure: tandem repeats of the sequence Tyr-Ser-Pro-Thr-Ser-Pro-Thr. The yeast enzyme has 26 tandem repeats and the mammalian enzyme has about 50. These can be phophorylated on Ser and Thr to give a highly charged CTD.

- RNA Pol IIa is not phosphorylated in the CTD.
- RNA Pol IIo is phosphorylated in the CTD.

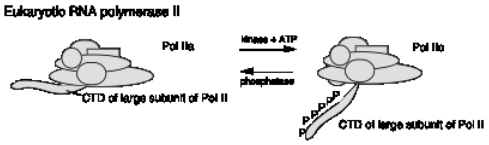
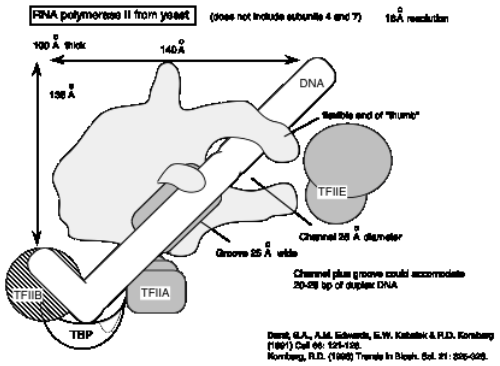


Figure 3.1.13 Diagram of yeast RNA polymerase II with some general transcription factors



4. **RNA polymerases in chloroplasts (plastids) and mitochondria**

- a. The RNA polymerase found in plastids is encoded on the plastid chromosome. In some species the mitochondrial RNA polymerase is encoded by the mitochondrial DNA.
- b. These organellar RNA polymerases are much more related to the bacterial RNA polymerases than to the nuclear RNA polymerases. This is a strong argument in favor of the origins of these organelles being bacterial, supporting the endo-symbiotnt model for acquisition of these organelles in eukaryotes.
- c. These RNA polymerases catalyze specific transcription of organellar genes.

E. General transcription factors for eukaryotic RNA polymerase II

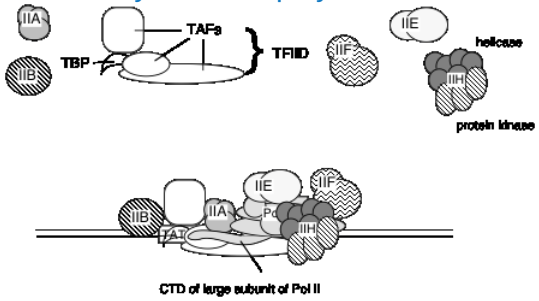


Figure 3.1.15. Ribbon diagram of TBP bound to DNA.

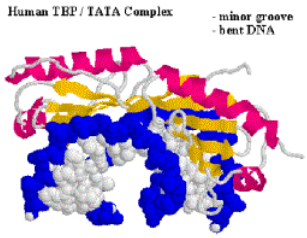


Image from crystal structures, provided by Dr. T. Nixon.

It is not known if the same set of TAFs are in the TFIID for all promoters transcribed by RNA polymerase II, or if some are used only for certain types of promoters. TFIID is the only sequence-specific general transcription factor so far characterized, and it binds in the minor groove of the DNA. It is also used at TATA-less promoters, so the role of the sequence -specific binding is still under investigation.

3. **Summary of general transcription factors for RNA polymerase II.**

Factors for RNA polymerase II (human cells)

Factor	No. of subunits	Molecular mass (kDa)	Functions	Functions to Recruit:
TFIID: TBP	1	38	Recognize core promoter (TATA)	TFIIB
TFIID: TAFs	12	15-250	Recognize core promoter (non-TATA); Positive and negative regulation	RNA Pol II?
TFIIA	2	12, 19, 35	Stabilize TBP-DNA binding; Anti-repression	
TFIIB	1	35	Select start site for RNA Pol II	RNA PolII-TFIIF
RNA Pol II	12	10-220	Catalyze RNA synthesis	TFIIE
TFIIF	2	30, 74	Target RNA PolII to promoter; destabilize non-specific interactions between PolII and DNA	
TFIIE	2	34, 57	Modulate TFIIH helicase, ATPase and kinase activities; Directly enhance promoter melting?	TFIIH
TFIIH	9	35-89	Helicase to melt promoter; CTD kinase; promoter clearance?	

Roeder, R.G. (1996) TIBS 21: 327-335.

4. **TFIIH is a multisubunit transcription factor also involved in DNA repair.**

Subunits of the human factor

Gene	Molec. mass of protein (kDa)	Function/ Structure	Proposed Role
XPB	89	helicase, tracks 3' to 5'	Unwind duplex for transcription/ Repair
XPD	80	helicase, tracks 5' to 3'	Unwind duplex, Repair
P62	62	unknown	
P52	52	unknown	
P44	44	Zn-finger	Binds DNA
P34	34	Zn-finger	
MAT1	32	CDK assembly factor	
Cyclin H	38	Cyclin partner for CDK7/MO15	
CDK7/MO15	32	Protein kinase	Kinase for CTD

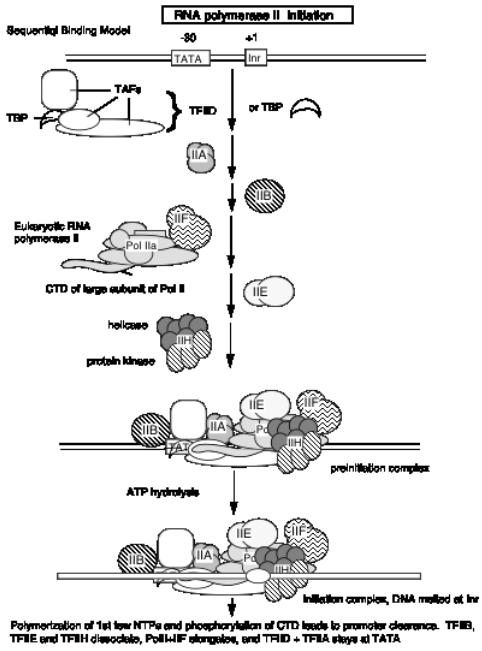


Figure 3.1.17).

Table 3.1.6. RNA polymerase II holoenzyme and mediator

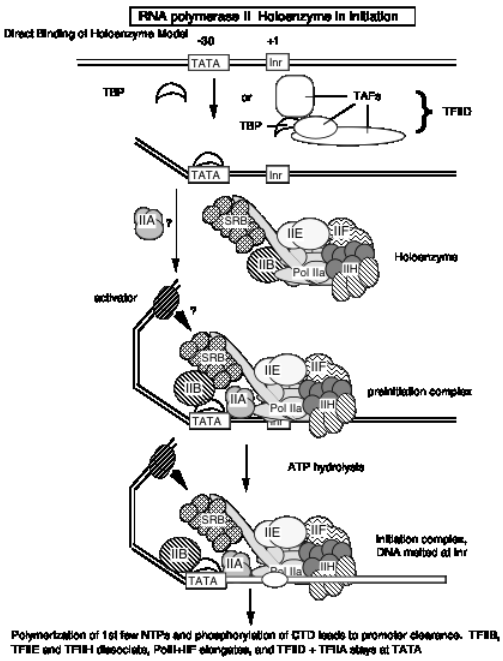
- **Holoenzyme**
  - RNA polymerase II + (TFIIB, E, F, H)+ (Srb2, 4, 5, 6) + (Rgr1, Gal11, others)
  - Correct initiation in presence of TBP (TFIID)
  - Responds to transcriptional activators
- **Mediator**
  - Complex needed for a response to transcriptional activators by purified RNA Pol II plus GTFs
  - Yeast Mediator has 20 subunits, including Srb2, 4, 5, 6; Srb7, Rgr1, Gal11, Med 1, 2, 6, 7, Pgd1, Nut 1, 2, and others
- **RNA Pol II + Mediator** (+ some GTFs?) = **Holoenzyme**

These studies show that RNA polymerase II can exist in several different states or complexes. One is in a very large holocomplex containing the mediator. In this state, it will accurately initiate transcription when directed by TFIID, and respond to activators (Table 3.1.6). The mediator subcomplex appears to be able to dissociate and reassociate with RNA polymerase II and GTFs. Indeed, this reassociation could be the step that was assayed in the identification of mediator. Without mediator, RNA polymerase II plus GTFs can initiate transcription at the correct place (as directed by TFIID), but they do not respond to activators. In the absence of GTFs, RNA polymerase II is capable of transcribing DNA templates, but it will not begin transcription at the correct site. Hence it is competent for elongation but not initiation.

Table 3.1.7. Expanding the functions of RNA polymerase II

Polymerase	Trans-cribe	Start at Promoter	Respond to Activator
RNA Pol II	Yes	No	No
RNA Pol II + GTFs	Yes	Yes	No
RNA Pol II holoenzyme + GTFs	Yes	Yes	Yes

Fig.3.1.17



If the holoenzyme is the primary enzyme involved in transcription initiation in eukaryotic cells, then the progressive assembly pathway observed *in vitro*(see section d above) may be of little relevance *in vivo*. Perhaps the holoenzyme will bind to promoters simply marked by binding of TBP (or TFIID) to the TATA box, in contrast to the progressive assembly model that has a more extensive, ordered assembly mechanism. In both models, TBP or TFIID binding is the initial step in assembly of the preinitiation complex. However, at this point one cannot rule out the possibility that the holoenzyme is used at some promoters, and progressive assembly occurs at others.

7. **Targets for the activator proteins**

The targets for transcriptional activator proteins may be some component of the initiation complex. One line of investigation is pointing to the TAFs in TFIID as well as TFIIB as targets for the activators. Thus the activators may facilitate the ordered assembly of the intiation complex by recruiting GTFs. However, the holoenzyme contains the "mediator" or SRB complex that can mediate response to activators. Thus the activators may serve to recruit the holoenzyme to the promoter. Further studies are required to establish whether one or the other is correct, or if these are separate paths to activation.

F. General transcription factors for eukaryotic RNA polymerases I and III

1. **General transcription factors for RNA polymerase I**

   a. Core promoter covers the start site of transcription, plus an upstream control element located about 70 bp further 5'.

   b. The factor UBF1 binds to a G+C rich sequence in both the upstream control element and in the core promoter.

   c. A multisubunit complex called SL1 binds to the UBF1-DNA complex, again at both the upstream and core elements.

   d. One of the subuntis of SL1 is TBP - the TATA-binding protein from TFIID!

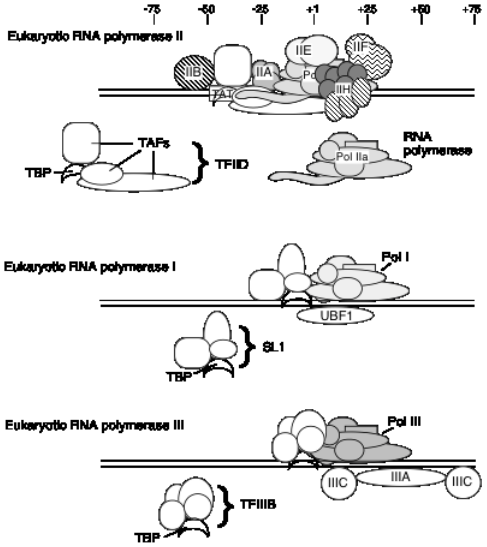
   e. RNA polymerase I then binds to this complex of DNA+UBF1+SL1 to initiate transcription at the correct nucleotide and the elongate to make pre-rRNA.

2. **General transcription factors for RNA Pol III**

- a. Internal control sequences are characteristic of genes transcribed by RNA Pol III (see below).
- b. TFIIIA: binds to the internal control region of genes that encode 5S RNA (type 1 internal promoter)
- c. TFIIIC: binds to internal control regions of genes for 5S RNA (alongside TFIIIA) and for tRNAs (type 2 internal promoters)
- d. TFIIIB: The binding of TFIIIC directs TFIIIB to bind to sequences (-40 to +11) that overlap the start site for transcription. One subunit of TFIIIB is TBP, even though no TATA box is required for transcription. TFIIIA and TFIIIC can now be removed

without affecting the ability of RNA polymerase III to initiate transcription. Thus TFIIIA and TFIIIC are assembly factors, and TFIIIB is the initiation factor.

Figure 3.1.18.



e. RNA polymerase III binds to the complex of TFIIIB+DNA to accurately and efficiently initiated transcription.

3 **Transcription factor used by all 3 RNA Pol'ases: TBP**

TBP seems to play a common role in directing RNA polymerase (I, II and III) to initiate at the correct place. The multisubunit factors that contain TBP (TFIID, SL1 and TFIIIB) may serve as positioning factors for their respective polymerases.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

10.E: Transcription: RNA polymerases (Exercises)

10.1 What is the role of the sigma factor in transcription, and how does it accomplish this?

10.2 Specific binding of *E. coli*RNA polymerase to a promoter. Which of the following statements are correct?

1. Completely envelopes the DNA duplex (both sides).
2. requires sigma factor to be part of the holoenzyme.
3. is enhanced by methylation of purine bases.
4. results in a temperature-dependent unwinding of about 10 base pairs.

10.3 (POB) RNA polymerase. How long would it take for the *E. coli*RNA polymerase to synthesize the primary transcript for *E. coli*rRNAs (6500 bases), given that the rate of RNA chain growth is 50 nucleotides per second?

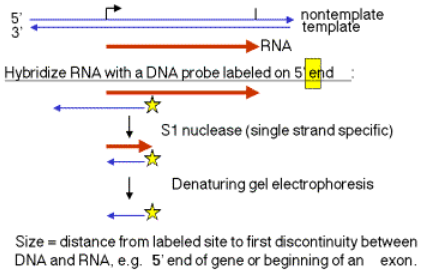
10.4 What is the maximum rate of initiation at a promoter, assuming that the diameter of RNA polymerase is about 204 Angstroms and the rate of RNA chain growth is 50 nucleotides per second?

10.5 Although three different eukaryotic RNA polymerases are used to transcribe nuclear genes, the enzymes and their promoters show several features in common. Are the following statements about common features of the polymerases and their mechanisms of initiation true or false?

1. All three purified polymerases need additional transcription factors for accurate initiation at promoter sequences.
2. All three polymerases catalyze the addition of a nucleotide "cap" to the 5' end of the RNA.
3. For all three polymerases, the TATA-binding protein is a subunit of a transcription factor required for initiation (not necessarily the same factor for each polymerase).
4. All three polymerases are composed of multiple subunits.

10.6 What is common and what is distinctive to the reactions catalyzed by DNA polymerase, RNA polymerase, reverse transcriptase, and telomerase?

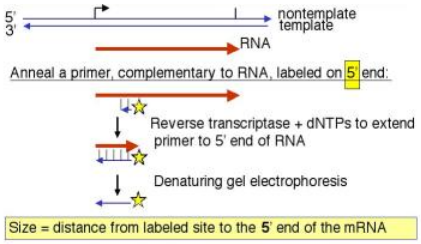
11: Transcription: Promoters, terminators and mRNA



1. "S1 protection assay"

This assay measures the distance between an end label (at a specific known site on DNA) and the end of a duplex between RNA and the labeled DNA. A fragment of DNA (complementary to the RNA) that extends beyond the 5' end of the RNA is labeled at a restriction site within the RNA-complementary region. The labeled DNA is hybridized to RNA and then digested with the single-strand specific nuclease S1. The resulting fragment of protected DNA is run on a denaturing gel to determine its size. Note that this fragment runs from the labeled site to the nearest interruption between the DNA and the RNA. This could be the beginning of the RNA, or it could be an intron, or it could be an S1 sensitive site.

Figure 3.2.3. Primer extension assay, another way to map the 5’ ends of genes



3. How do you label DNA at the ends?

- 5' end label: T4 polynucleotide kinase and [g 32P] ATP. The reaction is most efficient if the 5' phosphate is removed (by alkaline phosphatase) prior to the kinase treatment.
- 3' end label: Klenow DNA polymerase plus [a 32P] dNTP. The labeled dNTP is chosen to be complementary to the first position past the primer. A restriction fragment with a 5' overhang is ideal for this "fill-in" labeling.
- Digestion with a second restriction endonuclease will frequently work to remove the label at the "other" end. One can also use electrophoretic gels that separate strands.

4. A PCR-based technique to determine the 5’ ends of mRNAs and genes

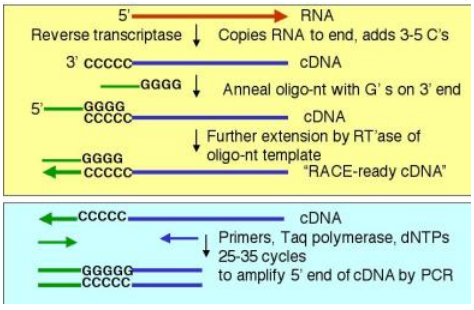


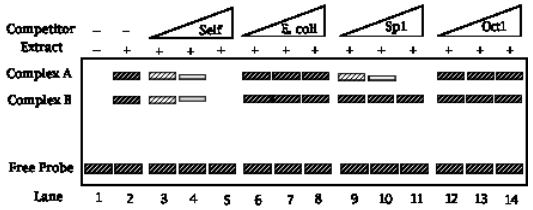
Figure 3.2.3.b, the fact that reverse transcriptase tends to add a few C residues to the 3’ end of the cDNA is used to design an artificial template that will anneal to those extra C nucleotides. Then reverse transcriptase copies the second template, thereby adding the artificial primer binding site. This artificial primer binding site is needed because the sequence of the 5’ end of the mRNA is not known in this experiment; indeed, that is what the experimenter is trying to determine. Once the artificial primer binding site has been added to the cDNA, then the modified cDNA serves as the template for PCR. The PCR product is sequenced and compared to an appropriate genomic DNA sequence. The first exon or exons of the genes will match the sequence of the PCR product, starting right after the first primer.

B. General methods for identifying the site for sequence-specific binding proteins

1. Does a protein bind to a particular region?

a. Electrophoretic mobility shift assay (EMSA), or gel retardation assay: This assay will test for the ability of a particular sequence to form a complex with a protein. Many protein-DNA complexes are sufficiently stable that they will remain together during electrophoresis through a (nondenaturing) polyacrylamide gel. A selected restriction fragment or synthetic duplex oligonucleotide is labeled (to make a probe) and mixed with a protein (or crude mixture of proteins). If the DNA fragment binds to the protein, the complex will migrate much slower in the gel than does the free probe; it moves with roughly the mobility of the bound protein. The presence of a slowly moving signal is indicative of a complex between the DNA probe and some protein(s). By incubating the probe and proteins in the presence of increasing amounts of competitor DNA fragments, one can test for specificity and even glean some information about the identity of the binding protein.

Figure 3.2.4. Diagram of results from an electrophoretic mobility shift assay



In this example, two proteins recognize sequences in the labeled probe, forming complexes A and B (lane 2). The proteins in complexes A and B recognize **specific**DNA sequences in the probe. This is shown by the competition assays in lanes 3-8. An excess of unlabeled oligonucleotide with the same sequence as the labeled probe (“self”) prevents formation of the complexes with labeled probe, whereas “nonspecific DNA” in the form of *E. coli* DNA does not compete effectively (compare lanes 6-7 with lanes 3-5).

This experiment also provides some information about the **identity** of the protein forming complex A. It recognizes an Sp1-binding site, as shown by the ability of an oligonucloetide with an Sp1-binding to compete for complex A, but not complex B (lanes 9-11). Hence the protein could be Sp1 or a relative of it.. The proteins forming complexes A and B do not recognize an Oct1-binding site (lanes 12-14).

b. **Nitrocellulose binding**: Free duplex DNA will not stick to a nitrocellulose membrane, but a protein-DNA complex will bind.

2. To what sequence in the probe DNA is the protein binding?

The presence of a protein will either protect a segment of DNA from attack by a nuclease or other degradative reagent, or in some cases will enhance cleavage (e.g. to an adjacent sequence that is distorted from normal B-form). An end-labeled DNA fragment in complex with protein is treated with a nuclease (or other cleaving reagent), and the protected fragments are resolved on a denaturing polyacrylamide gel, and their sizes measured.

- a**. Exonuclease protection assay:** The protein will block the progress of an exonuclease, so the protected fragment extends from the labeled site to the edge of the protein furtherest from the labeled site. One can use a combination of a 3' to 5' exonuclease (ExoIII) and a 5' to 3' exonuclease (l exonuclease) to map both edges.
- b. **DNase footprint analysis:** DNase I will cut at many (but not all) phosphodiester bonds in the free DNA. The protein-DNA complex is treated lightly with DNase I, so that on average each DNA molecule is cleaved once. The presence of a bound protein will block access of the DNase, and the bound region will be visible as a region of the gel that has no bands, i.e. that was not cleaved by the reagent.

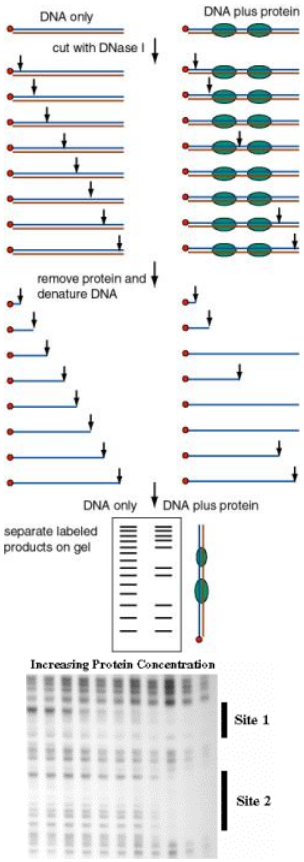


Figure 3.1.6. Methylation interference assay.

4. DNA sequence-affinity chromatography to purify DNA binding proteins

The specific binding sites (often 6 to 8 bp) can serve as an affinity ligand for chromatography. Multimers of the binding site are made by ligating together duplex oligonucleotides that contain the specific site. After a few crude initial steps (e.g. isolating all DNA-binding proteins on DNA-sepharose) the extract is applied to the affinity column. Most of the proteins do not bind, and subsequently the specifically bound proteins are eluted.

C. Promoters and the Initiation of Transcription: General Properties

1. A promoter is the DNA sequence required for correct initiation of transcription
2. **Phenotype of promoter mutants**

a. *cis*-acting: A *cis*-acting regulatory element functions as a segment of DNA to affect the expression of genes on the same chromosome that it is located on. *Cis*-acting elements do not encode a diffusible product. The promoter is a *cis*-acting regulatory element.

Compare the phenotypes of mutations in the gene encoding b-galactosidase (*lacZ*) versus mutations in its promoter (*p*).

Consider a heterozygote that is *p*+ *lacZ*- /*p*+ *lacZ*+ .

The phenotype is Lac+. *lacZ*+ complements *lacZ*- in *trans*. In this case, *lacZ*+ is dominant to *lacZ*-.

Consider a heterozygote that is *p*+ *lacZ*- /*p*- *lacZ*+ .

The phenotype is Lac-. *p*+ does not complement *p*- in *trans*.

p- operates in *cis* to prevent expression of *lacZ*+ on this chromosome. The mutant promoter is dominant over the wild-type when the mutant promoter is in *cis* to the wt *lacZ*.

Consider a heterozygote that is*p*+ *lacZ*+ /*p*- *lacZ*- .

The phenotype is Lac+. *lacZ*+ now complements *lacZ*- in *trans* because it is driven by a functional promoter in *cis*, *p*+

b. Dominance in *cis*: the promoter "allele" that is in *cis*to the wild-type structural gene (*lacZ*) is dominant over the other promoter allele.

c. Promoter mutations affect the amount of product from the gene but do not affect the structure of the gene product.

D. Bacterial promoters

1. Bacterial promoters **occur just 5' to and overlap the start site for transcription**(usually)
2. Bacterial promoters are the **binding site for *E.*** coli**RNA polymerase holoenzyme.** The promoter covers about 70 bp from about -50 to about +20.
3. **Consensus sequences in the *E. coli*promoter**

a. -35 and -10 sequences

-35 16-19 bp -10+1

--------TTGACA----------TATAAT---C**A**T

Recognition by Allows binary complex to convert

RNA polymerase from closed to open

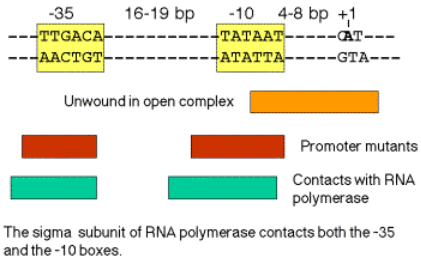
holoenzyme

b. The sequences are conserved in all *E. coli* genes transcribed by holoenzyme with s70

4. Promoter mutants

- a. Tend to fall into or close to one of these hexanucleotides
- b. Affect the level of gene expression, not the structure of the gene product
- c. Down promoter mutations: decrease the level of transcription. Tend to make the promoter sequence less like the consensus.
- d. Up promoter mutations: increase the level of transcription. Tend to make the promoter sequence more like the consensus.
- e. Down promoter mutations in the -35 sequence: decrease the rate of formation of the closed complex, indicating this is the sequence needed for intial recognition by the polymerase holoenzyme.
- f. Down promoter mutations in the -10 sequence: decrease the rate of conversion from the closed to the open complex, again supporting the proposed role for this A+T rich hexanucleotide.
- g. The critical contact points between RNA polymerase and the promoter tend to be in or immediately upstream from the consensus -35 and -10 boxes. (See Figure 3.2.7). Thus the biochemical and genetic data all support the importance of these conserved sequences.

Figure 3.2.7. Correlation of conserved sequences, location of promoter mutants, and regions of contact with polymerase at bacterial promoters



5. Alternate s factors can control the expression of sets of genes

- a. Alternative s factors make complexes with the core polymerase to direct the new holoenzyme to a particular set of promoters that differ in sequence from the general *E. coli* promoter sequence. Thus the polymerase can be directed to trancribe a new set of genes. This is one way to control gene expression.
- b. Examples include s factors for heat-shock response (s32), transcription of genes involved in chemotaxis and flagellar formation (s28), and nitrogen starvation (s54). The s factors are named by their size in kDa.
- c. Three of the *E.* colis factors have regions of sequence similarity (s70,s32, and s28) whereas s54 is a distinctly different molecule that works rather differently.

Factor	Gene	Use	-35	Separation	-10
s70	*rpoD*	General	TTGACA	16-19 bp	TATAAT
s32	*rpoH*	Heat shock	CCCTTGAA	13-15 bp	CCCGATNT
s28	*fliA*	Flagella	CTAAA	15 bp	GCCGATAA
s54	*rpoN*	Nitrogen starvation	CTGGNA	6 bp	TTGCA

E. Promoters for eukaryotic RNA polymerases

Promoters contain binding sites for nuclear proteins, but which of these binding sites have a function in gene expression? This requires a genetic approach for an answer.

1. Use of "surrogate genetics" to define the promoter

a. *In vitro* mutagenesis (deletions or point mutations)

1. Mutations of the binding sites for activator proteins lead to a decrease in the level of transcription of the gene. [Loss of function].
2. Addition of a DNA fragment containing these binding sites will activate (some) heterologous promoters. [Gain of function].
3. Sequences of the binding sites are frequently well conserved in promoters for homologous genes from related species.
4. A potential regulatory region is initially examined by constructing progressive deletions from the 5' end (with respect to the direction of transcription) and also from the 3' end. Subsequently one can make clusters of point mutations (e.g. by linker scanning mutagenesis) or individual point mutations.

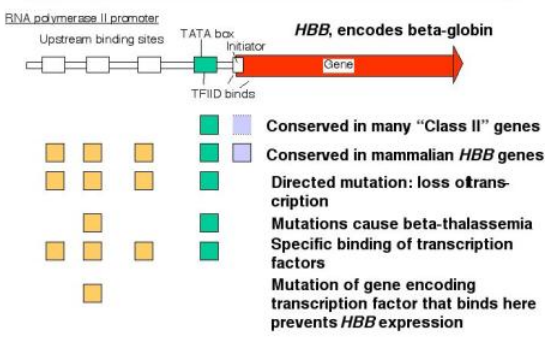


Figure 3.2.8. Evidence for an RNA polymerase II promoter.

b. Test in an expression assay

(1)The mutagenized promoter is linked to a reporter gene so that RNA or protein from that gene can be measured quantitatively

- (*a*) Gene itself - measure RNA production by S1 protection, primer extension, or other assay that is specific for a particular RNA
- (*b*) Heterologous reporter gene: encodes an enzyme whose activity is easy to measure quantitatively. Note that these measures of expression require both transcription and translation, in contrast to measurement of RNA directly. E.g., the genes encoding:

1. b-galactosidase: colorimetric assay, monitor the cleavage of o-nitrophenyl-b-galactoside
2. chloramphenicol (Cm) acetyl transferase (CAT): measure the acetylation of Cm, ususally use [14C] Cm; this is the enzyme that confers resistance to Cm in bacteria
3. luciferase: monitor the emmission of photons resulting from the ATP-dependent oxidation of luciferin; this is the enzyme that catalyzes light production in firefly tails

(2) The promoter-reporter DNA constructs are introduced into an assay system that will allow the reporter to be expressed.

(*a*) Whole cells

microinjection into Xenopus oocytes

transfection of cell lines: introduce the DNA via electroporation or by getting the cells to take up a precipitate of DNA and Ca phosphate by pinocytosis

(*b*) Whole animals = transgenic animals

Introduce the DNA into the germ line of an animal, in mammals by microinjecting into a fertilized egg and placing that into a pseudopregnant female. This technology allows one to examine the effects of the mutation throughout the development of the animal.

(*c*) Cell-free systems

Extracts of nuclei, or purified systems (i.e. with all the necessary components purified)

2. Promoter for RNA Pol II

a. The minimal promoter is needed for basal activity and accurate initiation.

1. Needed for assembly of the initiation complex at the correct site
2. DNA sequences

(a) TATA box

1. Initially identified as a well conserved sequence motif about 25 bp 5' to the cap site (The cap site is the usual start site for transcription)
2. The transcription factor TFIID binds to the TATA box
3. Mutations at the TATA box generates heterogeneous 5' ends of the mRNAs - indicative of a loss of start site specificity

(b) Initiator

1. Sequences at the start site for transcription have consensus YANWYY (Y = C or t, W = T or A)
2. Mode of action is still under investigation. Recent data indicate that TFIID also binds to the initiator; binds to one of the TAFs (see below).
3. TATA plus initiator is the simplest minimal promoter.

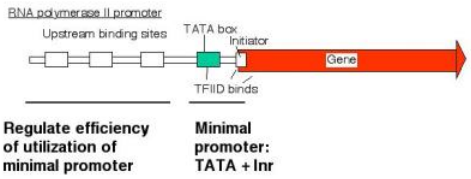


Figure 3.2.9. Two general parts of promoters for RNA polymerase II.

b. The amount of expression is regulated via upstream elements.

1. Proteins bind to specific sequences (usually) 5' to the TATA box to regulate the efficiency of utilization of the promoter.
2. These are frequently activators, but proteins that exert negative control are also being characterized.
3. Examples of activator proteins

Sp1: binds GGGGCGGGG = GC box

Oct*n*: binds ATTTGCAT = octamer motif

Oct1 is a general factor (ubiquitous)

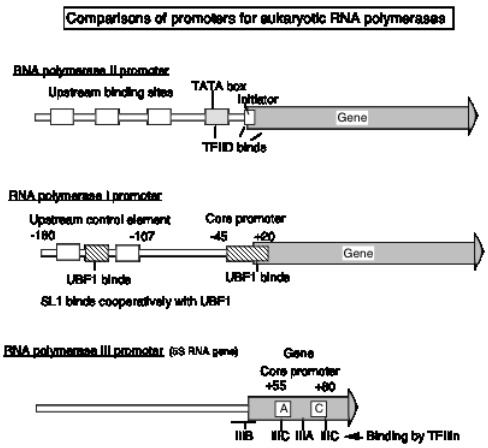
Oct2 is specific for lymphoid cells

CP1, CTF = NF1, C/EBP bind to CCAAT = CCAAT box (pronounced "cat" box)

These are different families of proteins, CP1 and CTF are found in many cell types, C/EBP is found in liver and adipose tissue.

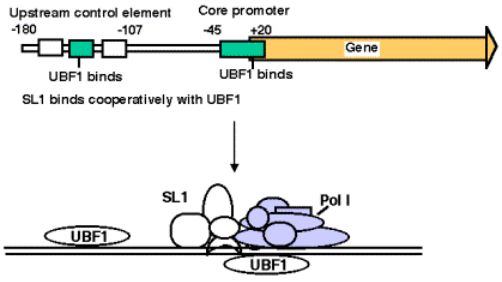
(4) These upstream control elements may be inducible (e.g. by hormones), may be cell-type specific, or they may be present and active in virtually all cell types (i.e. ubiquitous and constitutive).

Figure 3.2.10.



3. Promoter for RNA Pol I

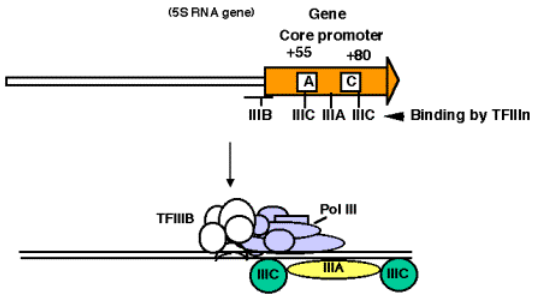
- a. The core promoter covers the start site of transcription, from about -40 to about +30. The promoter also contains an upstream control element located about 70 bp further 5', extending from -170 to -110.
- b. The factor UBF1 binds to a G+C rich sequence in both the upstream control element and in the core promoter. A multisubunit complex called SL1 binds to the UBF1-DNA complex, again at both the upstream and core elements. One of the subuntis of SL1 is TBP.
- c. RNA polymerase I then binds to this complex of DNA+UBF1+SL1 to initiate transcription at the correct nucleotide and the elongate to make pre-rRNA.



4. Promoter for RNA Pol III

a. This promoter has internal control sequences. Deletion of 5' flanking DNA still permits efficient transcription of (most) genes transcribed by RNA PolIII. Even the intial part of the gene is expendable, as is the 3' end. Sequences internal to the gene (e.g. +55 to +80 in 5S rRNA genes) are required for efficient initiation, in contrast to the familiar situation in bacteria, where most of the promoter sequences are 5' to the gene.

Figure 3.2.11. Binding of proteins for promoter for RNA polymearase III



F. Enhancers

1. Enhancers are **DNA sequences that cause an increase in the level of expression of a gene** with an intact promoter. They may act to **increase**the efficiency of utilization of a promoter, or they may increase the probability that a promoter is in a transcriptionally competent chromatin conformation. This will be explored further in Part Four.
2. They are operationally defined by their ability to act in either orientation and at a variety of positions and distances from a gene, i.e. **act independently of orientation and position**. This contrasts with promoters, that act (usually) in only one orientation and (usually) are at or close to the 5' end of the gene.
3. They consist of binding sites for specific activator proteins. Always have multiple binding sites, often for several different activator proteins.
4. Particular sets of genes can be regulated by their need for defined sets of activator proteins at their enhancers.

G. Elongation of transcription

1. RNA polymerase must be released from the initiation complex to transcribe the rest of the gene. Elongation must be highly processive, i.e. once the polymerase begins elongation, it must transcribe that template all the way to the end of the gene.
2. The factors required for initiation are not needed (and may inhibit) elongation, and they dissociate.

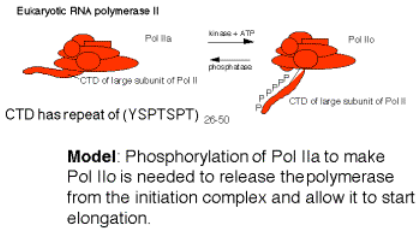
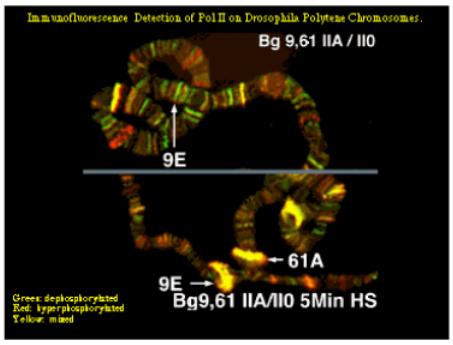


Figure 3.2.14. Supportive evidence: Immunofluoresence shows Pol IIa is on heat shock genes when quiescent (stalled polymerases), but Pol IIo is present once the genes are actively transcribed (elongating polymerases).



3. There is some indication that **factors that increase the processivity of the transcription complex**bind to the elongating polymerase. Examples include the following.

- NusA in bacteria
- GreA and GreB in bacteria
- TFIIS in eukaryotes, possibly many others.

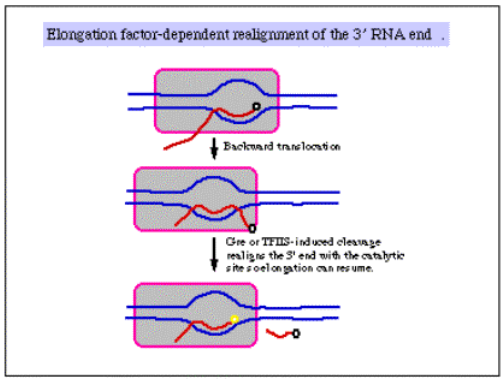
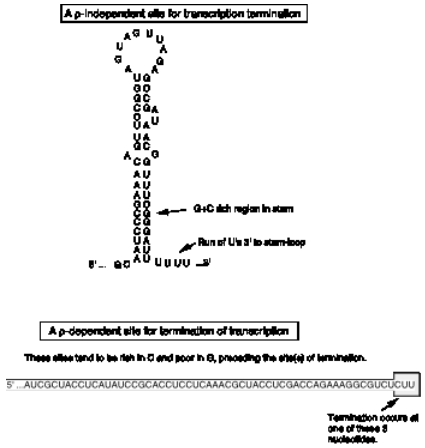


Figure 3.2.17.



a. r-independent sites [Note: r = rho]

1. Identified in vitro
2. G+C rich hairpin followed by about 6 U's
3. Hairpin is thought to be a site at which RNA polymerase pauses, and the weak rU-dA base pairs in the RNA-DNA heteroduplex allow melting of the duplex and termination.
4. Some of the best examples of r-independent terminators are integral parts of the mechanism of regulation. Examples include the attenuators in the *trp*operon and other amino acid biosynthetic operons. The r-independent terminators may be a specialized adaptation for regulation.

b. r-dependent sites

1. C-rich, G-poor stretch
2. Requires the action of the protein r both in vitro and in vivo
3. The r-dependent terminators are used at the 3' ends of many eubacterial genes.

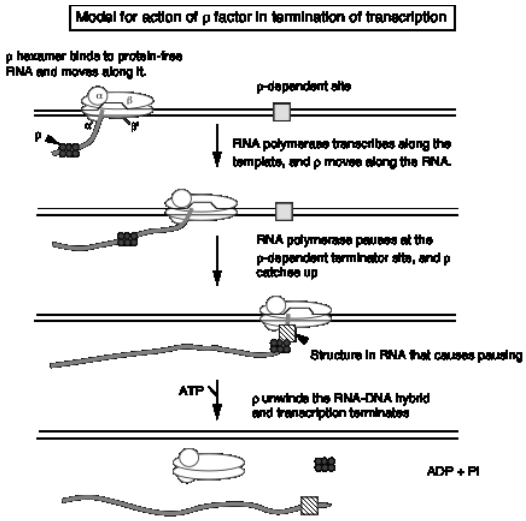
2. r factor

- a. Hexamer, each subunit 46 kDa
- b. RNA-dependent ATPase
- c. Gene for r is essential for *E. coli*

3. Model for action of r factor

- a. r binds to protein-free RNA and moves along it
- b. When it reaches a paused polymerase, it causes the polymerase to dissociate and unwinds the RNA-DNA duplex, thereby terminating transcription. This last step utilizes the energy of ATP hydrolysis. The protein r serves as the ATPase.

Figure 3.2.18.



I. Termination of transcription in eukaryotes

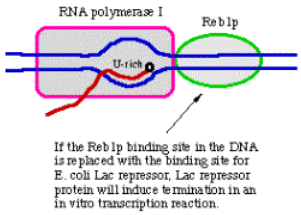
1. Termination by RNA Pol II

- a. No clear evidence for a discrete terminator for RNA polymerase II
- b. 3' end of mRNA is generated by cleavage and polyadenylation
- c. Signal for cleavage and polyadenylation:

  (1) AAUAAA, about 20 nt before the 3' end of the mRNA

  (2) Other sequences 3' to cleavage site

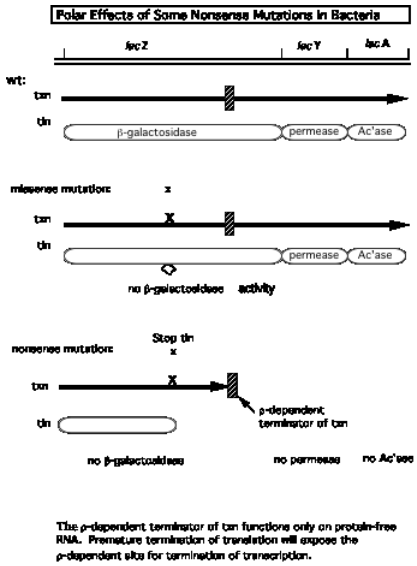
- d. Cleavage enzyme not well characterized at this point; the U4 snRNP may play a role in cleavage. A polyA polymerase has been identified.
- e. Polyadenylation is required for termination by RNA Pol II; possibly also pausing by the RNA polymerase



J. mRNA structure in bacteria

1. **Bacterial mRNA is often polycistronic**.

Figure 3.2.21.



2. Model for r action can explain why stopping translation can also lead to a cessation of transcription.

- a. Suppose a r-dependent terminator of transcription is present in the first gene of an operon. Normally it does not cause transcription to stop because it is covered by ribosomes translating the mRNA, and the subsequent genes in the operon are transcribed. Recall that r requires protein-free RNA to bind to and to move along.
- b. A nonsense mutation before the cryptic r-dependent terminator would cause the ribosomes to dissociate, now exposing the cryptic terminator in a protein-free stretch of RNA. The hexamer r can bind and move along the RNA, and when it encounters an RNA polymerase stalled, or paused, at the terminator, it will cause the RNA polymerase to dissociate and the RNA to be released, hence preventing transcription of the subsequent genes in the operon.

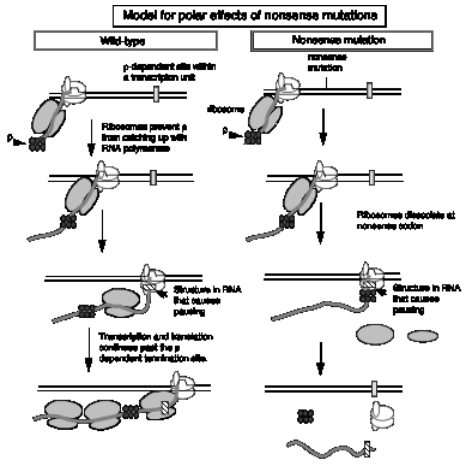
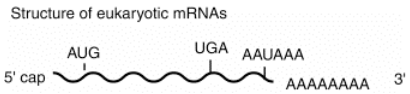


Figure 3.2.23.



This general structure is true for almost all eukaryotic mRNAs. The cap structure is almost ubiquitous. A few examples of mRNAs without poly A at the 3' end have been found. Some of the most abundant mRNAs without poly A encode the histones. However, most mRNAs do have the 3' poly A tail.

The poly A tail at the 3' end can be used to purify mRNAs from other RNAs. Total RNA from a cell (which is about 90% rRNA and less than 10% mRNA) can be passed over an oligo(dT)-cellulose column. The poly A-containing mRNAs will bind, whereas other RNAs will elute.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

11.E: Transcription: Promoters, terminators and mRNA (Exercises)

Q11.1 Determining the sequences that encode the ends of mRNAs

A gene that determines eye color in salamanders, called *almond*, is contained within a 2000 bp *Kpn*I fragment. After cloning the *Kpn*I fragment in a plasmid, it was discovered that it has a *Bgl*II site 500 bp from the left *Kpn*I site and an *Eco*RI site 300 bp from the right *Kpn*I site, as shown in the map below.

bp 0 500 1000 1500 2000

| | | | |

|_|_|_|

*Kpn*I *Bgl*II *Eco*RI *Kpn*I

In order to determine the positions that correspond to the 5' and 3' ends of the *almond*RNA, the *Eco*RI and *Bgl*II sites were labeled at the 5' or 3' end. The *Kpn*I to *Bgl*II fragments (500 and 1500 bp) and the *Kpn*I to *Eco*RI fragments (1700 and 300 bp) were isolated, hybridized to *almond*RNA and treated with the single-strand specific nuclease S1. The sizes of the probe fragments protected from digestion in the RNA-DNA duplex are shown below (in nucleotides); a 0 means that the probe was not protected by RNA.

5' end-labeled probe **3' end**-labeled probe

protected protected

probe fragment probe fragment

*Kpn*I-*Bgl*II* 5000 *Kpn*I-*Bgl*II* 500 100

**Bgl*II-*Kpn*I 1500 1300 **Bgl*II-*Kpn*I 1500 0

*Kpn*I-*Eco*RI* 17000 *Kpn*I-*Eco*RI* 1700 1300

**Eco*RI-*Kpn*I 300 100 **Eco*RI-*Kpn*I 300 0

The asterisk denotes the end that was labeled.

a. What is the direction of transcription of the *almond*gene, relative to the map above?
b. What position on the map corresponds to the 5' end of the mRNA?
c. What position on the map corresponds to the 3' end of the mRNA?

Q11.2: Determining the sequences that encode the ends of mRNAs

The gene for histone H2A from armadillo can be isolated as a 1400 bp *Pst*I fragment. The map is shown below; the armadillo *Pst*I fragment is shown by the double dashed line, and the vector DNA is denoted by the single dashed lines. Sizes are in base pairs. The H2A gene clone was cleaved with *Hind*III, treated with alkaline phosphatase, and incubated with polynucleotide kinase and [32P] ATP in an appropriate buffer to introduce a radiolabel at the 5’ ends of the DNA fragments. The DNA was then extracted with phenol to remove the kinase, and then cut again with *Pst*I. The labeled 600 bp and 800 bp *Pst*I-*Hind*III fragments were separated by gel electrophoresis and isolated. The isolated fragments were denatured, hybridized to histone mRNA, and treated with nuclease S1. The S1-resistant labeled DNA fragments were identified by gel electrophoresis followed by radioautography. A 200 nucleotide protected fragment was observed when the 600 bp fragment was used in the S1 protection assay, but no protected fragment was observed when the 800 bp fragment was used.

|--------------600----------------|-------------800----------------|

*Pst*I *Hind*III *Pst*I

---------|====================|====================|-----------

| | | |

0 500 1000 1400

a. What is the direction of transcription of the histone H2A gene (relative to the restriction map above)?
b. With reference to the numbers below the restriction map, what is the position of the 5' end of the histone H2A mRNA?

c. What is the position of the 3' end of the mRNA?

Q11.3

A 400 bp DNA fragment containing the start site for transcription of the *almond* gene was investigated to find transcriptional control signals. The start site (+1 in the coordinate system) is 100 bp from the right end. The 400 bp fragment is sufficient to drive transcription of a reporter gene (for luciferase) in an appropriate cell line. Two series of 5' and 3' deletions were made in the 400 bp fragment and tested for their ability to drive transcription of the luciferase reporter gene. Each fragment in the 5' deletion series has a different 5' end, but all are fused to the luciferase gene at +100 (see diagram below). Each fragment in the 3' deletion series has a common 5' end at -300, but each is fused to the luciferase gene at the designated 3' position. The amount of luciferase (a measure of the level of transcription) for each construct is shown in the first two pairs of columns in the table. The intact reporter construct, with *almond*DNA (the horizontal line) fused to the luciferase gene, is diagrammed immediately below.

-300-250 -200-150 -100-50 +1+50 +100

| | | | | | | | |

_____|Luciferase>

|

start>

To further investigate the function of different regions, sub-fragments of the *almond*DNA fragment were added to a construct in which the reporter gene was driven by a different promoter, as diagrammed below. The effects of the almond DNA fragments on this heterologous promoter are shown in the third pair of columns in the table.

Test fragment from *almond*DNA heterologous promoter Luciferase gene>

_____|----------|**********>

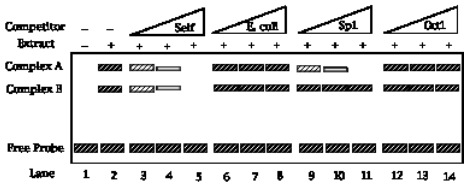
5' deletion endpoints	Amount of expression	3' deletion endpoints	Amount of expression		Test fragment of almond	Amount of expression
-300	100	-200	0		-300 to -250	100
-250	100	-150	0		-250 to -200	500
-200	50	-100	0		-200 to -150	100
-150	50	-50	0		-150 to -100	300
-100	25	+1	100		-100 to -50	300
-50	10	+50	100		-50 to -1	100
-1	0	+100	100		none	100

a. What do you conclude is the role of the -250 to -200 fragment?
b. What do you conclude is the role of the -200 to -150 fragment?
c. What do you conclude is the role of the -150 to -100 fragment?
d. What is the role of the -50 to -1 fragment of the *almond*gene?

Q11.4

An electrophoretic mobility shift assay was used to test for the ability of a short restriction fragment to bind to proteins from the nuclei of kidney cells. The restriction fragment was labeled at one end, mixed with an extract containing the nuclear proteins, and run on a non-denaturing polyacrylamide gel. Lane 1 (below) shows the free probe and lane 2 shows the the probe plus extract; electrophoresis is from the top to the bottom. Complexes between proteins and the labeled DNA probe move more slowly on the gel than does the free probe. Further tests of specificity are shown in the competition lanes, in which the labeled probe was mixed with an increasing excess of other DNA before mixing with the nuclear proteins to test for binding. Competitor DNAs included the unlabeled probe (self competition, lanes 3-5; the triangle above the lanes indicates that an increasing amount of competitor is used in successive lanes), a completely different DNA (sheared E. coli DNA) as a nonspecific competitor (lanes 6-8), and two different

duplex oligonucleotides, one containing the binding site for Sp1 (lanes 9-11) and the other containing the binding site for Oct1 (lanes 12-14). Thinner, less densely filled boxes denote bands of less intensity than the darker, thicker bands.



a. How many protein-DNA complexes are formed between the labeled DNA probe and the nuclear extract?
b. What do lanes 3-8 tell you about the protein-DNA complexes?
c. What do lanes 9-14 tell you about the protein-DNA complexes?

Q11.5

In order to determine the contact points between a regulatory protein and its binding site on the DNA, a small fragment of duplex DNA was end-labeled (at the 5' terminus of the left end as written below) and treated with dimethyl sulfate so that each molecule on average has one G nucleotide methylated. The regulatory protein was mixed with the preparation of partially methylated DNA, and protein-bound DNA was separated from unbound DNA. After cleaving the DNA at the methylated sites, the resultant fragments were resolved on a "sequencing gel". An autoradiogram of the results showed bands corresponding to all the G's in the labeled fragment for the unbound DNA, but the protein-bound DNA did not have bands corresponding to the G's at positions 14 and 16 below. When the left end of the fragment was labeled at the 3' terminus, no band corresponding to the G (bottom strand) at position 18 (same numbering system as for top strand) was seen in the preparation of protein-bound DNA.

.0 15 20 25 30

| | |

GATCCGCATGGATGAGTCACGTAACGTGTA

GCGTACCTACTCAGTGCATTGCACAT

What is the binding site for the regulatory protein?

Q11.6

Are the following statements about r and polar effects of some mutations in operons in *E. coli*true or false?

a. Nonsense mutations (terminating translation) in the first gene of an operon can have no effect on the transcription of subsequent gene in the operon.
b. Mutations in the gene for r (*rho*gene) can suppress polarity.
c. The hexameric protein r binds to protein-free RNA and moves along the RNA; when it encounters a stalled RNA polymerase it promoters termination of transcription.
d. The protein r is an RNA-dependent ATPase.

CHAPTER OVERVIEW

12: RNA processing

RNA processing refers to any covalent modification to the RNA that occurs after transcription. This includes specific cleavage, addition of nucleotides, methylation or other modification of the nucleotides, and removal of introns by splicing.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

12.0: Overview of RNA Processing

RNA processing refers to any covalent modification to the RNA that occurs after transcription. This includes specific cleavage, addition of nucleotides, methylation or other modification of the nucleotides, and removal of introns by splicing.

Overview

RNA	Precursor	Modification	Addition	Cleavage	Splicing
mRNA	pre-mRNA (hnRNA)	methylation on 2'-OH of ribose	5' cap 3' poly A	cut at site for poly A; excise viral mRNA	remove introns
rRNA	pre-rRNA	methylation on 2'OH of ribose	no	excise products fr. precursor	remove introns
tRNA	pre-tRNA	extensive and varied	CCA to 3' end	yes	remove introns
snRNAs	?	?	5' cap	?	?

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

12.1: Cutting and Trimming RNA

pre-rRNA

In *E. coli*, the *rrn* operon is transcribed into a 30S precursor RNA, containing 3 rRNAs and 2 tRNAs.

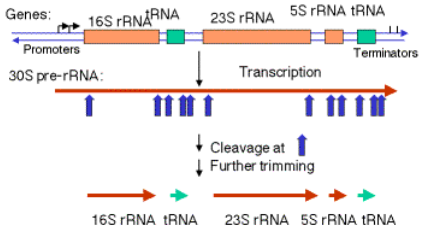


Figure 3.3.1. Excision of rRNAs and tRNAs from 30S precursor RNA

The segment containing 16S rRNA (small ribosomal subunit) and the one containing 23S rRNA (large ribosomal subunit) are flanked by inverted repeats that form stem structure in the RNA. The stems are cleaved by RNase III. There is no apparent single sequence at which RNase III cleaves - perhaps it recognizes a particular stem structure. This plus subsequent cleavage events (by an activity called M16) generates the mature 16S and 23S rRNAs. The rRNAs are also methylated.

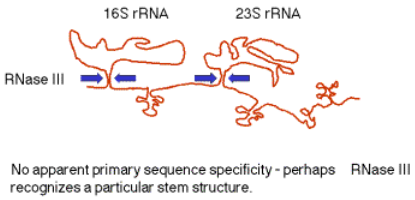


Figure 3.3.2. RNase III cuts in the stems of stem-loops in RNA

tRNA is liberated by RNases P and F and 5S rRNA is liberated by RNases E and M5

In eukaryotes

The initial precursor is 47S and contains ETS1, 18S rRNA, ITS1, 5.8S rRNA, ITS2, and 28S rRNA, where ETS = extragenic transcribed spacer and ITS = intragenic transcribed spacer. Specific cleavage events followed by methylations generate the mature products. Also, some rRNA genes in some species have introns that must be spliced out.

pre-tRNA in *E. coli*

Sequence specific cleavage by RNases P, F, D

1. RNase P is an endonuclease that cleaves the precursor to generate the 5' end of the mature tRNA.
2. RNase F is an endonuclease that cleaves the precursor 3 nucleotides past the 3' end of the mature tRNA.
3. RNase D is an exonuclease that trims in a 3' to 5' direction to generate the 3' end or the mature tRNA.

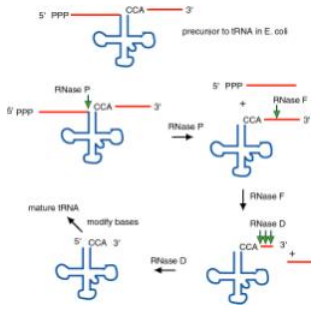


Figure 3.3.3. The ends of tRNA in E. coli are produced by the action of three nucleases that cleave the precursor to tRNA. A schematic of the pre-tRNA is shown at the top, with RNA extending from the 5' and 3' ends of the RNA that will become the mature tRNA (shown as a cloverleaf). The site of cleavage is indicated by the short vertical arrows above the lines denoting RNA, and they are labeled with the name of the enzyme cutting at that site. The enzymes catalyzing each reaction are listed above or adjacent to the reaction arrows.

The catalytic activity of RNase P is in the RNA component

1. RNAse P is composed of a 375 nt RNA and a 20 kDa protein.
2. The catalytic activity is in the RNA. The protein is thought to aid in the reaction, but is not required for catalysis. All enzymes are not proteins!
3. This was one of the first instances discovered of catalytic RNA, and Sidney Altman shared the Nobel Prize for this.

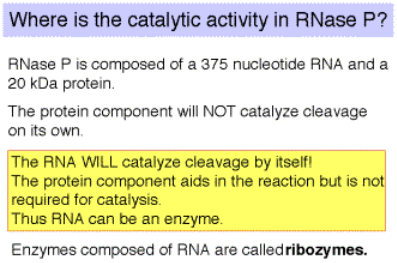


Figure 3.3.4. RNase P

The enzyme **tRNA nucleotidyl transferase**adds CCA to the 3' ends of pre-tRNAs.

1. Virtually all tRNAs end in CCA, forms the amino acceptor stem.
2. For most prokaryotic tRNA genes, the CCA is encoded at the 3' end of the gene.
3. No known eukaryotic tRNA gene encodes the CCA, but rather it is added posttranscriptionally by the enzyme tRNA nucleotidyl transferase. This enzyme is present in a wide variety of organisms, including bacteria, in the latter case presumably to add CCA to damaged tRNAs.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

12.2: Modifications at the 5' and 3' ends of mRNA

As discussed previously, eukaryotic mRNAs are capped at their 5' end and polyadenylated at their 3' end. In vitro assays for these reactions have been developed, and several of the enzymatic activities have been identified. These will be reviewed in this section. Polyadenylation is not limited to eukaryotes. Several mRNAs in E. coli are polyadenylated as well. This is a fairly new area of study.

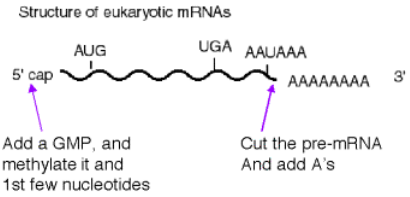


Figure 3.3.5. mRNAs can be modified on the 5’ and 3’ ends.

Modification at the 5' end: Cap Structure

The "cap" is a methylated 5'-GMP that is linked via its 5' phosphate to the b-phosphoryl of the initiating nucleotide (usually A); see Figure 3.3.6. Capping occurs shortly after transcription has begun. It occurs in a series of enzymatic steps (Figure 3.3.7):

1. Remove the g-phosphoryl of the initiating nucleotide (RNA triphosphatase)
2. Link a GMP to the b-phosphoryl of the initiating nucleotide (mRNA guanylyl transferase). The GMP is derived from GTP, and is linked by its 5' phosphate to the 5' diphosphate of the initiating nucleotide. Pyrophosphate is released.
3. The N-7 of the cap GMP is methylated (methyl transferase), donor is S-adenosyl methionine.
4. Subsequent methylations occur on the 2' OH of the first two nucleotides of the mRNA.

Capping has been implicated in having a role in efficiency of translation and in mRNA stability.

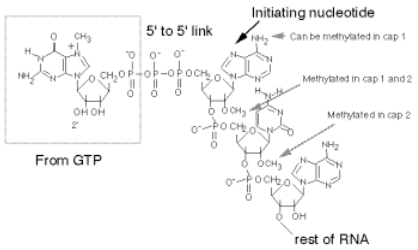


Figure 3.3.6. Structure of the 5’ cap on eukaryotic mRNAs.

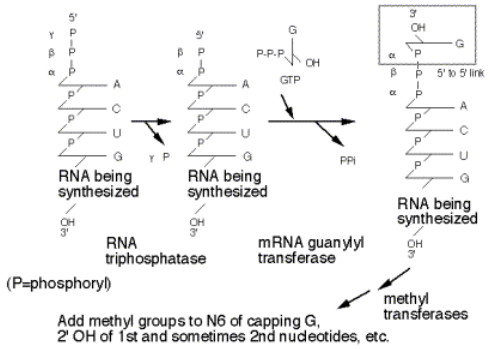


Figure 3.3.7. Stepwise synthesis of the 5’ cap.

Several proteins are required for cleavage and polyadenylation at the 3' end:

1. CPSF is a tetrameric specificity factor; it recognizes and binds to the AAUAAA polyadenylation signal.
2. CFI and CFII are cleavage factors.
3. PAP is the polyA polymerase.
4. CFI, CFII and PAP form a complex that binds to the nascent RNA at the cleavage site, directed by the CPSF specificity factor.
5. CstF is an additional protein implicated in this process in vitro, but its precise function is currently unknown.

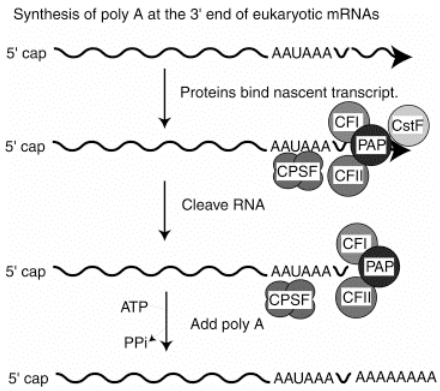


Figure 3.3.8

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

12.3: Multiple Mechanisms are Used for Splicing Different Types of Introns

Different Types of Introns

At least four distinct classes of introns have been identified: Introns in nuclear protein-coding genes that are removed by spliceosomes (spliceosomal introns) Introns in nuclear and archaeal transfer RNA genes that are removed by proteins (tRNA introns) Self-splicing group I introns that are removed by RNA catalysis.

1. pre-tRNA
2. group I, group II: Introns in fungal mitochondrial genes and in plastid (chloroplast) genes have been grouped into two different groups based on different consensus sequences found in the introns. As we will see below, the group II introns have a mechanism for splicing that is similar to that of pre-mRNA.
3. pre-mRNA

In all cases, splicing will remove the introns and join the exons to give the mature RNA.

Table. Features of splicing for different types of introns

Class	Distribution	Sequence	Distinguishing feature	Mechanism
pre-tRNA	yeast to mammals	very short (10-20 nucleotides)	requires ATP	cut, kinase, ligase
group I	fungal mitochondria, plastids, pre-rRNA in Tetrahymena	characteristic consensus	self-splicing, G nucleot(s)ide to initiate	phosphoester transfer
group II	fungal mitochondria, plastids	characteristic consensus	can self-splice, internal A nucleotide to initiate	phosphoester transfer
pre-mRNA	yeast to mammals	5' GU...AG 3'	spliceosome (ATP for assembly), internal A nucleotide to initiate	phosphoester transfer

Splicing of pre-tRNAs

Some precursor tRNAs contain short introns (only 10 to 20 nucleotides) with no apparent consensus sequences. These short introns are removed in a series of steps catalyzed by enzymes that include an endonuclease, a kinase and a ligase. Because the endonuclease generates a 2', 3' cyclic phosphodiester product, an additional phosphodiesterase is needed to open the cyclic phosphodiester to provide the 3' hydroxyl for the ligase reaction. In addition, the 2'-phosphate (product of the phosphodiesterase) must be removed by a phosphatase. This process uses two ATPs for every splicing event.

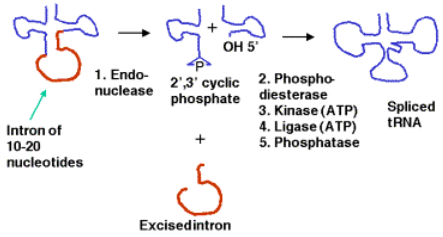


Figure 3.3.9. Steps in splicing of pre-tRNA.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

12.4: Self-splicing by group I introns (pre-rRNA of Tetrahymena)

An *in vitro* reaction was established to examine the removal of an intron from the precursor to rRNA in *Tetrahymena*. Suprisingly, it was discovered that the splicing of the pre-RNA occurred in the absence of any added protein!

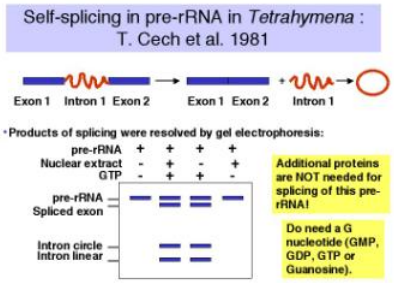


Figure 3.3.10. Discovery of self-splicing in T. Cech's lab, 1981

Further investigation revealed that the reaction requires a guanine nucleotide or nucleoside with a 3'-OH, plus mono- and divalent cations. GTP, GDP, GMP or guanosine will work to initiate splicing. There is no requirement for protein or high energy bond cleavage

Self-splicing occurs by a phosphoester transfer mechanism (Figure 3.3.11)

The 3'-OH of the guanine nucleotide is the nucleophile that attacks and joins to the 5' phosphate of the first nucleotide of the intron. This leaves the 3'-OH of the last nucleotide of the upstream exon available to attack and join the 5' phosphate of the first nucleotide of the downstream exon. These two phosphoester transfers result in a joining of the two exons and excision of the intron (with the initiating G nucleotide attached to the 5' end.) The excised intron is then circularized by attack of the 3'-OH of the last nucleotide of the intron on the phosphate between the 15th and 16th nucleotides of the introns. Further degradation effectively removes the intron from the reaction and helps prevent the reverse reaction from occurring. Note that the phosphoester transfers are readily reversible unless the products (excised intron) are removed. There is no increase or decrease in the number of phosphoester bonds during this splicing.

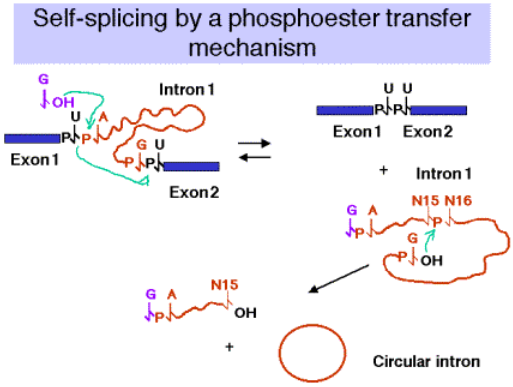


Figure 3.3.11.

The Intron is the Catalyst for Splicing in this System

RNA involvement in self-splicing is stoichiometric, but the excised intron does have a catalytic activity in vitro. After a series of intramolecular cyclization and cleavage reactions, the linear excised intron lacking 19 nucleotides (called L-19 IVS) can be used catalytically to add and remove nucleotides to an artificial substrate. For instance, C5, which is complementary to the internal guide sequences of the intron, can be converted to C4 + C6 and other products (Figure 3.3.12).

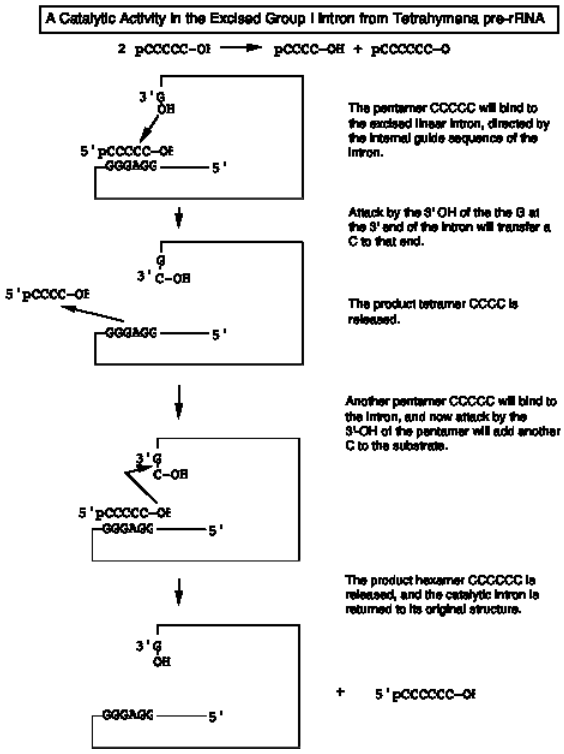
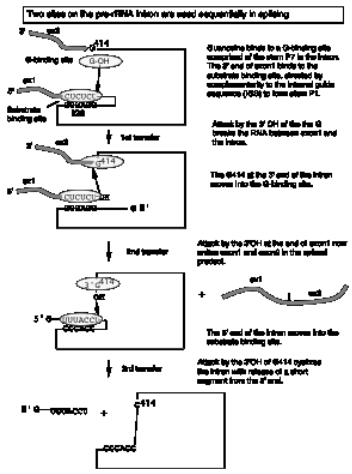


Figure 3.3.12.

The 3-D structure of the folded RNA is responsible for the specificity and efficiency of the reaction (analogous to the general ideas about proteins with enzymatic activity). The specificity of splicing is caused, at least in part, by base-pairing between the 3' end of the upstream exon and a region in the intron called the internal guide sequence. The initiating G nt also binds to a specific site in the RNA close to the 5' splice site. Thus two sites in the pre-rRNA intron are used sequentially in splicing (Figure 3.3.13 A and 3.3.13.B.).

Figure 3.3.13.A.



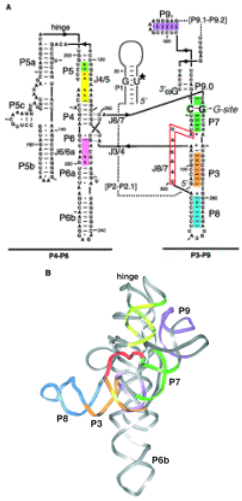


Figure 3.3.13.B. The catalytic domain of the group I intron from *Tetrahymena* pre-rRNA, shown in the RNA secondary structure view (left panel) and in a view of the tertiary structure (right panel).

The internal guide sequence (IGS) is not not required for catalysis but does confer specificity. Thus one can design RNAs for exon exchange in cells. This potential avenue for therapy for genetic disorders is called "exon replacement therapy.

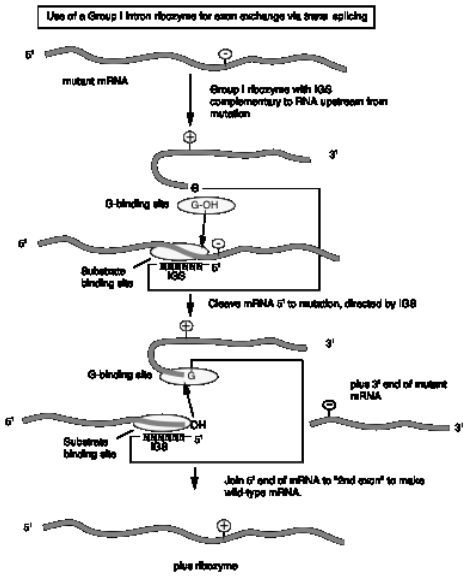


Figure 3.3.14

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

12.5: RNAs Can Function as Enzymes

Examples include the following:

- RNase P
- Group I introns (includes intron of pre-rRNA in Tetrahymena)
- Group II introns
- RNA: peptide bond formation

Hammerhead ribozymes

Viroids and virusoids have a self-cleaving activity that localized to a 58 nucleotide structure illustrated in Figure 3.3.15. The mechanism differs in some respects from the phosphoester transfer. A divalent metal hydroxide binds in the active site, and abstracts a proton from the 2' OH of the nucleotide at the cleavage site. This now serves as a nucleophile to attack the 3' phosphate and cleave the phosphodiester bond, generating a 2',3' cyclic phosphate and a 5' OH on the ends of the cleaved RNA.

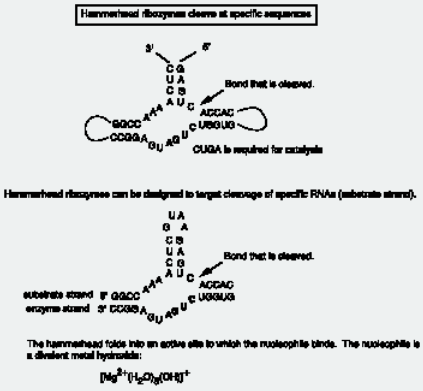


Figure 3.3.15.

One application currently being explored is the use of designed hammerheads to cleave a particular mRNA, thereby turning off expression of a particular gene. If over-expression or ectopic expression of a defined gene were the cause of some pathology (e.g. some form of cancer), then reducing its expression could have therapeutic value.

Other RNAs possibly involved in catalysis, such as the snRNAs involved in splicing pre-mRNA.

Even though RNAs can be sufficient for catalysis, sometimes they are assisted by proteins to improve efficiency. For instance, group I introns are capable of splicing introns by themselves in a cell-free reaction. However, some are not very efficient in this process, and in the cell they are assisted by proteins that themselves are not catalytic but they enhance the reaction. Examples are maturases, which are proteins that assist in the splicing of some group I introns found in yeast mitochondria.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

12.6: Splicing of introns in pre-mRNAs

1. Splice Sites

The sequence at the 5' and 3' ends of introns in pre-mRNAs is very highly conserved. Thus one can derive a **consensus sequence** for splice junctions.

5' exon...AG'**GU**RAGU.....YYYYYYYYYYNC**AG**'G....exon

The GU is the **5' splice site** (sometimes called the *donor splice site*) and the AG is the **3' splice site** (or acceptor splice site). GU is invariant at the 5' splice site, and AG is (almost) invariant at the 3' splice site for the most prevalent class of introns in pre-mRNA.

Effects of mutations at the splice junctions demonstrate their importance in the splicing mechanism. Mutation of the GT at the donor site in DNA to an AT prevents splicing (this was seen in a mutation of the b-globin gene that caused b0 thalassemia.) A different mutation of the b-globin gene that generated a new splice site caused an aberrant RNA to be made, resulting in low levels of b-globin being produced (b+ thalassemia).

2. The intron is excised as a lariat

The 2'-OH of an A at the "branch" point forms a phosphoester with the first G of the intron to initiate splicing. Splicing occurs by a series of phosphoester transfers (also called trans-esterifications). After the 2'-OH of the A at the branch has joined to the initial G of the intron, the 3'-OH of the upstream exon is available to react with the first nucleotide of the downstream exon, thereby joining the two exons via the phosphoester transfer mechanism.

c. Intron lariat is the equivalent of a "circular" intermediate.

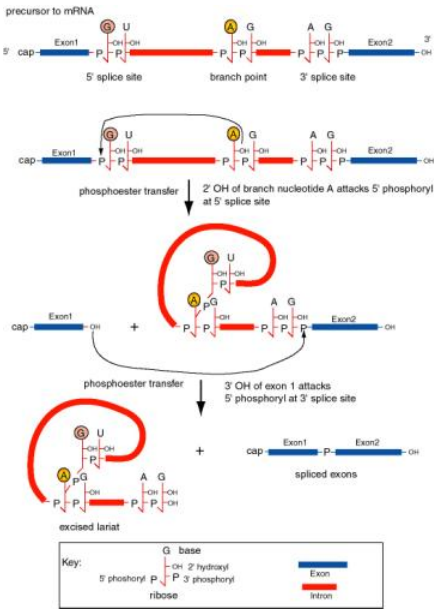


Figure 3.3.16. Splicing of precursor to mRNA excises the intron as a lariat structure. The chemical reactions are two phosphoester transfers. The first transfer is initiated by the 2’ hydroxyl of the adenine ribonucleoside at the branch point, which attacks the 5’ phosphoryl of the 5’ splice site. This generates a 3’ hydroxyl at exon 1 and joins the A at the branch point to the U at the 5’ splice site, producing a lariat in the intron. The second transfer is initiated by the attack of the newly exposed 3’ hydroxyl of exon 1 on the 5’ phosphoryl of exon 2. The latter reaction joins the two exons and releases the intron as a lariat.

The sequence at the branch point is only moderately conserved in most species; examination of many branch points gives the consensus YNYYRAG. It lies 18 to 40 nucleotides upstream of the 3' splice site.

3. Small nuclear ribonucleoproteins (or snRNPs) form the functional splicesome on pre-mRNA and catalyze splicing.

a. "U" RNAs and associated proteins. Small nuclear RNAs (**snRNA**s) are about 100 to 300 nts long and can be as abundant as 105 to 106 molecules per cell. They are named U followed by an integer. The major ones involved in splicing are U1, U2, U4/U6, and U5 snRNAs. They are conserved from yeast to human. The snRNAs are associated with proteins to form small nuclear

ribonucleoprotein particles, or **snRNPs**. The snRNPs are named for the snRNAs they contain, hence the major ones involved in splicing are U1, U2, U4/U6, U5 snRNPs.

One class of proteins common to many snRNPs are the **Sm proteins**. There are 7 Sm proteins, called B/B', D1, D2, D3, E, F, G. Each Sm protein has similar 3-D structure, consisting of an alpha helix followed by 5 beta strands. The Sm proteins interact via the beta strands, and may form circle around RNA.

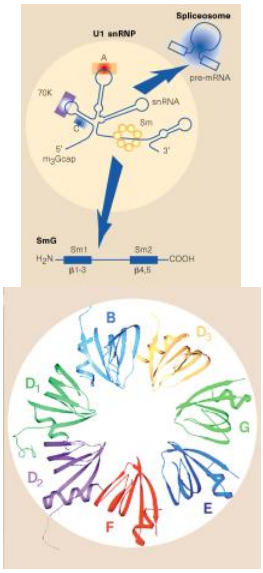


Figure 3.3.17). The right panel shows interactions of the Sm proteins through their beta-strands to make a ring with an inner portion large enough to encircle an RNA molecule. From Angus I. Lamond (1999) Nature 397, 655 - 656 "RNA splicing: Running rings around RNA."

A particular sequence common to many snRNAs is recognized by the Sm proteins, and is called the "Sm RNA motif".

b. Use of antibodies from patients with SLE. Several of the common snRNPs are recognized by the autoimmune serum called anti-Sm, initially generated by patients with the autoimmune disease Systemic Lupus Erythematosis. One of the critical early experiments showing the importance of snRNPs in splicing was the demonstration that anti-Sm antisera is a potent inhibitor of *in vitro*splicing reactions. Thus the targets of the antisera, i.e. Sm proteins in snRNPs, are needed for splicing.

c. The snRNPs assemble onto the pre-mRNA to make a large protein-RNA complex called a **spliceosome** (Figure 3.3.17). Catalysis of splicing occurs within the spliceosome. Recent studies support the hypothesis that the *snRNA components of the spliceosome actually catalyze splicing*, providing another example of ribozymes.

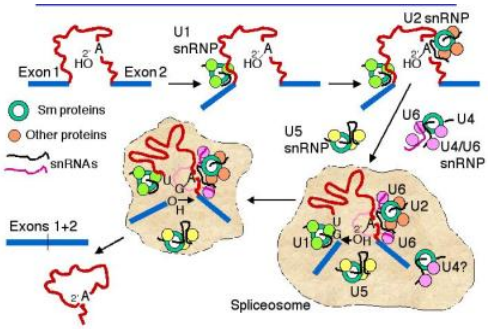


Figure 3.3.17. Spliceosome assembly and catalysis

d. U1 snRNP: Binds to the 5' splice site, and U1 RNA forms a base-paired structure with the 5' splice site.

e. U2 snRNP: Binds to the branch point and forms a short RNA-RNA duplex. This step requires an **a**uxiliary **f**actor (U2AF) and ATP hydrolysis, and commits the pre-mRNA to the splicing pathway.

f. U5 snRNP plus the U4, U6 snRNP now bind to assemble the functional spliceosome. Evidence indicates that U4 snRNP dissociates from the U6 snRNP in the spliceosome. This then allows U6 RNA to form new base-paired structures with the U2 RNA and the pre-mRNA that catalyze the transesterification reaction (phosphoester transfers). One model is that U6 RNA pairs with the

5' splice site and with U2 RNA (which itself is paired to the branch point), thus bringing the branch point A close to the 5' splice site. U5 RNA may serve to hold close together the ends of the exons to be joined.

4. *Trans*-splicing

All of the splicing we have discussed so far is between exons on the same RNA molecule, but in some cases exons can be spliced to other RNAs. This is very common in trypanosomes, in which a spliced leader sequence is found at the 5' ends of almost all mRNAs. A few examples of *trans*splicing have been described in mammalian cells.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

12.7: Splicing of group II introns

1. Similar mechanism as that for nuclear pre-mRNA splicing.
2. Can occur by self-splicing, albeit under rather artificial conditions.
3. Reaction can be reversible (as can splicing of group I introns), leading to the idea that these introns can be transposable elements.
4. The group II self-splicing may be the evolutionary ancestor to nuclear pre-mRNA splicing

I. Mechanistic similarties for splicing group I, group II and pre-mRNA introns

1. All involve transesterification = phosphoester transfers. No high energy bonds are utilized in the splicing process; the arrangement of phosphodiester bonds is reorganized, and as a result exons are joined together.

2. The initiating nucleophile is the 3' OH of a guanine nucleotide for Group I introns, whereas for Group II introns and introns in pre-mRNA, it is the 2' OH of an internal adenine nucleotide in the intron.

3. In all cases, particular secondary structures in the RNAs are utilized to bring together the reactive components (e.g. ends of exons and introns). These secondary structures may be intramolecular in the case of self-splicing Group I and Group II introns, or they may be intermolecular in the case of pre-mRNA and the snRNAs, e.g. those in the U1, U2, perhaps U6 snRNPs.

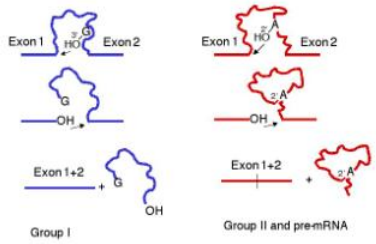


Figure 3.3.18. Common features of the mechanism of splicing in Group I introns and in Group II introns plus introns in precursor to mRNA.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

12.8: Alternative Splicing

For many genes, all the introns in the mRNA are spliced out in a unique manner, resulting in one mRNA per gene. But there is a growing number of examples of other genes in which certain exons are included or excluded from the final mature mRNA, a process called *alternative splicing*. Some exons may be included in some tissues and not others, or may be sex-specific, indicating some regulation over the selection of splice sites.

Alternative splicing of pre-mRNA means that a single gene may encode more than one protein product (e.g., sex determination in *Drosophila melanogaster).*

Drosophila melanogaster

The X to autosome ratio (X:A ratio) in the zygote will determine which of two different developmental pathways along which the fly will develop. If the X:A ratio is high (e.g. the female is XX and the X:A ratio is 1.0), the fly will utilize the female pathway; if the ratio is low (e.g. 0.5 since the male is XY), it will develop as a male.

The X:A ratio is determined by "counting" certain genes (or their expression) on the X chromosome (e.g. *sisterless a*, *sisterless b*, and *runt*) for the numerator and counting other genes (such as *deadpan*) for the denominator. All of the products of these genes are homologous to various calsses of transcription factors, consistent with at least part of the regulation of sex determination being transcriptional. However, as discussed below, alternative splicing plays a key role as well, at least in *Drosophila*.

The pathways have at least four steps that were defined genetically by mutations that caused, e.g. genetically female flies (high X:A) to develop as males. In each case, the same gene encodes both male and female-specific mRNAs (and proteins), but the sex-specific mRNAs (and proteins) differ as a result of alternative splicing.

In all cases, the default state is male development, and some new activity has to be present to establish and maintain the female pathway.

1. The target of the X:A signal is the *Sex-lethal*gene (*Sxl*), which serves as a master switch gene. In early development, an X:A ratio of 1 in females leads to the activiation of an embryo-specific promoter of the *Sxl*gene, whereas *Sxl*is not transcribed in male embryos. Later in development, *Sxl*is transcribed in both sexes. Now the high X:A ratio leads to the skipping of an exon in the splicing of pre-mRNA from the *Sex-lethal*gene. This produces a functional Sxl protein in females. In males (default pathway), the mRNA has an early termination codon, and no functional Sxl protein is made.

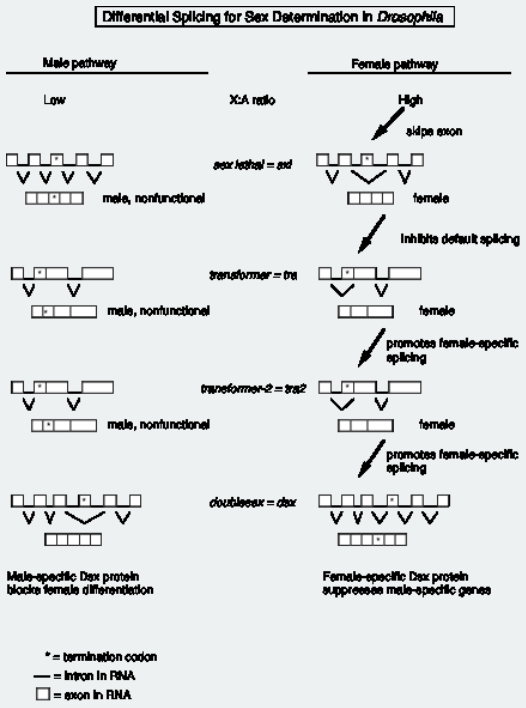


Figure 3.3.19.

2. A functional Sxl protein inhibits the default splicing of pre-mRNA from the *transformer*gene, to generate a functional Tra protein in female embryos. In the female-specific splicing of *tra*pre-RNA, a 5' splice site (common to both male and female splicing) is connected to an alternative 3' splice site, thereby removing a termination codon and allowing function Tra protein to be made (Figure 3.3.15).
3. The Tra protein promotes female-specific splicing of pre-mRNA from the *tra2*gene, again generating a functional Tra2 protein only in females.
4. Tra and Tra2 proteins promote female-specific splicing of pre-mRNA from the *doublesex*gene (*dsx*). In this case, the male-specific mRNA has skipped an exon (Figure 3.3.15). Skipping an exon requires an alteration in the splicing pattern at both the 3' splice site and the 5' splice sites surrounding the exon.
5. The male-specific Dsx protein blocks female differentiation and leads to male development. The female-specific Dsx protein represses expression of male genes and leads to female development.

Some clues about mechanism

1. Tra and Tra2 are RNA-binding proteins related to Splicing Factor 2 (SF2). This latter protein has a domain rich in the dipeptide Arg-Ser, which defines one type of RNA-binding domain. SF2 is required for early steps in spliceosome assembly. The related Tra and Tra2 proteins are not required for viability, but they do regulate the specific splicing events for pre-mRNA from dsx.
2. Tra2 binds in the female-specific exon of the dsx transcript, and presumably regulates splice site selection. The binding site for Tra2 within the exon is an example of a splicing enhancer. The mechanisms by which the binding of splicing regulatory proteins (e.g. Tra, Tra2) to splicing enhancers is a very active area of research currently.
3. Sxl is another RNA-binding protein that inhibits the default splicing pattern for tra pre-mRNA.

Figure 3.3.20.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

12.9: RNA editing

RNA editing refers to changing the sequence of RNA after transcription, either by adding nucleotides, taking them away, or substituting one for another. The extent of editing is dramatic in some mRNAs, e.g. in the mitochondria of trypansomes and Leishmania. For example, for some mRNAs 55% of the nucleotide sequence is added after transcription! In many of the cases characterized so far, a small number of U's are inserted at many places in the mRNA. Other examples of excising U's and adding C's are known for other mitochondrial genes from other organisms.

In at least some cases, the additional nucleotides are added under the direction of guide RNAs that are encoded elsewhere in the mitochondrial genome. A portion of the guide RNA is complementary to the mRNA in the vicinity of the position at which nucleotides will be added (Figure 3.3.16). The U at the 3' end of the guide RNA initiates a series of phosphoester transfer reactions to insert itself into the mRNA (see bottom of Figure 3.3.16). More U's at the 3' end of the guide RNA can be added, one at a time. Note the similarity in mechanism between these insertions of nucleotides (editing) and the self-splicing of Group I intron.

For a situation in which one segment of DNA encodes the unedited mRNA and two other segments of DNA encode the guide RNAs required for editing, the "gene" is encoded in three portions, mutations in which would complement in trans! This is a counter-example to one of our most powerful definitions of a gene.

In mammals, two different forms of apolipoprotein B are made, one in the liver and one in the intestine. The intestinal form is much shorter because of an earlier termination codon. Surprisingly, only one gene is found and it must encode both from of ApoB. A specifc enzyme must change one nucleotide of the mRNA for apolipoprotein B (a C in codon 2153, CAA) post-transcriptionally from a C to a U to generate the termination codon (UAA) found in the intestinal form.

This enzymatic activity is present in a protein with *no* apparent RNA component, and hence no obvious guide RNA. Thus it appears to operate by a distinctly different mechanism from the editing in protist mitochondria (see. e.g. Greeve, J. et al., 1991, Nucleic Acids Research 19: 3569-3576).

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

12.E: RNA Processing (Exercises)

12.1 Nucleoside triphosphates labeled with [32P] at the a, b, or g position are useful for monitoring various aspects of transcription. For the specific process listed in a-c, give the position of the label that is appropriate for examining that step.

a) Initiation by *E. coli*RNA polymerase.

b) Forming the 5' end of eukaryotic mRNA.

c) Elongation by eukaryotic RNA polymerase II.

12.2 (POB) RNA posttranscriptional processing.

Predict the likely effects of a mutation in the sequence (5')AAUAAA in a eukaryotic mRNA transcript.

12.3 A phosphoester transfer mechanism (or transesterification) is observed frequently in splicing and other reactions involving RNA. Are the following statements about these mechanisms true or false?

a) The mechanism requires the cleavage of high-energy bonds from ATP.

b) The initiating nucleophile for splicing of Group I introns (including the intron of pre-rRNA from Tetrahymena) is the 3' hydroxyl of a guanine nucleotide.

c) The initiating nucleophile for splicing of nuclear pre-mRNA is the 2' hydroxyl of an internal adenine nucleotide.

d) The individual reactions in the phosphoester transfers are reversible, but the overall process is essentially irreversible because of circularization (includes lariat formation) of the excised intron.

12.4 What properties are shared by the splicing mechanism of Tetrahymena pre-rRNA and Group II fungal mitochrondrial introns?

12.5 Please answer these questions on splicing of precursors to mRNA.

a) What dinucleotides are almost invariably found at the 5’ and 3’ splice sites of introns?

b) Which splicing component binds at the 5' splice junction?

c) What nucleotides are joined by the branch structure in the intron during splicing?

d) What is ATP used for during splicing of precursors to mRNA?

12.6 (POB) RNA splicing.

What is the minimum number of transesterification reactions needed to splice an intron from an mRNA transcript? Why?

12.7 Match the following statements with the appropriate eukaryotic splicing process listed in parts a-c.

1) A guanine nucleoside or nucleotide initiates a concerted phosphotransfer reaction.

2) The consensus sequences at splice junctions are AG'GUAAGU...YYYAG'G (' is the junction, Y = any pyrimidine).

3) Splicing occurs in two separate steps, cutting to generate a 3'-phosphate followed by an ATP dependent ligation.

4) Splicing requires no protein factors.

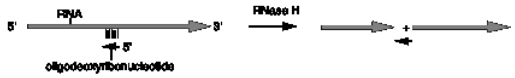
5) Splicing requires U1 small nuclear ribonucleoprotein complexes.

a) Splicing of pre-mRNA.

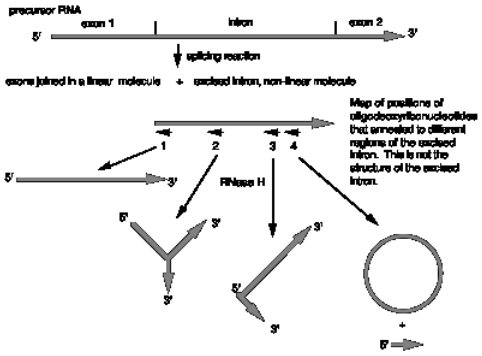
b) Splicing of pre-tRNA in yeast

c) Splicing of pre-rRNA in Tetrahymena

12.8 The enzyme RNase H will cleave any RNA that is in a heteroduplex with DNA. Thus one can cleave a single-stranded RNA in any specific location by first annealing a short oligodeoxyribonucleotide that is complementary to that location and then treating with RNase H.



This approach is useful in determining the structure of splicing intermediates. Let's consider a hypothetical case shown in the figure below. After incubation of radiolabeled precursor RNA (exon1-intron-exon2) with a nuclear extract that is capable of carrying out splicing, the products were analyzed on a denaturing polyacrylamide gel. The results showed that the exons were joined as linear RNA, but the excised intron moved much slower than a linear RNA of the same size, indicative of some non-linear structure. The excised intron was annealed to a short oligodeoxyribonucleotide that is complementary to the region at the 5' splice site (labeled oligo 1 in the figure), treated with RNase H and analyzed on a denaturing polyacrylamide gel. The product ran as a linear RNA with the size of the excised intron (less the length of the RNase H cleavage site). As summarized in the figure, the excised intron was analyzed by annealing (separately) with three other oligodeoxyribonucleotides, followed by RNase H treatment and gel electrophoresis. Use of oligodeoxyribonucleotide number 2 generated a Y-shaped molecule, use of oligodeoxyribonucleotide number 3 generated a V-shaped molecule with one 5' end and 2 3' ends, and use of oligodeoxyribonucleotide number 4 generated a circle and a short linear RNA.



(a) What does the result with oligodeoxyribonucleotide 2 tell you?

(b) What does the result with oligodeoxyribonucleotide 4 tell you?

(c) What does the result with oligodeoxyribonucleotide 1 tell you?

(d) What does the result with oligodeoxyribonucleotide 3 tell you?

(e) What is the structure of the excised intron? Show the locations of the complementary oligos on your drawing.

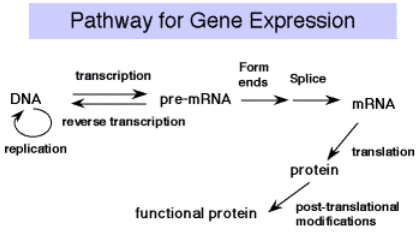
13: Genetic code

Once transcription and processing of rRNAs, tRNAs and snRNAs are completed, the RNAs are ready to be used in the cell - assembled into ribosomes or snRNPs and used in splicing and protein synthesis. But the mature mRNA is not yet functional to the cell. It must be translated into the encoded protein. The rules for translating from the "language" of nucleic acids to that of proteins is the **genetic code**.

Introduction

Experiments testing the effects of frameshift mutations showed that the deletion or addition of 1 or 2 nucleotides caused a loss of function, whereas deletion or addition of 3 nucleotides allowed retention of considerable function. This demonstrated that the coding unit is 3 nucleotides. The nucleotide triplet that encodes an amino acid is called a **codon**. Each group of three nucleotides encodes one amino acid. Since there are 64 combinations of 4 nucleotides taken three at a time and only 20 amino acids, the code is **degenerate** (more than one codon per amino acid, in most cases). The adaptor molecule for translation is **tRNA**. A charged tRNA has an amino acid at one end, and at the other end it has an anticodon for matching a codon in the mRNA; ie. it "speaks the language" of nucleic acids at one end and the "language" of proteins at the other end. The machinery for synthesizing proteins under the direction of template mRNA is the **ribosome**.

Figure 3.4.1. tRNAs serve as an adaptor for translating from nucleic acid to protein



A. Size of a codon: 3 nucleotides

1. Three is the minimum number of nucleotides per codon needed to encode 20 amino acids.

   a. 20 amino acids are encoded by combinations of 4 nucleotides

   b. If a codon were two nucleotides, the set of all combinations could encode only

      4x4 = 16 amino acids.

   c. With three nucleotides, the set of all combinations can encode

      4x4x4 = 64 amino acids (i.e. 64 different combinations of four nucleotides taken three at a time).

2. Results of combinations of frameshift mutations show that the code is in triplets. Length-altering mutations that add or delete one or two nucleotides have severe defective phenotype (they change the reading frame, so the entire amino acid sequence after the mutation is altered.). But those that add or delete three nucleotides have little or no effect. In the latter case, the reading frame is maintained, with an insertion or deletion of an amino acid at one site. Combinations of three different single nucleotide deletions (or insertions), each of which has a loss-of-function phenotype individually, can restore substantial function to a gene. The wild-type reading frame is restored after the 3rd deletion (or insertion).

B. Experiments to decipher the code

1. Several different cell-free systems have been developed that catalyze protein synthesis. This ability to carry out translation in vitro was one of the technical advances needed to allow investigators to determine the genetic code.

   a. Mammalian (rabbit) reticulocytes: ribosomes actively making lots of globin.

   b. Wheat germ extracts

   c. Bacterial extracts

2. The ability to synthesize random polynucleotides was another key development to allow the experiments to decipher the code. S. Ochoa isolated the enzyme polynucleotide phosphorylase, and showed that it was capable of linking nucleoside **di**phosphates

(NDPs) into polymers of NMPs (RNA) in a reversible reaction.

nNDP n + nPi

The physiological function of polynucleotide phosphorylase is to catalyze the reverse reaction, which is used in RNA degradation. However, in a cell-free system, the forward reaction is very useful for making random RNA polymers.

3. Homopolymers program synthesis of specfic homo-polypeptides (Nirenberg and Matthei, 1961).

a. If you provide only UDP as a substrate for polynucleotide phosphorylase, the product will be a homopolymer poly(U).

b. Addition of poly(U) to an in vitro translation system (e.g. E. coli lysates), results in a newly synthesized polypeptide which is a polymer of polyphenylalanine.

c. Thus UUU encodes Phe.

d. Likewise, poly(A) programmed synthesis of poly-Lys; AAA encodes Lys.

Poly(C) programmed synthesis of poly-Pro; CCC encodes Pro.

Poly(G) programmed synthesis of poly-Gly; GGG encodes Gly.

4. Use of mixed co-polymers

a. If two NDPs are mixed in a known ratio, polynucleotide phosphorylase will make a mixed co-polymer in which nucleotide is incorporated at a frequency proportional to its presence in the original mixture.

b. For example, consider a 5:1 mixture of A:C. The enzyme will use ADP 5/6 of the time, and CDP 1/6 of the time. An example of a possible product is:

AACAAAAACAACAAAAAAAACAAAAAACAAAC...

Table 3.4.1. Frequency of triplets in a poly(AC) (5:1) random copolymer

Composition	Number	Probability	Relative frequency
3 A	1	0.578	1.0
2 A, 1 C	3	3 x 0.116	3 x 0.20
1 A, 2 C	3	3 x 0.023	3 x 0.04
3 C	1	0.005	0.01

c. So the frequency that AAA will occur in the co-polymer is

(5/6)(5/6)(5/6) = 0.578.

This will be the most frequently occurring codon, and can be normalized to 1.0 (0.578/0.578 = 1.0)

d. The frequency that a codon with 2 A's and 1 C will occur is

(5/6)(5/6)(1/6) = 0.116.

There are three ways to have 2 A's and 1 C, i.e. AAC, ACA and CAA. So the frequency of occurrence of all the A2C codons is 3 x 0.116.

Normalizing to AAA having a relative frequency of 1.0, the frequency of A2C codons is 3 x (0.116/0.578) = 3 x 0.2.

e. Similar logic shows that the expected frequency of AC2 codons is 3 x 0.04, and the expected fequency of CCC is 0.01.

Table 3.4.2. Amino acid incorporation with poly(AC) (5:1) as a template

Radioactive	Precipitable cpm		Observed	Theoretical
amino acid	- template	+ template	incorporation	incorporation
Lysine	60	4615	100.0	100
Threonine	44	1250	26.5	24
Asparagine	47	1146	24.2	20
Glutamine	39	1117	23.7	20

Radioactive	Precipitable cpm		Observed	Theoretical
Proline	14	342	7.2	4.8
Histidine	282	576	6.5	4

These data are from Speyer et al. (1963) Cold Spring Harbor Symposium in Quantitative Biology, 28:559. The theoretical incorporation is the expected value given the genetic code as it was subsequently determined.

f. When this mixture of mixed copolymers is used to program in vitro translation, Lys is incorporated most frequently, which can be expressed as 100. This confirms that AAA encodes Lys.

g. Relative to Lys incorporation as 100, Thr, Asn, and Gln are incorporated with values of 24 to 26, very close to the expectation for amino acids encoded by one of the A2C codons. However, these data do not show which of the A2C codons encodes each specific amino acid. We now know that ACA encodes Thr, AAC encodes Asn, and CAA encodes Gln.

h. Pro and His are incorporated with values of 6 and 7, which is close to the expected 4 for amino acids encoded by AC2 codons. E.g. CCA encodes Pro, CAC encodes His. ACC encodes Thr, but this incorporation is overshadowed by the “26.5” units of incorporation at ACA. Or, more accurately, “26.5” @ 20 (ACA) + 4 (ACC) for Thr.

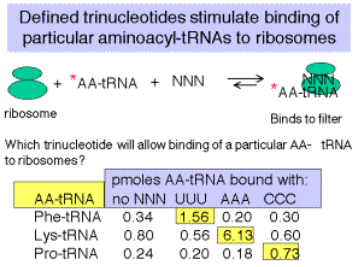
5. Defined trinucleotide codons stimulate binding of aminoacyl-tRNAs to ribosomes

a. At high concentrations of Mg^{2+} cations, the normal initation mechanism, requiring f-Met-tRNAf, can be overriden, and defined trinucleotides can be used to direct binding of particular, labeled aminoacyl-tRNAs to ribosomes.

b. E.g. If ribosomes are mixed with UUU and radiolabeled Phe-tRNAphe, under these conditions, a ternary complex will be formed that will stick to nitrocellulose ("Millipore assay" named after the manufacturer of the nitrocellulose).

c. One can then test all possible combinations of triplet nucleotides.

Figure 3.4.2. Data from Nirenberg and Leder (1964) Science 145:1399.



6. Repeating sequence synthetic polynucleotides (Khorana)

a. Alternating copolymers: e.g. (UC)n programs the incorporation of Ser and Leu. So UCU and CUC encode Ser and Leu, but cannot tell which is which. But in combination with other data, e.g. the random mixed copolymers in section 4 above, one can make some definitive determinations. Such subsequent work showed that UCU encodes Ser and CUC encodes Leu.

b. poly(AUG) programs incorporation of poly-Met and poly-Asp at high Mg concentrations. AUG encodes Met, UGA is a stop, so GUA must encode Asp.

C. The genetic code

By compiling observations from experiments such as those outlined in the previous section, the coding capacity of each group of 3 nucleotides was determined. This is referred to as the **genetic code**. It is summarized in Table 3.4.4. This tells us **how the cell translates from the "language" of nucleic acids** (polymers of nucleotides) **to that of proteins** (polymers of amino acids).

Table 3.4.4. The Genetic Code

	Position in Codon											
1st	2nd .											3rd
	U .		C .		A .		G .					

U	UUU	Phe		UCU	Ser		UAU	Tyr		UGU	Cys		U
	UUC	Phe		UCC	Ser		UAC	Tyr		UGC	Cys		C
	UUA	Leu		UCA	Ser		UAA	Term		UGA	Term		A
	UUG	Leu		UCG	Ser		UAG	Term		UGG	Trp		G
C	CUU	Leu		CCU	Pro		CAU	His		CGU	Arg		U
	CUC	Leu		CCC	Pro		CAC	His		CGC	Arg		C
	CUA	Leu		CCA	Pro		CAA	Gln		CGA	Arg		A
	CUG	Leu		CCG	Pro		CAG	Gln		CGG	Arg		G
A	AUU	Ile		ACU	Thr		AAU	Asn		AGU	Ser		U
	AUC	Ile		ACC	Thr		AAC	Asn		AGC	Ser		C
	AUA	Ile		ACA	Thr		AAA	Lys		AGA	Arg		A
	AUG *	Met		ACG	Thr		AAG	Lys		AGG	Arg		G
G	GUU	Val		GCU	Ala		GAU	Asp		GGU	Gly		U
	GUC	Val		GCC	Ala		GAC	Asp		GGC	Gly		C
	GUA	Val		GCA	Ala		GAA	Glu		GGA	Gly		A
	GUG *	Val		GCG	Ala		GAG	Glu		GGG	Gly		G

* Sometimes used as initiator codons.

2. Of the total of 64 codons, 61 encode amino acids and 3 specify termination of translation.

3. Degeneracy

The **degeneracy** of the genetic code refers to the fact that most amino acids are specified by more than one codon. The exceptions are methionine (AUG) and tryptophan (UGG). The degeneracy is found primarily the third position. Consequently, single nucleotide substitutions at the third position may not lead to a change in the amino acid encoded. These are called **silent** or **synonymous** nucleotide substitutions and do not alter the encoded protein. This is discussed in more detail below.

The pattern of degeneracy allows one to organize the codons into "**families**" and "**pairs**". In 9 groups of codons, the nucleotides at the first two positions are *sufficient* to specify a unique amino acid, and any nucleotide (abbreviated N) at the third position encodes that same amino acid. These comprise 9 codon "families". An example is ACN encoding threonine.

There are 13 codon "pairs", in which the nucleotides at the first two positions are sufficient to specify two amino acids. A purine (R) nucleotide at the third position specifies one amino acid, whereas a pyrimidine (Y) nucleotide at the third position specifies the other amino acid.

These examples add to more than 20 (the number of amino acids) because leucine (encoded by UUR and CUN), serine (encoded by UCN and AGY) and arginine (encoded by CGN and AGR) are encoded by both a codon family and a codon pair. The UAR codons specifying termination of translation were counted as a codon pair. The three codons encoding isoleucine (AUU, AUC and AUA) are half-way between a codon family and a codon pair.

4. Chemically similar amino acids often have similar codons.

5. The major codon specifying initiation of translation is AUG

Bacteria can also use GUG or UUG, and very rarely AUU and possibly CUG. Using data from the 4288 genes identified by the complete genome sequence of *E. coli*, the following frequency of use of codons in initiation was determined:

- AUG is used for 3542 genes.

- GUG is used for 612 genes.
- UUG is used for 130 genes.
- AUU is used for 1 gene.
- CUG may be used for 1 gene.

Regardless of which codon is used for initiation, the first amino acid incorporated during translation is f-Met in bacteria.

6. Three codons specify termination of translation: UAA, UAG, UGA.

Of these three codons, UAA is used most frequently in *E. coli*, followed by UGA. UAG is used much less frequently.

- UAA is used for 2705 genes.
- UGA is used for 1257 genes.
- UAG is used for 326 genes.

7. The genetic code is almost universal

In the rare exceptions to this rule, the differences from the genetic code are fairly small. For example, one exception is RNA from mitochondrial DNA, where both UGG and UGA encode Trp.

D. Differential codon usage

1. Various species have different patterns of codon usage: E.g. one may use 5' UUA to encode Leu 90% of the time (determined by nucleotide sequences of many genes). It may never use CUR, and the combination of UUG plus CUY may account for 10% of the codons.

2. tRNA abundance correlates with codon usage in natural mRNAs: In this example, the tRNALeu with 3' AAU at the anticodon will be the most abundant.

3. The pattern of codon usage may be a *predictor*of the level of expression of the gene: In general, more highly expressed genes tend to use codons that are frequently used in genes in the rest of the genome. This has been quantitated as a "codon adaptation index". Thus in analyzing complete genomes, a previously unknown gene whose codon usage profile matches the preferred codon usage for the organism would score high on the codon adaptation index, and one would propose that it is a highly expressed gene. Likewise, one with a low score on the index may encode a low abundance protein.

The observation of a gene with a pattern of codon usage that differs substantially from that of the rest of the genome indicates that this gene may have entered the genome by horizontal transfer from a different species.

4. The preferred codon usage is a useful consideration in "reverse genetics": If you know even a partial amino acid sequence for a protein and want to isolate the gene for it, the family of mRNA sequences that can encode this amino acid sequence can be determined easily. Because of the degeneracy in the code, this family of sequences can be very large. Since one will likely use these sequences as hybridization probes or as PCR primers, the larger the family of possible sequences is, the more likely that one can get hybridization to a target sequence that differs from the desired one. Thus one wants to limit the number of possible sequences, and by referring to a table of codon preferences (assuming they are known for the organism of interest), then one can use the preferred codons rather than all possible codons. This limits the number of sequences that one needs to make as hybridization probes or primers.

E. Wobble in the anticodon

This flexibility at the "wobble" position allows some tRNAs to pair with two or three codons, thereby reducing the number of tRNAs required for translation. The following “wobble” rules mean that the 61 codons (for 20 amino acids) can be read by as few as 31 anticodons (or 31 tRNAs).

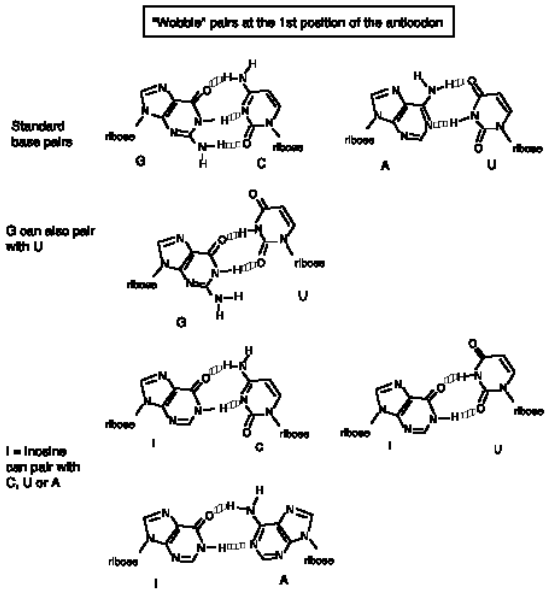
In addition to the usual base pairs, **one can have G-U pairs and I in the anticodon 1st position can pair with U, C or A** (wobble rules).

5' base of the anticodon =	3' base of the codon =
first position in the tRNA	third position in the mRNA
C	G
A	U

U A or G

G C or U

I U, C or A



F. Types of mutations

Base substitutions

This has already been covered in Part Two, DNA Repair. Just as a reminder, there are two types of base substitutions.

1. **Transitions**: A purine substitutes for a purine or a pyrimidine substitutes for another pyrimidine. The same class of nucleotide remains. Examples are A substituting for G or C substituting for T.
2. **Transversions**: A purine substitutes for a pyrimidine or a pyrimidine substitutes for a purine. A different class of nucleotide is placed into the DNA, and the helix will be distorted (especially with a purine-purine base pair). Examples are A substituting for T or C, or C substituting for A or G.

Over evolutionary time, the rate of accumulation of transitions exceeds the rate of accumulation of transversions.

Effect of mutations on the mRNA

1. **Missense mutations** cause the replacement of an amino acid. Depending on the particular replacement, it *may or may not* have a detectable phenotypic consequence. Some replacements, e.g. a valine for an leucine in a position that is important for maintaining an a-helix, may not cause a detectable change in the structure or function of the protein. Other replacements, such as valine for a glutamate at a site that causes hemoglobin to polymerize in the deoxygenated state, cause significant pathology (sickle cell anemia in this example).
2. **Nonsense mutations** cause premature termination of translation. They occur when a substitution, insertion or deletion generates a stop codon in the mRNA within the region that encodes the polypeptide in the wild-type mRNA. They *almost always* have serious phenotypic consequences.
3. **Frameshift mutations** are insertions or deletions that change the reading frame of the mRNA. They *almost always* have serious phenotypic consequences.

Not all base subsitutions alter the encoded amino acids

1. The base substitution may lead to an alteration in the encoded polypeptide sequence, in which case the substitution is called **nonsynonymous** or **nonsilent**.
2. If the base substitution occurs in a degenerate site in the codon, so that the encoded amino acid is not altered, it is called a synonymous or silent substitution.

Example:

ACU -> AAU is a nonsynonymous substitution that results in Thr → Asn

while,

ACU -> ACC is a synonymous substitution that results in Thr → Thr

3. Examination of the patterns of degeneracy in the genetic code shows that nonsynonymous substitutions occur mostly in the first and second positions of the codon, whereas synonymous substitutions occur mostly in the third position. However, there are several exceptions to this rule.
4. In general, the rate of fixation of synonymous substitutions in a population is significantly greater that the rate of fixation of nonsynonymous substitutions. This is one of the strongest supporting arguments in favor of model of neutral evolution, or evolutionary drift, as a principle cause of the substitutions seen in natural populations.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

13.E: Genetic Code (Exercises)

Questions for Chapter 13. Genetic Code

13.1 How does the enzyme polynucleotide phosphorylase differ from DNA and RNA polymerases?

13.2 A short oligopeptide is encoded in this sequence of RNA

5' GACUAUGCUCAUAUUGGUCCUUUGACAAG

a) Where does it start and stop, and how many amino acids are encoded?

b) Which codon position usually shows degeneracy?

The template strand of a sample of double-helical DNA contains the sequence:

(5')CTTAACACCCCTGACTTCGCGCCGTCG

a) What is the base sequence of mRNA that can be transcribed from this strand?

c) Suppose the other (nontemplate) strand of this DNA sample is transcribed and translated. Will the resulting amino acid sequence be the same as in (b)? Explain the biological significance of your answer.

13.5 The Basis of the Sickle-Cell Mutation.

b) Leu can be converted to either Ser, Val, or Met by a single nucleotide substitution (a different nucleotide substitution for each amino acid replacement). What is the codon for Leu?

b) valine?

b) 5'-G-A-U-3'

13.10 (POB) Identifying the Gene for a Protein with a Known Amino Acid Sequence.

H_3N^+-Ala-Pro-Met-Thr-Trp-Tyr-Cys-Met-Asp-Trp-Ile-Ala-Gly-Gly-Pro-Trp-Phe-Arg-Lys-Asn-Thr-Lys---

13.11 Let's suppose you are in a lab on the Starship Enterprise. One of the “away teams” has visited Planet Claire and brought back a fungus that is the star of this week's episode. While the rest of the crew tries to figure out if the fungus is friend or foe (and gets all the camera time), you are assigned to determine its genetic code. With the technologies of two centuries from now, you immediately discover that its proteins are composed of only eight amino acids, which we will call simply amino acids 1, 2, 3, 4, 5, 6, 7, and 8. Its genetic material is a nucleic acid containing only three nucleotides, called K, N and D, which are not found in earthly nucleic acids.

The results of frameshift mutations confirm your suspicion that the smallest possible coding unit is in fact used in this fungus. Insertions of a single nucleotide or three nucleotides into a gene cause a complete loss of function, but insertions or deletions of two nucleotides have little effect on the encoded protein.

You make synthetic polymers of the nucleotides K, N and D and use them to program protein synthesis. The amino acids incorporated into protein directed by each of the polynucleotide templates is shown below. Assume that the templates are read from left to right.

Template Amino acid(s) incorporated

K_n = KKKKKKKKKK 1

N_n = NNNNNNNNN 2

D_n = DDDDDDDDDD 3

$(KN)_n$ = KNKNKNKNKN 4 and 5

$(KD)_n$ = KDKDKDKDKD 6 and 7

$(ND)_n$ = NDNDNDNDND 8

$(KND)_n$ = KNDKNDKNDKND 4 and 6 and 8

Please report your results on the genetic code used in the fungus from Planet Claire.

a) What is size of a codon?

b) Is the code degenerate?

Amino acid	Codon(s)
1	
3	
5	
7	

e) What is the mutation that will change a codon for amino acid 6 to a codon for amino acid 5? Show both the initial codon and the mutated codon.

f) What is the mutation that will change a codon for amino acid 8 to a codon for amino acid 7? Show both the initial codon and the mutated codon.

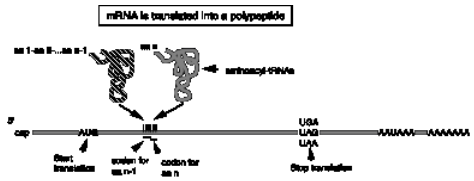
14: Translation (Protein synthesis)

A reminder: mRNA encodes the polypeptide with each amino acid designated by a string of three nucleotides. tRNAs serve as the adaptors to translate from the language of nucleic acids to that of proteins. Ribosomes are the factories for protein synthesis.

A. tRNAs

1. The transfer RNAs, or tRNAs serve as adaptors to align the appropriate amino acids on the mRNA templates.

Figure 3.5.1.



2. **Primary structure of tRNAs**

   a. tRNAs are short, being only 73 to 93 nts long.

   b. All tRNAs have the trinucleotide CCA at the 3' end.

      1. The amino acid is attached to the terminal A of the CCA.
      2. In most prokaryotic tRNA genes, the CCA is encoded at the 3' end of the gene. No known eukaryotic tRNA gene encodes the CCA, but rather it is added posttranscriptionally by the enzyme tRNA nucleotidyl transferase.

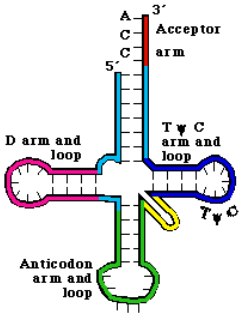


Figure 3.5.2. Secondary structure of tRNA.

3. **The secondary structure of tRNA is a cloverleaf**

   a. tRNAs have 4 arms with 3 loops (see Figure 3. 5.2. for yeast phenylalanine tRNA)

   b. Theamino acid acceptor arm is formed by complementary base-pairing between the intial 7 nts of tRNA and a short segment near the 3' end. Again, the amino acid will be added to the terminal A.

   c. The D arm ends in the D loop. It contains several dihydrouridines, which are abbreviated "D".

   d. The anticodon arm ends in anticodon loop. The anticodon is located in the center of the loop. It will align 3' to 5' with the mRNA (reading 5' to 3').

   e. The variable loop varies in size in different tRNAs. The difference in size between the 73 nt versus 93 nt tRNAs is found in the variable loop.

   f. The TyC arm is named for this highly conserved motif found in the loop.

4. **The tertiary structure of tRNA is a "fat L". (**See Fig 3.5.3.)

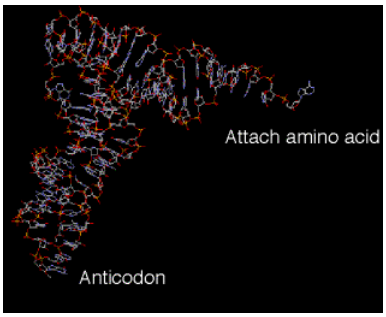


Figure 3.5.9.

F. The polarity of translation is from the amino (N) terminus to the caboxy (C) terminus.

This was demonstrated in a classic experiment by Dintzis.

1. Actively translating proteins were labeled with radioactive amino acids for a brief time (short relative to the time required to complete synthesis).
2. Completed polypeptides were collected, digested with trypsin, and the amount of radioactivity in tryptic fragments was determined.
3. Tryptic fragments from the C-terminal end of the polypeptide had radioactivity at the earliest times of labeling.
4. As the period of labeling was increased (longer pulse), tryptic fragments closer to the N terminus were labeled.
5. This shows that the direction of polypeptide growth is from the N teminus to the C terminus, i.e. translation begins at the N terminal amino acid. This corresponds to mRNA chain growth in a 5' to 3' direction.
6. Note that this experimental protocol is also used to map origins of replication, as we covered in Part Two of the course.

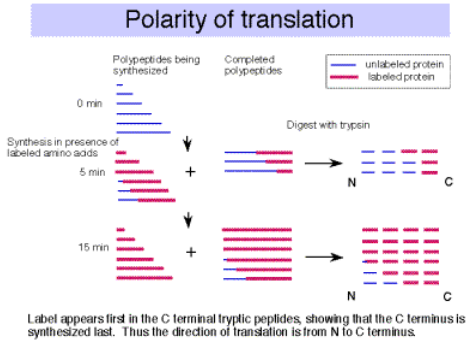


Figure 3.5.12.

b. **IF3**= Initiation Factor 3

1. An antiassociation factor; prevents association between the large and small ribosomal subunits.
2. It also must be associated with the small subunit for it to form an initiation complex, i.e. for the small subunit to correctly bind mRNA and fmet-tRNAf.
3. It dissociates prior to binding of the large subunit.

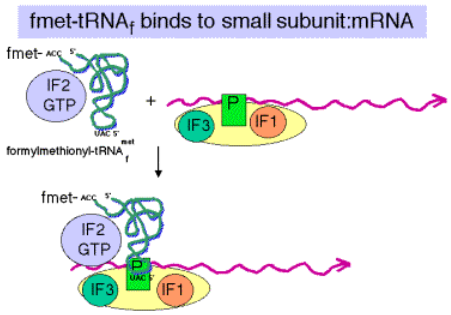


Figure 3.5.13.

c. **IF2**

1. Brings **fmet-tRNAf** to the partial P site on the small subunit.
2. At least in eukaryotes, it does this in a ternary complex with IF2, fmet-tRNAf and GTP. In bacteria, the GTP may bind the initiation complex separately. [In some texts, such as MBOG, p. 412, the GTP-IF2 complex binds to the 30S subunit separately from fmet-tRNAf. How would you test the differences in these two models?]
3. IF2 activates a GTPase activity in the small subunit. The resulting change in conformation may allow the large subunit to bind.

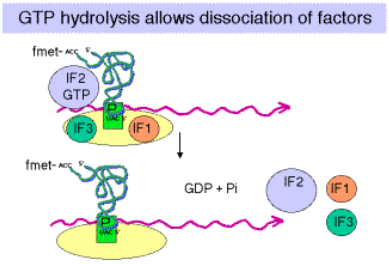


Figure 3.5.15.

3. **Binding of 50S (large) subunit to initiation complex gives a complete ribosome ready for the elongation phase of translation**. Note that f-met-tRNAfmet is positioned at the P site. It has recognized the initiator AUG in the mRNA.

4. **Identification of initiator AUG in eukaryotes**

a. **Bases around AUG influence efficiency of initiation.**

1. The most important effects are from a purine 3 nt before AUG and a G after it. The preferred context is **R**NNAUG**G**.
2. The consensus sequence for a large number of mRNAs is GCC**R**CCAUG**G**, but these other nucleotides have little effect in mutagenesis experiments.

a. **Modified scanner model**

(1) The mRNA is "prepared" for binding to the ribosome by the action of eukaryotic initiation factor 4, abbreviated eIF4 (Figure 3.5.16). eIF4 is a multisubunit factor; it includes a cap-binding protein, eIF4F, that recognizes the 5' cap structure. It also includes proteins eIF4A and eIF4B. These are RNA helicases, which unwind secondary structures in the 5' untranslated region of the mRNA at the expense of ATP hydrolysis.

The mRNA then binds to the small ribosomal subunit. The met-tRNAi has already been brought to the small ribosomal subunit by eIF2, in a complex with GTP.

eIF3 keeps the small ribosomal subunit apart from the large subunit during the process of binding the mRNA.

(2) The small subunit, with associated factors, scans along the mRNA until it reaches (usually) the first AUG. Factors eIF1 and eIF1A help move the preinitiation complex to the AUG start.

Figure 3.5.16.

H. The elongation cycle during translation

1. **Binding of aminoacyl-tRNA to the A site**

Recent review: Weijland, A. and A. Parmeggiani (1994) TIBS 19:188-193. Schroeder, R. (1994) Nature 370:597.

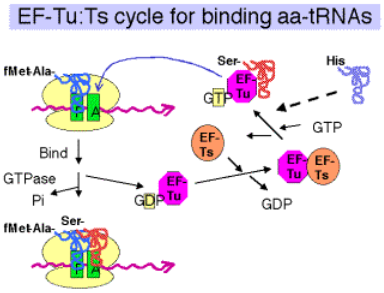
a. Elongation factor **EF-Tu**

1. The ternary complex of aminoacyl-tRNA, EF-Tu, and GTP brings the aminoacyl-tRNA to the A site on the 70S ribosome (fig. 3.5.17).
2. After the aminoacyl-tRNA is deposited at the A site of the ribosome, the GTP is cleaved to GDP + Pi. The binary complex of EF-Tu and GDP dissociates from the ribosome.
3. This is one of many examples of guanine-nucleotide-binding proteins that are active when GTP is bound and inactive when GDP is bound.

The general model is that the GTP-bound state of EF-Tu adopts a conformation with a high affinity for aminoacyl-tRNA. The conformation (shape, charge density, etc.) of the resulting ternary complex (containing EF-Tu,GTP, and aminoacyl-tRNA) then

allows it to bind to the A site of the ribsosome. **Hydrolysis of GTP to form GDP and inorganic phosphate causes the EF-Tu to adopt a different conformation**. The aminoacyl-tRNA now has a lower affinity for EF-Tu in the GDP bound state, and presumably a higher affinity for the A site on the ribosome, so it stays on the ribosome when EF-Tu in the GDP bound state dissociates (both from aminoacyl-tRNA and from the ribosome).

Figure 3.5.17.



(4) EF-Tu is one of the most abundant proteins in E. coli, at 70,000 copies per cell. This is almost equal to the number of aminoacyl-tRNAs per cell, so most of the aminoacyl-tRNAs are likely to be in the ternary complex when the concentration of GTP is sufficiently high.

b. **GTP**

1. Required for binding aminoacyl-tRNA.
2. Hydrolysis promotes dissociation of the complex EF-Tu plus GDP from the ribosome.

c. **EF-Ts**

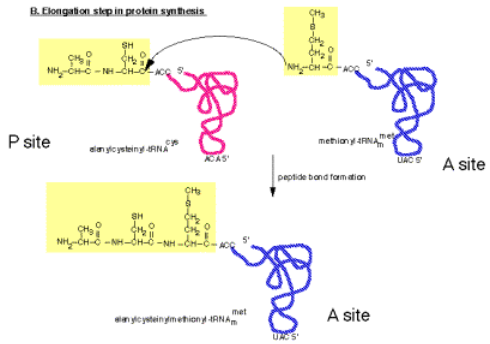
1. Aids in the recycling of EF-Tu by GDP-GTP exchange.
2. EF-Ts binds to EF-Tu complexed with GDP, causing dissociation of GDP. GTP can now bind to the EF-Tu-Ts complex, causing EF-Ts to dissociate and leaving EF-Tu complexed with GTP. This latter binary complex is ready to bind another aminoacyl-tRNA.

d. The antibiotic kirromycin prevents release of EF-Tu-GDP, thereby blocking elongation. This demonstrates that one step must be completed before the next can take place, and illustrates the importance of the EF-Tu-GTP/GDP cycle.

2. **Peptidyl transferase** on the large ribosomal subunit

a. The peptidyl transferase reaction occurs via **nucleophilic displacement**. The amino group from aminoacyl-tRNA (position *n*) attacks the "C-terminal" carboxyl group of peptidyl-tRNA (position *n-1*in the mRNA). This results in cleavage of the high energy peptidyl-tRNA ester linkage, thereby providing the free energy to drive the reaction. The resulting products of the reaction are deacylated tRNA at the P site and peptidyl-tRNA at the A site.

Figure 3.5.18. Peptidyl transferase reaction



b. **Role of rRNA in catalysis**

It is likely that rRNA provides the catalytic center for the peptidyl transferase activity, with perhaps some ribosomal proteins aiding in holding the rRNA in the correct conformation for catalysis. This conclusion is supported by several lines of investigation, some

of which are listed below.

1. No protein, singly or in combination with other proteins, has been shown to catalyze peptide bond formation.
2. Specific regions of 16S rRNA (in the small subunit) interact with the anticodon regions of tRNA in both the A and P sites. In contrast, 23S rRNA in the large subunit interacts with the CCA terminus of peptidyl-tRNA, thus placing it in the right location for peptidyl transferase.
3. The antibiotics erythromycin and chloramphenicol block peptidyl transferase. Some mutations that confer resistance to them map to the 23S rRNA sequence (others map to some 50S ribosomal proteins).
4. A preparation consisting of 23S rRNA and some remnants of large subunit proteins retains peptidyl transferase activity. For more information, see Noller et al. (1992) Unusual resistance of peptidyl transferase to protein extraction procedures. Science 256: 1416-1419.
5. Ribozyme RNAs can be selected that catalyze peptide bond formation. In this experiment, the investigators started with a pool of 1.3 ´ 1015 different RNAs of 72 nucleotides, flanked by constant regions. They let this large population of RNAs catalyze a peptide bond formation that adds a biotinyl-labeled amino acid (in a chemical mimic of a P site) to an amino acid connected to the RNA (in a chemical mimic of an A site). The RNAs that successfully catalyzed the reaction were extremely rare, but were now covalently attached to a biotin label. Thus they could be selected from the population by binding to streptavidin. PCR was used to amplify the successful RNAs, and the procedure repeated 19 times. At this point, the investigators characterized 9 RNAs that catalyzed the reaction. They found that these RNAs increased the reaction rate by a factor of 106 over the uncatalyzed reaction.
6. The three-dimensional structure of the ribosome shows that the active site is comprised of RNA. The structure of a ribosome crystallized with an active site directed inhibitor has been determined, as well as the structure without the inhibitor. This allowed researchers to see precisely where the peptidyl transferase active site is within the structure. Only RNA is seen around this site. The nearest protein is 20 Angstroms away, too far to participate in catalysis.

3. **Translocation**

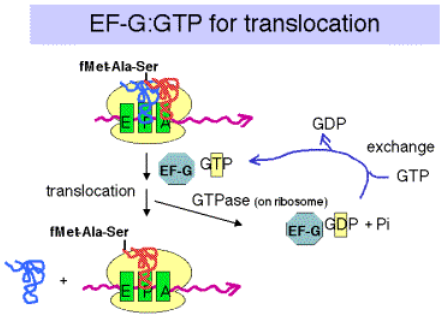
a. The translocation step moves the ribosome another 3 nucleotides downstream (one codon) and moves peptidyl-tRNA to the P site (position *n*), deacylated tRNA exits through the E site, and the A site (position *n+1*) is vacant for another round of elongation.

b. Elongation Factor G = **EF-G**

1. This is another very abundant protein, with about 20,000 copies per cell, which is equivalent to the number of ribosomes.
2. EF-G-GTP binds to the ribosome to aid translocation, and is released upon GTP hydrolysis (GTPase is from some ribosomal component).
3. Recent structural studies (from A. Dahlberg and colleagues) show that EF-G in the GTP-bound state has a shape similar to that of the ternary complex of EF-Tu, GTP and aminoacyl-tRNA. Like the latter ternary complex, EF-G in the GTP-bound state also has a high affinity for the A site on the ribosome. This may help drive the movement of the peptidyl-tRNA from the A site to the P site, replacing it with EF-G (GTP) in the A site.

c. Hydrolysis of GTP is required for dissociation of EF-G after translocation. The GTPase is part of the ribosome, not EF-G.

Figure 3.5.20.



d. Action of fusidic acid revealed the need for release of EF-G-GDP. In the presence of fusidic acid, EF-G-GTP binds the ribosome, GTP is hydrolyzed, and the ribosome moves three nucleotides. But the ribosome-EF-G-GDP complex is stabilized by this compound, and translation is halted.

e. Ribosomes cannot bind EF-Tu and EF-G simultaneously. EF-Tu must finish its action before EF-G can act, and EF-G must complete its cycle before EF-Tu can act again to bring in another aminoacyl-tRNA.

f. Effect of diptheria toxin

1. The eukaryotic analog to EF-G is eEF2, which is also a translocase dependent on GTP hydrolysis. It is also is blocked by fusidic acid.
2. Diptheria toxin will catalyze the addition of ADP-ribose (from substrate NAD+) to eEF2, thereby inactivating it. The target for ADP-ribosylation is modified histidine found in eEF2 from many species.

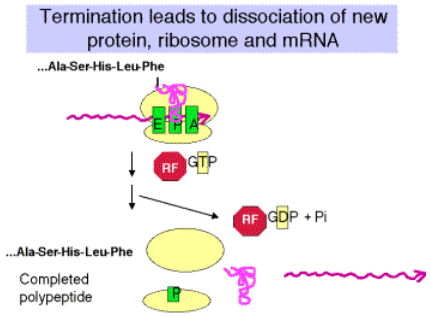
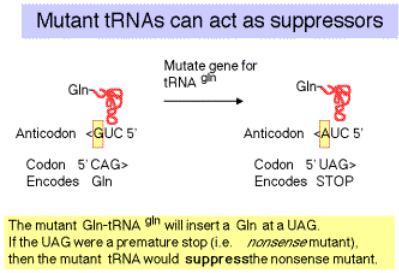


Figure 3.5.21.



c. The "down side" to nonsense suppression is that the suppressor tRNA can act at any amber codon, Therefore it competes with the releasing factors in recognizing the normal termination codons. When the suppressor tRNA is used instead of releasing factors, translation proceeds further down the mRNA than it is supposed to, leading to production of aberrant proteins. Suppressor strains of *E. coli*can be pretty sick (i.e. they don't grow as well as wild type strains).

d. Two other amber suppressors are encoded by the *supD*gene, which encodes a tRNA that will insert Ser at a UAG, and *supF*, which will insert Tyr.

3. Missense suppressors: These are mutant tRNAs that lead to the insertion of an amino acid that is compatible with the wild type amino acid (altered by the original mutation).

4. Frameshift suppressors: These are mutant tRNAs whose anticodon has been expanded (or contracted?) to match the length-altering mutation in the mRNA.

E.g. Consider an original mutation 5'GGG -> 5'GGGG (insert a G).

A frameshift suppressor would also have an additional C in the anticodon.

wt tRNA anticodon 3'CCC --> suppressor tRNA 3'CCCC.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

14.E: Translation - Protein synthesis (Exercises)

14.1 (POB) Methionine Has Only One Codon.

> Methionine is one of the two amino acids having only one codon. Yet the single codon for methionine can specify both the initiating residue and interior Met residues of polypeptides synthesized by *E. coli*. Explain exactly how this is possible.

14.2 Are the following statements concerning aminoacyl-tRNA synthetase true or false?

> a) Two distinct classes of the enzymes have been defined that are not very related to each other.
>
> b) The enzymes scan previously-synthesized aminoacyl-tRNAs and cleave off any amino acids that are linked to the incorrect tRNA.
>
> c) Proofreading can occur at the formation of either the aminoacyl-adenylate intermediate (in some synthetases) or at the aminoacyl-tRNA (in other synthetases) to insure that the correct amino acid is attached to a given tRNA.
>
> d) The product of the reaction has a high-energy ester bond between the carboxyl of an amino acid and a hydroxyl on the terminal ribose of the tRNA.

14.3 A preparation of ribosomes in the process of synthesizing the polypeptide insulin was incubated in the presence of all 20 radiolabeled amino acids, tRNA's, aminoacyl-tRNA synthetases and other components required for protein synthesis. All the amino acids have the same specific radioactivity (counts per minute per nanomole of amino acid). It takes ten minutes to synthesize a complete insulin chain (from initiation to termination) in this system. After incubation for 1 minute, the completed insulin chains were cleaved with trypsin and the radioactivity of the fragments determined.

> a) Which tryptic fragment has the highest specific activity?
>
> b) In the same system described above, the insulin polypeptide chains still attached to the ribosomes after ten minutes were isolated, cleaved with trypsin, and the specific activity of each tryptic peptide determined. Which peptide has the highest specific activity?

14.4 Which component of the protein synthesis machinery of *E. coli*carries out the function listed for each statement.

> a) Translocation of the peptidyl-tRNA from the A site to the P site of the ribosome.
>
> b) Binding of f-Met-tRNA to the mRNA on the small ribosomal subunit.
>
> c) Recognition of the termination codons UAG and UAA.
>
> d) Holds the initiator AUG in register for formation of the initiation complex (via base pairing).

14.5 a) In the initiation of translation in *E. coli*, which ribosomal subunit does the mRNA initially bind to?

> b) What nucleotide sequences in the mRNA are required to direct the mRNA to the initial binding site on the ribosome?
>
> c) What other factors are required to form an initiation complex?

14.6 What steps in the activation of amino acids and elongation of a polypeptide chain require hydrolysis of high energy phosphate bonds? What enzymes catalyze these steps or which protein factors are required?

14.7(POB) Maintaining the Fidelity of Protein Synthesis

> The chemical mechanisms used to avoid errors in protein synthesis are different from those used during DNA replication. DNA polymerases utilize a 3' ® 5' exonuclease proofreading activity to remove mispaired nucleotides incorrectly inserted into a growing DNA strand. There is no analogous proofreading function on ribosomes; and, in fact, the identity of amino acids attached to incoming tRNAs and added to the growing polypeptide is never checked. A proofreading step that hydrolyzed the last peptide bond formed when an incorrect amino acid was inserted into a growing polypeptide (analogous to the proofreading step of DNA polymerases) would actually be chemically impractical. Why? (Hint: Consider how the link between the growing polypeptide and the mRNA is maintained during the elongation phase of protein synthesis.)

14.8 (POB) Expressing a Cloned Gene.

You have isolated a plant gene that encodes a protein in which you are interested. What are the sequences or sites that you will need to get this gene transcribed, translated, and regulated in *E. coli*.)?

14.9 The three codons AUU, AUC, and AUA encode isoleucine. They correspond to "hybrid" between a codon family (4 codons) and a codon pair (2 codons). The single codon AUG encodes methionine. Given the prevalence of codon pairs and families for other amino acids, what are hypotheses for how this situation for isoleucine and methionine could have evolved?

14.10 Use the following processes to answer parts a-c:

1. synthesis of aminoacyl-tRNA from an amino acid and tRNA.
2. binding of aminoacyl-tRNA to the ribosome for elongation.
3. formation of the peptide bond between peptidyl-tRNA and aminoacyl-tRNA on the ribosome.
4. translocation of peptidyl-tRNA from the A site to the P site on the ribosome.
5. assembly of a spliceosome for removal of introns from nuclear pre-mRNA.
6. removal of introns from nuclear pre-mRNA after assembly of a spliceosome.
7. synthesis of a 5' cap on eukaryotic mRNA.

(a) Which of the above processes require ATP?

(b) Which of the above processes require GTP?

(c) For which of the above processes is there evidence that RNA is used as a catalyst?

CHAPTER OVERVIEW

Unit IV: Regulation of Gene Expression

Regulation is the controlled expression of gene functions. This can be done in many ways, but these can be grouped into two classes. The level of enzyme activity can be regulated by noncovalent or covalent modification of a protein. The amount of the protein can also be regulated. This latter class of regulation can be exerted at any step in the pathway of gene expression or during protein turnover. For many (perhaps most) genes, the principal level of regulation of expression is at transcription, and Part Four of this course will focus primarily on this. However, post-transcriptional control is also important in many genes, and this will also be discussed.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

15: Positive and negative control of gene expression

Operons

An **operon** is a cluster of coordinately regulated genes. It includes **structural genes** (generally encoding enzymes), **regulatory genes** (encoding, e.g. activators or repressors) and **regulatory sites** (such as promoters and operators). The type of control is defined by the response of the operon when no regulatory protein is present. In the case of negative control, the genes in the operon are expressed unless they are switched off by a repressor protein. Thus the operon will be turned on constitutively (the genes will be expressed) when the repressor in inactivated. In the case of positive control, the genes are expressed only when an active regulator protein, e.g. an activator, is present. Thus the operon will be turned off when the positive regulatory protein is absent or inactivated.

	Regulatory protein is present	Example of regulatory protein	Mutate regulatory gene to lose function
Positive control	Operon ON	Activator	Operon OFF
Negative control	Operon OFF	Repressor	Operon ON

Table 4.1.1. *Positive vs. negative control*

Catabolic versus Biosynthetic Operons

Catabolic pathways catalyze the breakdown of nutrients (the substrate for the pathway) to generate energy, or more precisely ATP, the energy currency of the cell. In the absence of the *substrate*,there is no reason for the catabolic enzymes to be present, and the operon encoding them is repressed. In the presence of the substrate, when the enzymes are needed, the operon is induced or de-repressed.

Table 4.1.2. Comparison of catabolic and biosynthetic operons

Operon encodes	Absence of	Effect	Presence of	Effect
catabolic enzymes	substrate	repressed	substrate	derepressed (induced)
biosynthetic enzymes	product	induced	product	repressed

For example, the lac operon encodes the enzymes needed for the uptake (lactose permease) and initial breakdown of lactose (the disaccharide b-D-galactosyl-1->4-D-glucose) into galactose and glucose (catalyzed by b-galactosidase). These monosaccharides are broken down to lactate (principally via glycolysis, producing ATP), and from lactate to CO2 (via the citric acid cycle), producing NADH, which feeds into the electron-transport chain to produce more ATP (oxidative phosphorylation). This can provide the energy for the bacterial cell to live. However, the initial enzymes (lactose permease and b-galactosidase) are only needed, and only expressed, in the presence of lactose and in the absence of glucose. In the presence of the substrate lactose, the operon in turned on, and in its absence, the operon is turned off.

Anabolic, or biosynthetic, pathways use energy in the form of ATP and reducing equivalents in the form of NAD(P)H to catalyze the synthesis of cellular components (the product) from simpler materials, e.g. synthesis of amino acids from small dicarboxylic acids (components of the the citric acid cycle). If the cell has plenty of the product already (in the presence of the *product*), the the enzymes catalyzing its synthesis are not needed, and the operon encoding them is repressed. In the absence of the product, when the cell needs to make more, the biosynthetic operon is induced. E.g., the *trp*operon encodes the enzymes that catalyze the conversion of chorismic acid to tryptophan. When the cellular concentration of Trp (or Trp-tRNAtrp) is high, the operon is not expressed, but when the levels are low, the operon is expressed.

Inducible versus repressible Operons

Inducible operons are turned on in reponse to a metabolite (a small molecule undergoing metabolism) that regulates the operon. E.g. the lac operon is induced in the presence of lactose (through the action of a metabolic by-product allolactose). Repressible operons are switched off in reponse to a small regulatory molecule. E.g., the *trp*operon is repressed in the presence of tryptophan. Note that in this usage, the terms are defined by the reponse to a small molecule. Although lac is an inducible operon, we will see conditions under which it is repressed or induced (via derepression).

Inducible vs. repressible operons

Defined by response of operon to a metabolite (small molecule).

Type of operon	Presence of	Effect	Examples: Metabolite	Examples: Operon
Inducible	metabolite	ON	lactose	*lac*
Repressible	metabolite	OFF	Trp	*trp*

Table 4.1.3.

Map of the E. colilac operon

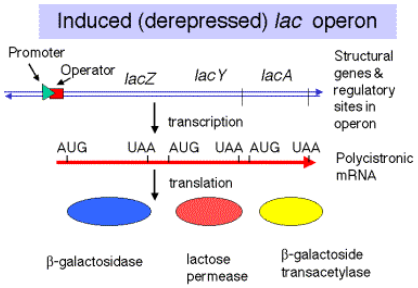


Figure 4.1.1.

1. Promoters = *p*= binding sites for RNA polymerase from which it initiates transcription. There are separate promoters for the *lacI*gene and the *lacZYA*genes.
2. Operator = *o* = binding site for repressor; overlaps with the promoter for *lacZYA*.
3. Repressor encoded by *lacI*gene
4. Structural genes: *lacZYA*

    lacZ encodes b-galactosidase, which cleaves the disccharide lactose into galactose and glucose.

    *lacY*encodes the lactose permease, a membrane protein that faciltitates uptake of lactose.

    *lacA*encodes b-galactoside transacetylase; the function of this enzymes in catabolism of lactose is not understood (at least by me)

C. Negative control

The *lac* operon is under both negative and positive control. The mechanisms for these will be considered separately.

1. In negative control, the *lacZYA*genes are switched off by repressor when the inducer is absent (signalling an absence of lactose). When the repressor tetramer is bound to *o*, *lacZYA*is not transcribed and hence not expressed.

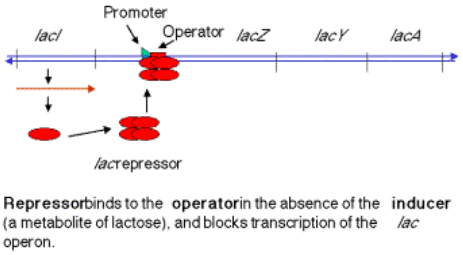


Figure 4.1.2. Repressed ***lac* operon**

2. When inducer is present (signalling the presence of lactose), it binds the repressor protein, thereby altering its conformation, decreasing its affinity for *o*, the operator. The dissociation of the repressor-inducer complex allows *lacZYA*to be transcribed and therefore expressed.

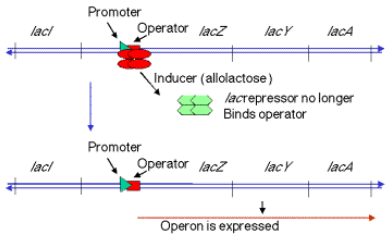


Figure 4.1.3. Induction of the lac operon by derepression.

Inducers

The natural inducer (or antirepressor), is allolactose, an analog of lactose. It is made as a metabolic by-product of the reaction catalyzed by b-galactosidase. Usually this enzyme catalyzes the cleavage of lactose to galactose + glucose, but occasionally it will catalyze an isomerization to form allolactose, in which the galacose is linked to C6 of glucose instead of C4.

A gratuitous inducer will induce the operon but not be metabolized by the encoded enzymes; hence the induction is maintained for a longer time. One of the most common ones used in the laboratory is a synthetic analog of lactose called isopropylthiogalactoside (IPTG). In this compound the b-galactosidic linkage is to a thiol, which is not an efficient substrate for b-galactosidase.

E. Regulatory mutants

Regulatory mutations affect the amount of all the enzymes encoded by an operon, whereas mutations in a structural gene affects only the activity of the encoded (single) polypeptide.

Repressor mutants

- a. Wild-type strains (*lacI+*) are inducible.
- b. Most strains with a defective repressor (*lacI-*) are constitutive, i.e. they make the enzymes encoded by the lac operon even in the absence of the inducer.
- c. Strains with repressor that is not able to interact with the inducer (*lacIS*) are noninducible. Since the inducer cannot bind, the repressor stays on the operator and prevents expression of the operon even in the presence of inducer.
- d. Deductions based on phenotypes of mutants

Table 4.1.4. Phenotypes of repressor mutants

	b-galactosidase		transacetylase		
Genotype	-IPTG	+IPTG	-IPTG	+IPTG	Conclusion
I+Z+A+	<0.1>	100	<1>	100	Inducible
I+Z-A+	<0.1>	<0.1>	<1>	100	*lacZ*encodes b-galactosidase
I-Z+A+	100	100	90	90	Constitutive
I+Z-A+ /F' I-Z+A+	<0.1>	100	<1>	200	*I+ >I-* in *trans*
IsZ+A+	<0.1>	<1>	<1>	<1>	Noninducible
IsZ+A+ /F' I+Z+A+	<0.1>	1	<1>	1	*Is>.I+* in *trans*

1. The wild-type operon is inducible by IPTG.
2. A mutation in *lacZ*affects only b-galactosidase, not the transacetylase (or other products of the operon), showing that *lacZ*is a structural gene.
3. A mutation in *lacI*affects both enzymes, hence *lacI*is a regulatory gene. Both are expressed in the absence of the inducer, hence the operon is constitutively expressed (the strain shows a constitutive phenotype).
4. In a merodiploid strain, in which one copy of the lac operon is on the chromosome and another copy is on an F' factor, one can test for dominance of one allele over another. The wild-type *lacI+*is dominant over *lacI-*in*trans*. In a situation where the only functional *lacZ*gene is on the same chromosome as *lacI-*, the functional *lacI* still causes repression in the absence of inducer.
5. The *lacIS*allele is noninducible.
6. In a merodiploid, the *lacIS*allele is dominant over wild-type in *trans*.

e. The fact that the product of the *lacI*gene is *trans*-acting means that it is a diffusible molecule that can be encoded on one chromosome but act on another, such as the F' chromosome in example (d) above. In fact the product of the *lacI*gene is a repressor protein.

2. Operator mutants

a. Defects in the operator lead to constitutive expression of the operon, hence one can isolate operator constitutive mutations, abbreviated *oc*. The wild-type *o+*is inducible.

b. Mutations in the operator are *cis*-acting; they only affect the expression of structural genes on the same chromosome.

(1)The merodiploid *I+ocZ+/I+o+Z-* [this is an abbreviation for *lacI+oclacZ+/lacI+o+lacZ-*] expresses b-galactosidase constitutively. Thus *oc* is dominant to *o+* when *oc* is in *cis*to *lacZ+*.

(2)The merodiploid *I+ocZ-/I+o+Z+* is inducible for b-galactosidase expression. Thus *o+* is dominant to *oc* when *o+* is in *cis*to *lacZ+*.

(3)The allele of *o*that is in *cis*to the active reporter gene (i.e., on the same chromosome as *lacZ+* in this case) is the one whose phenotype is seen. Thus the operator is *cis*-acting, and this property is referred to as *cis*-dominance. As in most cases of *cis*-regulatory sequences, these are sites on DNA that are required for regulation. In this case the operator is a binding site for the *trans*-acting repressor protein.

Interactions between Operator and Repressor

Sequence of operator

The operator overlaps the start the site of transcription and the promoter. It has a dyad symmetry centered at +11. Such a dyad symmetry is commonly found within binding sites for symmetrical proteins (the repressor is a homotetramer). The sequence of DNA that consititutes the operator was defined by the position of *oC* mutations, as well as the nucleotides protected from reaction with, e.g. DMS, upon binding of the repressor.

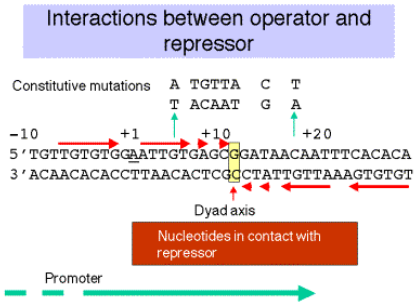


Figure 4.1.4.

2. **Repressor**

a. Purification

(1)Increase the amount of repressor in the starting material by over-expression.

A wild-type cell has only about 10 molecules of the repressor tetramer. Isolation and purification of the protein was greatly aided by use of mutant strain with up-promoter mutations for *lacI*, so that many more copies of the protein were present in each cell. This general strategy of over-producing the protein is widely used in purification schemes. Now the gene for the protein is cloned in an expression vector, so that the host (bacteria in this case) makes a large amount of the protein - often a substantial fraction of the total bacterial protein.

(2)Assays for repressor

[1]Binding of radiolabeled IPTG (gratuitous inducer) to repressor

[2]Binding of radiolabeled operator DNA sequence to repressor. This can be monitored by the ability of the protein-DNA complex to bind to nitrocellulose (whereas a radiolabeled mutant operator DNA fragement, *oc*, plus repressor will not bind). Electrophoretic mobility shift assays would be used now in many cases.

[3]This ability of particular sequences to bind with high affinity to the desired protein is frequently exploited to rapidly isolate the protein. The binding site can be synthesized as duplex oligonucleotides. These are ligated together to form multimers, which are then attached to a solid substrate in a column. The desired DNA-binding protein can then be isolated by affinity chromatography, using the binding site in DNA as the affinity ligand.

b. The isolated, functional repressor is a tetramer; each of the four monomers is the product of the *lacI* gene (i.e. it is a homotetramer).

c. The **DNA-binding domain**of the *lac* repressor folds into a **helix-turn-helix**domain. We will examine this structural domain in more in Chapter III. It is one of the most common DNA-binding domains in prokaryotes, and a similar structural domain (the homeodomain) is found in some eukaryotic transcriptional regulators.

3. **Contact points between repressor and operator**

a. Investigation of the contact points between repressor and the operator utiblized the same techniques that we discussed previously for mapping the binding site of RNA polymerase on the promoter, e.g. electrophoretic mobility shift assays (does the DNA fragment bind?), DNase footprints (where does the protein bind?) and methylation interference assays (methylation of which purines will prevent binding?). Alternative schemes will allow one to identify sites at which methylation is either prevented or enhanced by the binding of the repressor. These techniques provide a biochemical defintion of the operator = binding site for repressor.

b. The key contact points (see Figure 4.1.4.):

(1)are within the dyad symmetry.

(2)coincide (in many cases) with nucleotides that when mutated lead to constitutive expression. Note that the latter is a genetic definition of the operator, and it coincides with the biochemically-defined operator.

(3)tend to be distributed symmetrically around the dyad axis (+11).

(4)are largely on one face of the DNA double helix.

c. The partial overlap between the operator and the promoter initially suggested a model of steric interference to explain the mechanism of repression. As long a repressor was bound to the operator, the polymerase could not bind to the promoter. But, as will be explored in the next chapter, this is *not*the case. RNA polymerase *can*bind to the *lac*promoter even when repressor is boudn to the *lac* operator. However, the polymerase *cannot initiate*transcription when juxtaposed to the repressor.

4. **Conformational shift in repressor when inducer binds**

a. The repressor has two different domains, one that binds to DNA ("headpiece" containing the helix-turn-helix domain) and another that binds to the inducer (and other subunits) (called the "core). These are connected by a "hinge" region.

b. These structural domains can be distinguished by the phenotypes of mutations that occur in them.

*lacI-d*prevents binding to DNA, leads to constitutive expression.

*lacIS*prevents binding of inducer, leads to a noninducible phenotype.

c. Binding of inducer to the "core" causes an allosteric shift in the repressor so that the "headpiece" is no longer able to form a high affinity complex with the DNA, and the repressor can dissociate (go to one of the many competing nonspecific sites).

Positive control: "catabolite repression"

1. Catabolite repression

a. Even bacteria can be picky about what they eat. Glucose is the preferred source of carbon for *E. coli*; the bacterium will consume the available glucose before utilizing alternative carbon sources, such as lactose or amino acids.

b. Glucose leads to repression of expression of *lac*and some other catabolic operons. This phenomenon is called catabolite repression.

2. Two components are needed for this form of regulation

a. cAMP

[1]In the presence of glucose, the [cAMP] inside the cell decreases from 10-4 M to 10-7 M. A high [cAMP] will relieve catabolite repression.

[2]cAMP synthesis is catalyzed by adenylate cyclase (product of the *cya*gene)

$$ATP \rightarrow cAMP + PP_i$$

b. Catabolite Activator Protein = CAP

[1]Product of the *cap*gene, also called *crp*(cAMP receptor protein).

[2]Is a dimer

[3]Binds cAMP, and then the cAMP-CAP complex binds to DNA at specific sites

3. Binding site for cAMP-CAP

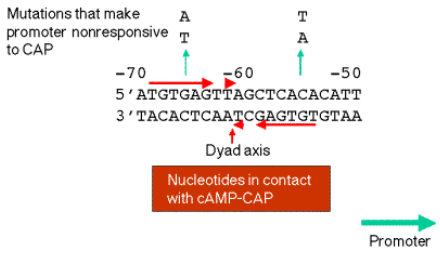
a. In the lac operon, the binding site is a region of about 20 bp located just upstream from the promoter, from -52 to -72.

b. The pentamer TGTGA is an essential element in recognition. For the lac operon, the binding site is a dyad with that sequence in both sides of the dyad.

c. Contact points betwen cAMP-CAP and the DNA are close to or coincident with mutations that render the *lac*promoter no longer responsive to cAMP-CAP.

d. cAMP-CAP binds on one face of the helix.

Figure 4.1.5. Binding site for cAMP-CAP



4. Binding of cAMP-CAP to its site will enhance efficiency of transcription initiation at promoter

a. The *lac*promoter is not a particularly strong promoter. The sequence at -10, TATGTT, does not match the consensus (TATAAT) at two positions.

b. In the presence of cAMP-CAP, the RNA polymerase will initiate transcription more efficiently.

c. The *lac*UV5 promoter is an up-promoter mutation in which the -10 region matches the consensus. The lac operon driven by the UV5 promoter will achieve high level induction without cAMP-CAP, but the wild-type promoter requires cAMP-CAP for high level induction.

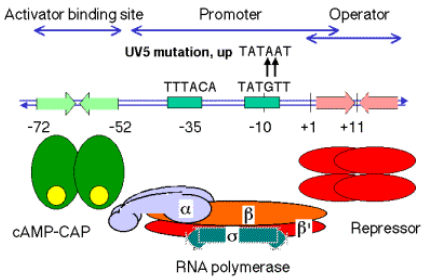


Figure 4.1.6. Regulatory region of lac operon, including CAP binding site

5. **Mode of action of cAMP-CAP**

   a. Direct positive interaction with RNA polymerase. The C-terminus of the a subunit is required for RNA polymerase to be activated by cAMP-CAP. This will be explored in more detail in Chapter 16.

   b. cAMP-CAP bends the DNA about 90o.

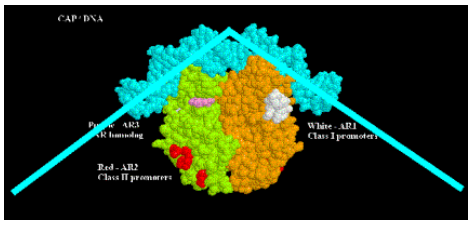


Figure 4.1.7. DNA (top helical structure) is bent by the CAP dimer .

Some generalities

- Repressors, activators and polymerases interact primarily with one face of the DNA double helix.
- Regulatory proteins, such as activators and repressors, are frequently symmetrical and bind symmetrical sequences in DNA.
- RNA polymerases are not symmetrical, and the promoters to which they bind also are asymmetrical. This confers directionality on transcription.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

15.E: Positive and negative control of gene expression (Exercises)

Q15.1

Amber mutations are one class of nonsense mutations. They lead to premature termination of translation by alternation of an amino acid-encoding codon to a UAG terminator, e.g. CAG (Gln) may be changed to UAG (stop; amber). The phenotype of such amber mutants can be suppressed by amber-suppressor genes, which are mutant tRNA genes that encode tRNAs that recognize UAG codons and allow insertion of an amino acid during translation. Which genes or loci in the *lac* operon can give rise to amber-suppressible mutations?

15.2 (POB) Negative regulation.

In the *lac*operon, describe the probable effect on *lacZ* gene expression of:

a. Mutations in the *lac*operator
b. Mutations in the *lac*I gene
c. Mutations in the promoter

Q15.3

Consider a negatively controlled operon with two structural genes (A and B, for enzymes A and B) an operator gene (0) and a regulatory gene (R). In the wild-type haploid strain grown in the absence of inducer, the enzyme activities of A and B are both 1 unit. In the presence of an inducer, the enzyme activities of A and B are both 100 units. For parts a-d, choose the answer that best describes the enzyme activities in the designated strains.

Uninduced Induced

Enz A Enz B Enz A Enz B

a) R+0CA+B+ a) 1 1 100 100

b) 1 100 100 1

c) 50 50 100 100

b) R-0+A+B- a) 1 1 100 100

b) 100 100 100 100

c) 100 0 100 0

c) R+0CA+B+/R+0+A+B+ a) 2 2 200 200

b) 51 51 200 200

c) 200 2 2 200

d) R-0+A+B+/R+0+A+B+ a) 2 2 200 200

b) 2 101 2 101

c) 200 200 200 200

Q15.4 (POB) Positive regulation.

A new RNA polymerase activity is discovered in crude extracts of cells derived from an exotic fungus. The RNA polymerase initiates transcription only from a single, highly specialized promoter. As the polymerase is purified, its activity is observed to decline. The purified enzyme is completely inactive unless crude extract is added to the reaction mixture. Suggest an explanation for these observations.

Q15.5

Consider a hypothetical regulatory scheme in which citrulline induces the production of urea cycle enzymes. Four genes (citA, citB, citC, citD) affecting the activity or regulation of the enzymes were analyzed by assaying the wild-type and mutant strains for argininosuccinate lyase activity and arginase activity in the absence (-cit) or presence (+cit) of citrulline. In the following table,

wild-type alleles of the genes are indicated by a + under the letter of the cit gene and mutant alleles are indicated by a - under the letter. The activities of the enzymes are given in units such that 1 = the uninduced wild-type activity, 100 = the induced activity of a wild-type gene, and 0 = no measurable activity. In the diploid analysis, one copy of each operon is present in each cell.

Strain lyase activity arginase act.

number genes - cit + cit - cit + cit

Haploid: A B C D

1 + + + + 1 100 1 100

2 - + + + 100 100 100 100

3 + - + + 0 0 1 100

4 + + - + 100 100 100 100

5 + + + - 1 100 0 0

Diploid: A B C D / A B C D

6 + + + - / + - + + 1 100 1 100

7 - + + + / + - + + 1 100 2 200

8 + + - + / + - + - 100 100 100 100

9 + - - + / + + + - 1 100 100 100

Use the data in the table to answer the following questions.

a) What is the phenotype of the following strains with respect to lyase and arginase activity? A single word will suffice for each phenotype.

Lyase activity Arginase activity

Strain 2 _____ _____

Strain 3 _____ _____

Strain 4 _____ _____

Strain 5 _____ _____

Strain 6 _____ _____

b) What can you conclude about the roles of *citB* and *citD* in the activity or regulation of the urea cycle in this organism? Brief answers will suffice.

c) What is the relationship (recessive or dominant) between wild-type and mutant alleles of *citA* and *citC*? Be as precise as possible in your answer.

d) What can you conclude about the roles of *citA* and *citC* in the activity or regulation of the urea cycle in this organism? Brief answers will suffice.

Q15.6

Consider a hypothetical operon responsible for synthesis of the porphyrin ring (the heterocyclic ring that is a precursor to heme, cytochromes and chlorophyll). Four genes or loci, *porA*, *porB*, *porC*, and *porD* that affect the activity or regulation of the biosynthetic enzymes were studied in a series of haploid and diploid strains. In the following table, wild-type alleles of the genes or loci are indicated by a + under the letter of the *por* gene or locus and mutant alleles are indicated by a — under the letter. The activities of two enzymes involved in porphyrin biosynthesis, d-aminolevulinic acid synthetase and d-aminolevulinic acid dehydrase (referred to in the table as ALA synthetase and ALA dehydrase), were assayed in the presence or absence of heme (one product of the pathway). The units of enzyme activity are 100 = non-repressed activity of the wild-type enzyme, 1 = repressed activity of the wild-type enzyme (in the presence of heme), and 0 = no measurable activity. In the diploid analysis, one copy of each operon is present in each cell.

Strain ALA synthetase ALA dehyd.

number *por* *- heme* *+heme* *- heme* *+ heme*

Haploid: A B C D

1 + + + + 100 1 100 1

2 - + + + 100 100 100 100

3 + - + + 0 0 100 1

4 + + - + 100 1 0 0

5 + + + - 100 100 100 100

Diploid: A B C D / A B C D

6 + - + + / + + - + 100 1 100 1

7 - + + + / + + - + 200 101 100 100

8 + + + - / + + - + 200 2 100 1

9 - + - + / + - + - 100 100 100 1

Use the data in the table to answer the following questions.

a) Describe the phenotype of the following the strains with respect to ALA synthetase and ALA dehydrase activities. A single word will suffice for each phenotype.

ALA synthetase ALA dehydrase

Strain 2 _____ _____

Strain 3 _____ _____

Strain 4 _____ _____

Strain 5 _____ _____

Strain 6 _____ _____

b) What is the relationship (dominant or recessive) between wild-type and mutant alleles of the four genes, and which strain demonstrates this? Please answer in a sentence with the syntax in this example: "Strain 20 is repressible, which shows that mutant *grk*1 is dominant to wild-type."

porA Strain ___ is _____, which shows that _____

porA is _____.

porB Strain ___ is _____, which shows that _____

porB is _____.

porC Strain ___ is _____, which shows that _____

porC is _____.

porD Strain ___ is _____, which shows that _____

porD is _____.

c) What is the role of each of the genes in activity or regulation of porphyrin biosynthesis? Brief phrases will suffice.

d) Is this operon under positive or negative control?

16: Transcription regulation via effects on RNA polymerases

[Dr. Tracy Nixon made major contributions to this chapter.]

A. The multiple steps in initiation and elongation by RNA polymerase are targets for regulation

RNA Polymerase has to

- bind to promoters,
- form an open complex,
- initiate transcription,
- escape from the promoter,
- elongate , and
- terminate transcription.

See Figure 4.2.1.

Summarizing a lot of work, we know that strong promoters have high KB, high kf, low kr, and high rates of promoter clearance. Furthermore, weak promoters have low KB, low kf, high kr, and low rates of promoter clearance and moderate promoters have one or more "weak" spots.

To learn these facts, we need genetic data to identify which macromolecules (DNA and proteins) interact in a specific regulation event, and to determine which base pairs and amino acid residues are needed for that regulation event. We also need biochemical data to describe the binding events and chemical reactions that are affected by the specific regulation event. Ideally, we would determine all forward and reverse rate constants, or equilibrium constants (which are a function of the ratio of rate constants) if rates are inaccessible. Although, in reality, we cannot get either rates or equilibrium constants for many of the steps, some of the steps are amenable to investigation and have proved to be quite informative about the mechanisms of regulation.

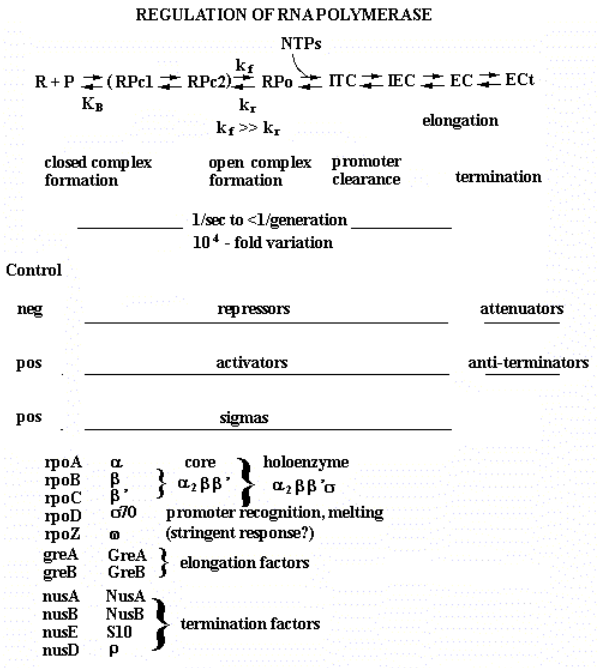


Figure 4.2.1. Pc = closed promoter complex, Po = open promoter complex, ITC = initial transcribing complex, IEC = initial elongating complex, EC = elongation complex, ECt = terminating elongation complex.

B. Methods exist for measuring rate constants and equilibrium constants, and newer, more accurate methods are now being used

Classical methods of equilibrium studies and data analysis use low concentrations of enzymes and make assumptions that simplify complex reactions so that they can be treated by definite integrals of chemical flux equations. They also manipulate an equation into a form that can be plotted as a linear function, and derive parameter estimates by slope and intercept values

Driven by the success of recombinant DNA and protein purification technology, and by the increased computational power in desktop computers, the classical methods are being replaced by

- using of large amounts of enzymes to directly include them in kinetic studies. In this approach, the enzymes are used in substrate level quantities.
- numerical integrations of chemical flux equations (Kinetic Simulation)
- more rigorous methods based on NonLinear, Least Squares (NLLS) regression, and
- analyzing data from multiple experiments of different design simultaneously (global NLLS analysis).

These changes increase the steps in a reaction that can be examined experimentally, replace the limited set of simple mechanisms that can be analyzed with essentially any mechanism and increase knowledge of error, permitting conclusions to be drawn with more confidence

Synonymous and related terms

$K_B = K_b = K_{eq}$ = equilibrium constant for binding
$K_S = K_B$ for binding of protein to a *specific* DNA sequence
$K_{NS} = K_B$ for binding of protein to ***non****specific* DNA

[P] = [P_2] = molar concentration of protein
[R_4] = molar concentration of repressor

[D] = molar concentration of free DNA
[D_S] = concentration of free specific DNA
[D_{NS}] = concentration of free **non**specific DNA

[DP] = molar concentration of DNA-protein complex
[R_4D_S] = concentration of repressor-operator

Box 1: The equations used in this chapter come from several different sources that use different names for the same thing. The following lists some of these synonyms.

C. Experimental approaches to macromolecular binding reactions

Several methods are available for measuring the amount of protein that binds specifically to a DNA molecule. We have already encountered these as methods for localizing protein-binding sites on DNA, and all are amenable to quantization. Major methods include **nitrocellulose filter binding**, **electrophoretic mobility shift assays**, and **DNase protection assays**.

Which Experimental Technique is Best?

- The kind of observations that can be made about the system differ for different experimental approaches.
- These differences lead to specific problems with each technique,
- Each technique depends on combining the analysis of more than one experiment to obtain enough information to resolve intrinsic binding free energy from cooperativity energy.

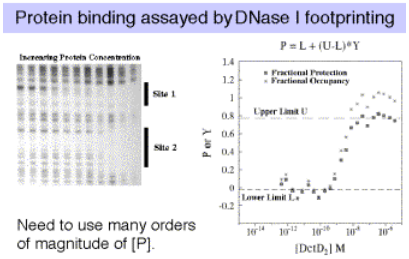


Figure 4.2.2 Data courtesy of Dr. Tracy Nixon

The most robust technique is DNase I footprinting. If you are studying the binding of multiple, interacting proteins, then it is possible that these proteins are showing cooperativity in their binding to DNA. When analyzing such cooperativity by DNase I footprinting, the resolution is limited to cooperativities >0.5 kcal/mole, and is subject to some critical assumptions. Gel-shifts (also called electrophoretic mobility shift assays, or EMSAs) are useful when there is no cooperativity, or when cooperativity is large relative to site heterogeneity. Filter binding studies require knowledge about filter retention efficiencies for the different protein-DNA complexes, which can only be empirically determined. And always keep in mind that flanking sequences do affect binding affinities, and even point mutations can have distant effects.

In any of these assays, we are devising a physical means for measuring a quantity that is related to fractional occupancy.

D. Measurement of equilibrium constants in macromolecular binding reactions

Classical methods with their linear transformation are **not** as accurate as the NonLinear, Least Squares (NLLS) regression analysis, but **they can serve to show the general approach**.

a.The binding constants can be determined by titrating labeled DNA binding sites with increasing amounts of the repressor, and measuring amount of protein-bound DNA and the amount of free DNA. Typical techniques are electrophoretic mobility shift assays or nitrocellulose filter binding.

Note that for a simple equilibrium of a single protein binding to a single site on the DNA, the equilibrium constant for binding (K_B) is approximated by the inverse of the protein concentration at which the concentration of DNA bound to protein equals the concentration of free DNA (Figure 4.2.3).

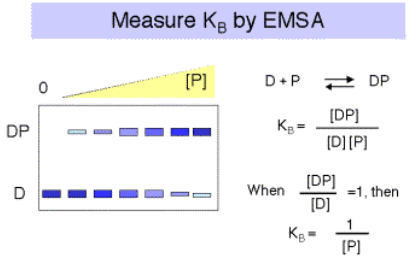


Figure 4.2.3

If it were possible to reliably determine both the concentration of DNA bound to protein (i.e. [DP]) and the concentration of free DNA ([D]), then one could plot the ratio of bound DNA to free DNA at each concentration of repressor. If the results were linear, then the slope of the line would give the equilibrium binding constant, KB . See Figure 4.2.4.

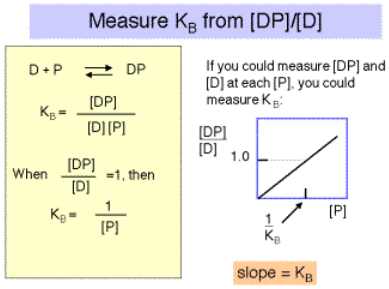


Figure 4.2.4

However, the error associated with determining very low concentrations of free or bound DNA is substantial, and a more reliable measurement is that of the ratio of bound DNA to total DNA, i.e. $[DP]/[D]_{tot}$, as illustrated in Figure 4.2.5. The equation describing this binding curve has a form equivalent to the Michelis-Menten equation for steady-state enyzme kinetics. Note that the concentration of protein at which half the DNA is bound to protein is the inverse of K_B . You can show this for yourself by substituting 0.5 for $[DP]/[D]_{tot}$ in the equation. At this point, $[P] = 1/K_B$.

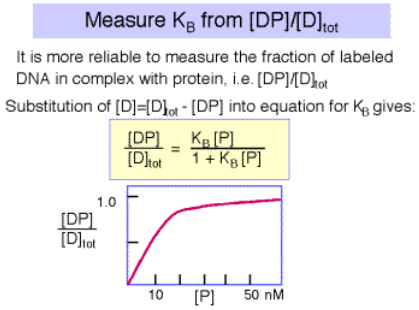


Figure 4.2.5.

2. **Problems with the classical approach**.

In this classical approach, experiments were designed such that

o one or more concentrations could be assumed to be unchanging, and

o observations were manipulated mathematically (transformed) to a linear equation so that one could

+ plot the transformed data,

+ decide where to draw a straight line, and

+ use the slope and intercepts to estimate the parameters in question. (Scatchard plots, Lineweaver-Burke plots, etc).

* Two problems are associated with the older technique

o Deciding where to draw the straight line is an arbitrary decision for each person doing the analysis (and using a linear regression to find the "best fit" line is not justified, as two of the assumptions about your data that are needed to justify such a regression are not true)

o There is no accurate estimate of the error in the estimate of the parameter value

3. These limitations have been overcome in the last 5 or so years, aided by the advent of recombinant DNA techniques that allow the production of large amounts of the proteins being analyzed, and the availability of powerful microcomputers that can carry out the large number of computations required for **nonlinear, least squares regression analysis (NLLS)**.

a. We can model binding reactions by

- tabulating the different states that exist in a system,
- associating each state with a fractional probability based on the Boltzmann partition function and the Gibb's free energy for that state (DGs),
- and determine the probability of any observed measurement by the ratio of
    - the sum of fractional probabilities that give the observation, and
    - the sum of the fractional probabilities of all possible states.

Where *j*is the number of ligands bound, the fractional probability of a particular state is given by this equation for *fs*.

$$f_s = \frac{e^{-\Delta G_s / RT} \times [P_2]^j}{\sum_{s_j} e^{-\Delta G_s / RT} \times [P_2]^j}$$

As an example, consider a one-site system, such as an operator that binds one protein. There are two states, the 0 state with no protein bound to the operator and the 1 state with one protein bound. Thus one can write the equation for *f0* and for *f1*.

If we expand the fractional probabilities for each of these fractional occupancy equations, we derive equations relating fractional occupancy, $\bar{Y}$, to a function of Gibb's free energies for binding (DG), protein concentration ([P2]), and complex stoichiometry (j).

For a single site system, we have the following equations:

$$\bar{Y} = \frac{f_1}{\sum f_s}$$

$$\bar{Y} = \frac{e^{-\Delta G / RT} \times [P_2]}{1 + e^{-\Delta G / RT} \times [P_2]}$$

Since Gibb's free energy is also related to the equilibrium constant for reactions:

$$\Delta G = -RT \ln(K_{eq}) \tag{16.1}$$

these free energies can be re-cast as equilibrium constants, as follows.

$$\bar{Y} = \frac{K_b \times [P_2]}{1 + (K_b \times [P_2])}$$

A more complete presentation of this method, including a treatment of multiple binding sites, can be obtained at the BMB Courses web site (www.bmb.psu.edu/courses/default.htm) by clicking on BMB400 "Nixon Lectures."

b. Analyzing the data

After collecting the binding data, we are in a position to analyze the observed data to find out what values for DG or Kb make the function best predict the observations. Statisticians have developed Maximum Likelihood Theory to allow using the data to find, for each parameter, the value that is most likely to be correct. For biochemical data the approach that is most appropriate (most of the time) is global, nonlinear, least squares (NLLS) regression.

• Fortunately, desktop computers are now powerful enough to do these calculations in a few minutes, for one experiment, or even for many experiments combined in a global analysis. This method has several advantages. It gives you:

o the same parameter estimates, no matter what program or method you or someone else uses, provided that the program is written correctly and used correctly.

o much more rigorous estimates of error.

This last point is worth emphasizing:

• is it not true that $100 (minus $50) is much less attractive as a fee for your time than is $100 (minus $0.01)? The same can be true for estimates of binding free energies, or equilibrium constants .

• Moreover, when several experiments are required to estimate a parameter, the error in each experiment should be included in the estimate of the parameter. Without a global analysis that determines a conglomerate error, it is not possible to carefully carry forward the error of one experiment to the analysis of data from additional ones.

c. This analysis produces a plot of the variance of fit, or error, over a wide range of possible values for the parameter being measured, such as the DG for binding. **The DG value with the smallest error is the most accurate value**.

An example of this analysis is shown in Figure 4.2.6. The raw data shown in Figure 4.2.2 (left panel) produced the binding curves shown on right panel of that figure. These data were then subjected to non-linear least-squares analysis. The errors (or variance of fit) for each possible value of DG are plotted in Figure 4.2.6. For example, note that the lowest variance of fit for DG_1 is about –9.5 kcal/mole.

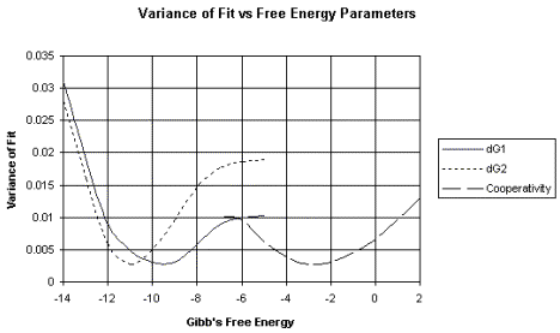


Figure 4.2.6.

dG1 = DG1 = Gibb's free energy for binding to the first site of a two-site system.

dG2 = DG2 = Gibb's free energy for binding to the second site of a two-site system.

The variance of fit for the DG for the cooperativity between proteins bound at the two sites is also plotted.

These data were kindly provided by Dr. Tracy Nixon.

As indicated above, once a value for DG is available, one can calculate Keq from

DG = -RT ln (Keq)

Figure 4.2.7.

Some key references for NLLS:

Senear and Bolen, 1992, Methods Enzymol. 210:463

Koblan et al, 1992, Methods Enzymol. 210:405.

Senear et al 1991, J. Biol. Chem. 266:13661

E. Insights into the mechanism of lac regulation by measuring binding constants

Having gone through both classical and non-linear least squares analysis for measuring binding constants, let's look at an example of how one uses these measurements to better understand the mechanism of gene regulation. We know that transcription of the *lac*operon is increased in the presence of the inducer, but how does this occur? One could list a number of possibilities, each with different predictions about how the inducer may affect the binding constant of repressor for operator, K_B.

a. Does the inducer change the conformation of the lac repressor so that it now activates transcription? This could occur with no effect on K_B.
b. Does inducer cause the repressor to dissociate from the operator DNA and remain free in solution? This predicts a decrease in K_B for specific DNA, but no binding to nonspecific DNA.
c. Does inducer cause the repressor to dissociate from the operator and redistribute to nonspecific sites on the DNA? This predicts a decrease in K_B for specific DNA, but proposes that most of the repressor is bound to non-operator sites.

Measurement of the equilibrium constants for *lac* repressor binding to operator and to nonspecific DNA, in the absence and presence of the inducer, shows that possibility **c** above is correct. This section of the chapter explores this result in detail.

In the *absence* of inducer, the repressor, or R_4, will bind to **specific sites** (in this case the operator) with **high affinity** and to **nonspecific sites** (other DNA sequences) with lower affinity (Figure 4.2.8). This is stated quantitatively in the following values for the equilibrium association constant. Either equilibrium constant can be abbreviated Keq or KB. We will use the term KS to refer to KB at specific sites and KNS for the KB at nonspecific sites.

$$KS = 2 \times 10^{13} M - 1 KNS = 2 \times 10^{6} M - 1 \tag{16.2}$$

[A detailed presentation of some representative data and how to use them to determine these binding constants for the lac repressor is in Appendix A at the end of this chapter. This Appendix goes through the classic approach to measuring binding constants.]

3. The binding constant of *lac* repressor to its operator changes in the presence of inducer. (Figure 4.2.8)

Binding of the inducer to the repressor **lowers**the affinity of the repressor for the **operator**1000 fold, but does **not** affect the affinity of repressor for **nonspecific sites**.

For *R*4 with *inducer* :

*K*S = 2 x 1010 *M* -1 *KNS*= 2 x 106 *M* -1

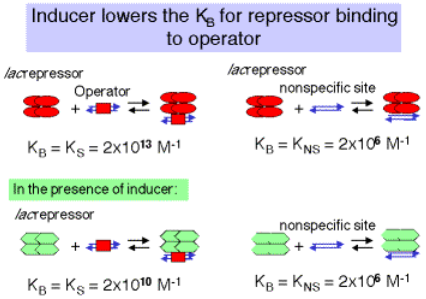


Figure 4.2.8.

4. The difference in affinity for specific versus nonspecific sites can be described by the **specificity parameter**, which is the ratio between the equilibrium constant for specific binding and the equilibrium constant for nonspecific binding.

Specificity = $\frac{K_S}{K_{NS}} = 10^7$ *in* ***absence****of inducer*

$\frac{K_S}{K_{NS}} = 10^4$ *in* ***presence****of inducer*

Note the in the presence of the inducer, the specificity with which the *lac*repressor binds to DNA is decreased 1000-fold.

Even though the repressor still has a higher affinity for specific DNA in the presence of the inducer, there are *so many nonspecific sites*in the genome that the repressor stays bound to these nonspecific sites rather than finding the operator. Hence in the presence of the inducer, the operator is largely unoccupied by repressor, and the operon is actively transcribed.

The regulation of the *lac* operon via redistribution of the repressor to nonspecific sites in the genome is covered in more detail in the next two sections. They show the effect of having a large number of nonspecific, low affinity sites competing with a single, high affinity site for a small number of repressor molecules.

5.Distribution of repressor between operator and nonspecific sites

Although repressor has a much higher affinity for the operator than for nonspecific sites, there are so many more nonspecific sites (4.6 x 106, since essentially every nucleotide in the *E. coli*genome is the beginning of a nonspecific binding site) than specific sites (one operator per genome) that virtually all of the repressor is bound to DNA, even if only nonspecific sites are present.

a. We use the binding constants above, and couple them with a calculation that the concentration of repressor (10 molecules per cell) is 1.7 x 10-8 M and the concentration of nonspecific sites (4.6 x 106 per cell) is 7.64 x 10-3 M. These values for [R4] and [DNS] are essentially constant. With this information, we can compute that the ratio of free repressor to that bound to nonspecific sites is less that 1 x 10-4 (it is about 6.6 x 10-5), as shown in the box below. Thus only about 1 in 15,000 repressor molecules is not bound to DNA.
b. This analysis shows that the *lac* repressor is partitioned between nonspecific sites and the operator. When it is not bound to the operator, it is bound elsewhere to any of about 4.6 million sites in the genome. Almost none of the repressor is unbound to DNA in the cell.
c. Box 2 (below) goes through these calculations in more detail.

Box 2. Effectively all repressor protein is bound to DNA.

$$[R_4]_{tot} = \frac{10\ molecules}{cell} = \frac{10\ molec/6.02\times10^{23}\ molec\ mole^{-1}}{10^{-15}\ L} = 1.7\times10^{-8}\ M$$

$$[D_{NS}] = \frac{4.6\times10^{6}\ sites}{cell} = \frac{4.6\times10^{6}\ sites\ /\ 6.02\times10^{23}\ molecules\ /\ mole}{10^{-15}\ L} = 7.64\times10^{-3}\ M$$

$$K_{NS} = \frac{[R_4D_{NS}]}{[R_4][D_{NS}]} = 2\times10^{6}\ M^{-1}$$

6. Regulation of the lac operon via redistribution of the repressor to nonspecific sites in the genome

a. The high specificity of repressor for the operator means that in the absence of inducer, the operator is bound by the repressor virtually all the time. This is true despite the huge excess of nonspecific binding sites.

b. The specificity parameter described above (Ks/Kns) allows one to evaluate the simultaneous equilibria (repressor for operator and repressor for nonspecific sites on the DNA). We want to calculate the ratio of repressor-bound operators to free operators. Values for KS, KNS, and [DNS] are already known, and the concentration of repressor not bound to DNA is negligible.

Box 3. Specificity parameter is related to ratio of bound to free operator sites.

Specificity = $\frac{K_S}{K_{NS}} = \frac{\frac{[R_4D_S]}{[R_4][D_S]}}{\frac{[R_4D_{NS}]}{[R_4][D_{NS}]}} = \frac{[R_4D_S]}{[D_S]} \times \frac{[D_{NS}]}{[R_4D_{NS}]}$

ratio of Bound:Free

operator sites

Now we need a value for [R4DNS]. This is obtained by realizing that under conditions that saturate specific sites, the concentration of repressor bound to nonspecific sites is closely approximated by [repressor]total - [operator], or $[R_4]_{total}$ - $[D_s]_{total}$ in the equations in Box 4.

Box 4.

$$[R_4D_{NS}] = [R_4]_{tot} - [R_4D_S] - [R_4]_{free}$$

[*R*4]*free* is negligible (see above).

Under conditions that saturate specific sites,

$$[R_4D_S] = [D_S]_{tot}$$

Thus [*R4DNS*] = $[R_4]_{total}$ - $[D_S]_{total}$

$$[D_S]_{tot} = \frac{1\ site}{cell} = \frac{1\ molec/6.02\times10^{23}\ molec\ mole^{-1}}{10^{-15}\ L} = 1.7\times10^{-8}\ M$$

$$[D_{NS}] = 7\times10^{-3}\ M$$

c. After making these simplifying assumptions, we now have a value for every variable and constant in the equation, except the ratio of bound:free operator sites. Thus we can compute the desired ratio.

Box 5. Equation relating specificity to the ratio of bound to free operator and a set of constants.

Specificity = $\frac{K_f}{K_{ns}} = \frac{[R_4 D_f]}{[D_f]} \times \frac{[D_{ns}]}{[R_4]_{total} - [D_f]_{total}}$

already want to constants

measured determine

d. Now that we have the equation in Box 5, we can calculate the ratio of free operator to operator bound by repressor can be calculated in the absence and presence of inducer.

(1) In the absence of inducer:

Specificity = $\frac{K_f}{K_{ns}} = 10^7$

$$\frac{[D_f]}{[R_4 D_f]} = \frac{K_{ns}}{K_f} \times \frac{[D_{ns}]}{[R_4]_{total} - [D_f]_{total}} = \frac{1}{10^7} \times \frac{7.64 \times 10^{-3} M}{17 \times 10^{-9} M - 1.7 \times 10^{-9} M}$$

$$\frac{[D_f]}{[R_4 D_f]} = \frac{1}{10^7} \times 4.99 \times 10^5 = 0.0499 = 0.050$$

i.e. the ratio of free operators to operators bound by repressor is 0.05. *R*4 is bound to the operator ~ 95% of the time. Thus the operon is not expressed.

(2) In the presence of inducer:

Specificity = $\frac{K_f}{K_{ns}} = 10^4$

$$\frac{[D_f]}{[R_4 D_f]} = \frac{1}{10^4} \times 4.99 \times 10^5 = 50 \quad or \quad \frac{[R_4 D_f]}{[D_f]} = 0.02$$

i.e. in the presence of inducer, only about 2% of the operators are bound by repressor, or *R*4 is bound to the operator ~ 2% of the time. Thus the operon is expressed.

In summary, these calculations show that in the absence of inducer, 95% of the operators are occupied (*o* is bound by *R*4 95% of the time). In the presence of inducer, the repressor re-distributes to nonspecific sites on the DNA, leaving only 2% of the operators bound by *R*4. Thus the **operon is expressed**in most of the cells.

An additional example of the use of the measured binding constants and the specificity parameter is in Appendix B at the end of this chapter. This example explores the effects of operator mutants.

F.Mechanism of repression and induction for the *lac*operon

1.Effect of lac repressor on the ability of RNA polymerase to bind to the promoter

The analysis in the previous section showed how the inducer affects the partitioning of the repressor between specific and nonspecific sites. Now let's examine the effect that repressor bound to the operator has on the function of the **polymerase**at the promoter

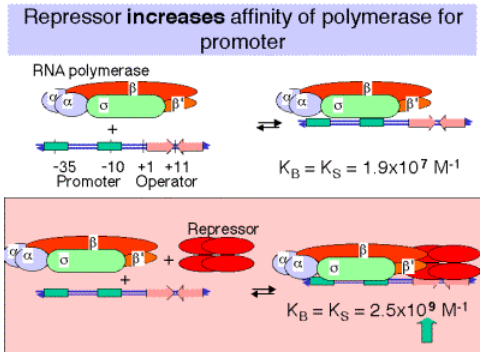


Figure 4.1.9

a.Binding of repressor to the operator actually **increases the affinity**of the RNA polymerase for the promoter!

Consider the following equilibrium:

RNA polymerase + promoter RNA polymerase-promoter (closed complex)

In the absence of repressor on the operator, the affinity of RNA polymerase for the *lac*promoter is

KB = 1.9 x 107 M-1

In the presence of repressor on the operator, the affinity is

KB = 2.5 x 109 M-1

b.Repressor bound to the operator *increases*the affinity of RNA polymerase for the *lac*promoter about 100 fold, so the closed complex is formed much more readily. The repressor essentially holds the RNA polymerase in storage at the promoter, but transcription is not initiated.

c.Upon binding of the inducer to the repressor, the repressor dissociates and the RNA polymerase-promoter complex can shift to the open complex and initiate transcription, thus switching on the operon.

d.Thus the effect of repressor bound to the operator is not on Kb for the polymerase-promoter interaction, but rather is on kf for the conversion from closed to open complex.

G. Kinetic measurements of the abortive initiation reaction allow one to calculate kf.

1. **Abortive Transcription Assay**

The initial transcribing complex (ITC) that exists after open complex formation frequently fails to transform into the initial elongating complex (IEC). The RNA product is released, and the system initiates again. The rate at which the aborted transcripts accumulates can provide a measure of promoter strength, and experiments have been devised to use such an assay to estimate KB for polymerase binding to the promoter region, and kf for isomerization from closed to open complex form. Polymerase, promoter DNA, and nucleotides are mixed such that a radiolabeled phosphate will be introduced into transcripts that are made and aborted. The amount of radioactivity in the short transcripts is then counted as a function of time.

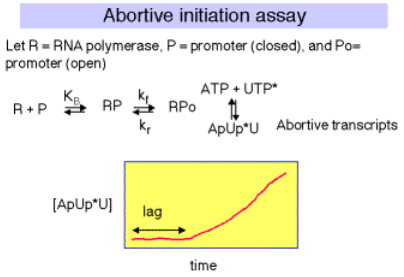
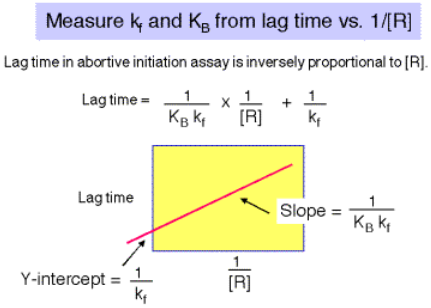


Figure 4.2.10

There is a lag between mixing reagents, and optimal rate of abortive transcript production. The length of this lag is inversely proportional to the [RNAP]. A plot of lag-time vs 1/[RNAP] gives a straight line plot, with slope equal to 1/[KB ´ kf] and y-intercept of 1/kf.

Figure 4.2.11.



H. Activation of transcription by the CAP protein of E. coli

1. **Activation of transcription by the CAP protein of *E. coli*illustrates several general regulatory principles**.

We will focus on the point that in different contexts (different promoters), a single protein can directly interact with RNAP via at least 2 distinct contact surfaces. Depending on the context, CAP can affect KB or kf for RNA polymerase-promoter interactions.

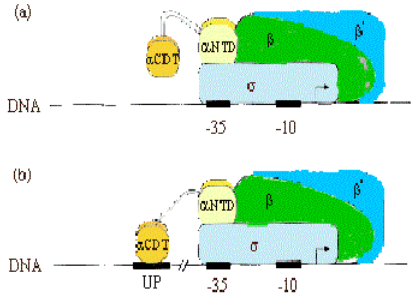
An additional discussion of the ability of CAP to affect the architecture of a protein-DNA complex which contains precise contacts between RNAP and an additional regulatory protein (MalT), by bending DNA, is at the BMB400 Web site, under "Nixon Lectures." This latter point will not be covered in detail here.

2. **aSubunit of RNA polymerase**

a. Recall from Part Three that the a subunit of RNA polymerase has two separate domains. The amino terminal domain (aNTD) is essential for dimerization and assembly of polymerase, and the carboxy terminal domain (aCTD) is needed for binding to DNA and for communication with many, but not all, transcription factors.

Most RNA polymerase (~60%) is associated with rRNA or tRNA genes. This is accomplished by a special sequence upstream of the promoter elements (i.e. the –35 and –10 boxes), called the UP element (-57)5'-AAAATTATTTT-3'(-47), which binds a2 dimers, and increases occupancy by polymerase by ~10-fold.

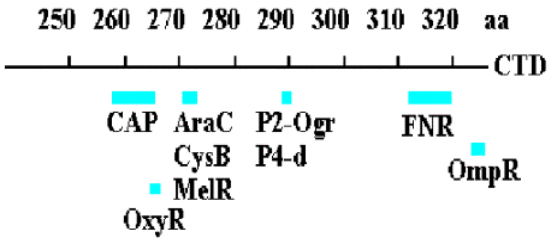
Figure 4.2.12.



b. Much of the communication between activators and *E. coli*RNA polymerase is mediated between the CTD of a and these factors.

(see Ebright and Busby, 1995, Curr. Opinion in Gen. & Dev. 5:197-203)

Figure 4.2.13.



3. Summary & Distinctions between Cap at Class I and Cap at Class II Promoters

(For reviews see Mol Micro 23:853-859 and Cur. Opin. Genet. Dev. 5:197-203).

Class I promoters have CAP binding sites centered at -62, -83, or -93.

At **class II promoters**, it is centered at -42 and overlaps the -35 determinant of the promoter.

Figure 4.2.14. CAP binding to class I and class II promoters.

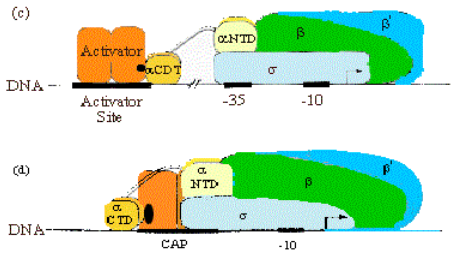


Figure 4.2.14. The dimeric CAP protein is labeled "Activator". Binding to a class I promoter is shown in panel (c) and binding to a class II promoter is shown in panel (d).

4. **CAP has at least two Activation Regions (ARs)**:

- AR1 (residues 156-164)

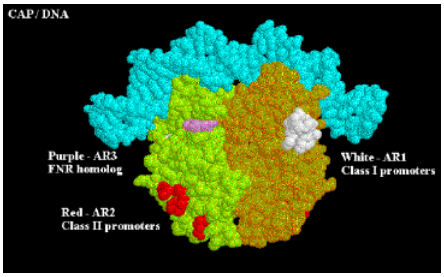
  At class I promoters, AR1 in the downstream subunit of CAP "sees" residues 258-265 of CTD of a. This interaction increases KB for polymerase binding to the promoter.

  At class II promoters, CAP displaces the aCTD (decreasing KB), which is overcome by increasing KB via upstream subunit AR1-aCTD interaction

- AR2 (residues 19, 21, 96, 101)

  At class II promoters, the downstream subunit "sees" aNTD residues 162-165, increasing kf for isomerization from closed to open complexes.

Figure 4.2.15. Activation Regions on CAP



At both class I and class II promoters, CAP AR1 interacts with the CTD of a. It is clear that for class I promoters, residues 258-265 of the a subunit are the target of AR1 of CAP; it is not clear if these are the same residues needed for interaction at class II promoters. At class I promoters, this interaction provides "true" direct activation: the interaction is between the downstream subunit of CAP, and appears to only be used to increase KB for the binding of RNA polymerase to the promoter region (perhaps substituting for the lack of an UP sequence). At class II promoters, AR1 in the upstream subunit contacts the alpha subunit, but it does not appear to cause direct stimulation of transcription. Instead, it overcomes inhibition of polymerase that is hypothesized to arise from CAP displacing the alpha subunit from its preferred position near -45. This is evidenced by the following observations:

- aCTD binds to -40 to -55 region at class II promoters in the absence of CAP, but binds to the -58 to -74 region in its presence
- AR1 mutants in CAP decrease KB for RNA polymerase at class II promoters, but have no affect on kf .
- Removal of the a CDT eliminates the need for CAP AR1 in class II promoters, and has no negative affect.
- In contrast, removal of the a CDT prevents activation by CAP at class I promoters.

In addition to overcoming a decrease in KB by AR1, at class II promoters CAP also exerts a "direct" activation. This occurs between CAP residues 19, 21, 96 and 101 (AR2) in the downstream subunit of CAP, and residues 162-165 of the a subunit NTD. This interaction increases the kf and has no affect on KB. Region 162-165 is between regions 30-55 / 65-75 and 175-185 / 195-210 which are essential for contact with the b and b' subunits of polymerase, respectively. AR2 is not needed for CAP to work at class I promoters.

Appendix A for Chapter 17 (Part Four., section II)

Measurement of equilibrium constants for binding of *lac*repressor to specific and nonspecific sites in DNA

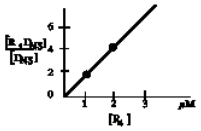
*R*4 = Repressor

D S = Specific DNA site Þ operator

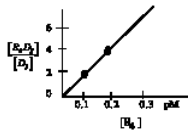
D NS = Nonspecific DNA site Þ all other sites in genome

R4+ DSR4DS R4+ DNSR4DNS

$K_s = \frac{[R_4 \cdot D_S]}{[R_4][D_S]}$ $K_{ns} = \frac{[R_4 D_{ns}]}{[R_4][D_{ns}]}$



$\frac{Bound}{Free} = \frac{[R_4 \cdot D_S]}{[D_S]} = K_S[R_4]$ $\frac{[R_4 D_{ns}]}{[D_{ns}]} = K_{ns}[R_4]$



$slope = K_S = \frac{2}{1 \times 10^{-13} M} = 2 \times 10^{13} M^{-1}$ $K_{ns} = \frac{2}{1 \times 10^{-6} M} = 2 \times 10^{6} M^{-1}$

The *lac*repressor will bind to its specific site, the **operator**, with **very high affinity,**

Keq = KS = 2 x 1013 M-1, where Ks is the equilibrium association constant for binding to a specific site

and it will bind to other DNA sequences, or **nonspecific sites**, with a l**ower affinity**.

Keq = KNS = 2 x 106 M-1, where Kns is the equilibrium association constant for binding to a nonspecific site.

Measurements in the laboratory:

Since it can be difficult to measure the amount of bound or free probe at very low concentrations, it is more reliable to measure the fraction of probe bound as a function of [R4]. The fraction of probe bound is

= .

By substituting [Ds] = [Ds]total- [R4Ds] into the equation for Ks, you can derive the following relationship between the fraction of probe bound by repressor and the concentration of the repressor:

=

{Since the [R4] is usually much greater than the [Ds]total in these assays, the

[R4]free >> [R4Ds], and [R4] is well approximated by [R4]total .}

This equation has the form of the classic Michaelis-Menten equation for steady-state enzyme kinetics, and it is also useful in analysis of many binding assays. Once is plotted against [R4], one can do curve fitting to derive a value for Ks. One can also get a value for Ks by measuring the [R4] at which half the probe is bound. At this point, [R4] = . {This can be seen simply by substituting = 0.5 into the equation above. The algebra is exactly the same as is done for the determination of Km by the Michaelis-Menten analysis.}

Appendix B. Use of binding constants and the equations relating the specificity parameter to the ratio of bound to free operator sites to study the effects of operator mutants.

The same equations used in section E of this chapter also can be used to examine the effects of **operator mutants**. The following analysis shows that a mutation that decreases the affinity of the operator 20-fold for the repressor will result in about half the operators being free of repressor (or the operon being expressed about half the time).

$K_f = 2 \times 10^{13}\,M^{-1}$ for wild-type

$\therefore K_f = \frac{2 \times 10^{13}\,M^{-1}}{20} = 1 \times 10^{12}\,M^{-1}$ for the mutant

Specificity = $\frac{K_f}{K_{ns}} = \frac{1 \times 10^{12}\,M^{-1}}{2 \times 10^{6}\,M^{-1}} = 0.5 \times 10^{6} = 5 \times 10^{5}$

$$\frac{[D_f]}{[R_1 D_f]} = \frac{K_{ns}}{K_f} \times \frac{[D_{ns}]}{[R_1]_{free} \cdot [D_f]_{free}} = \frac{1}{5 \times 10^{5}} \times 4.99 \times 10^{5}$$

$$\frac{[D_f]}{[R_1 D_f]} = 0.998 = 1.0$$

This says that the operator is essentially equally distributed between the bound and free form.

$$[D_f]_{total} = [D_f] + [R_1 D_f]$$
$$\frac{[D_f]}{[D_f]_{total} - [D_f]} = 1.0$$
$$[D_f] = [D_f]_{total} - [D_f]$$
$$2[D_f] = [D_f]_{total}$$
$$\frac{[D_f]}{[D_f]_{total}} = \frac{1}{2} = 0.50$$

50% of the operators are not occupied by repressor, thus only about half of the operons will be expressed (in a population of bacteria), or any particular operon will be expressed about half the time.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

16.E: Transcription regulation via effects on RNA polymerases (Exercises)

16.1 The ratio [RDs]/[Ds] is the concentration of a hypothetical repressor (R) bound to its specific site on DNA divided by the concentration of unbound DNA, i.e. it is the ratio of bound DNA to free DNA. When the measured [RDs]/[Ds] is plotted versus the concentration of free repressor [R], the slope of the plot showed that the ratio [RDs]/[Ds] increased linearly by 60 for every increase of 1x10-11 M in [R]. What is the binding constant Ks for association of the repressor with its specific site?

16.2 The binding of the protein TBP to a labeled short duplex oligonucleotide containing a TATA box (the probe) was investigated quantitatively. The following table gives the fraction of total probe bound (column 2) and the ratio of bound to free probe (column 3) as a function of [TBP]. These data are provided courtesy of Rob Coleman and Frank Pugh.

	[TBP] nM		
	0.10	0.040	0.042
	0.20	0.16	0.19
	0.30	0.33	0.5
	0.40	0.44	0.78
	0.50	0.52	1.1
	0.70	0.62	1.6
	1.0	0.71	2.45
	2.0	0.83	4.88
	3.0	0.87	6.69
	5.0	0.93	14
	10	0.97	32.3
	20	0.99	99

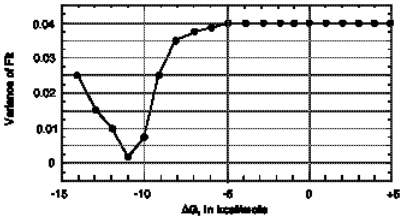
Plot the data for the two different measures of bound probe. Note that since the denominator for column 2 is a constant, the ratio of bound to total probe will level off, whereas the amount of free probe can continue to decrease with increasing [TBP], and thereby getting a continuing increase in the ration of bound to free probe.

What is the equilibrium constant for TBP binding to the TATA box?

16.3 What is the fate of the *lac* repressor after it binds the inducer?

16.4 How does the *lac* repressor prevent transcription of the *lac*operon?

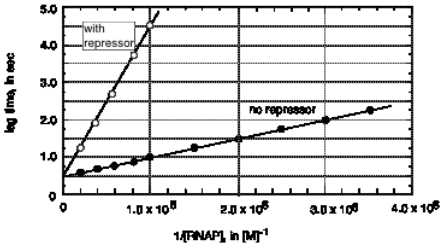
For the next two questions, let's imagine that you mixed increasing amounts of the DNA binding protein called AP1 with a constant amount of a labeled duplex oligonucleotide containing the binding site (TGACTCA). After measuring the fraction of DNA bound by AP1 (i.e. the fractional occupancy) as a function of [AP1], the data were analyzed by nonlinear, least squares regression analysis at a wide range of possible values for DG. The error associated with the fit of each of those values to experimental data is shown below; the higher the variance of fit, the larger the error.



16.5 What is the most accurate value of DG for binding of AP1 to this duplex oligonucleotide?

16.6 What is the most accurate measure of the equilibrium constant, Ks, for binding of AP1 to this duplex oligonucleotide?

For the next two problems, consider a hypothetical eubacterial operon in which the operator overlaps the -10 region of the promoter. Measurement of the lag time before production of abortive transcripts (in an abortive initiation assay) as a function of the inverse of the RNA polymerase concentration (1/[RNAP]) gave the results shown below. The filled circles are the results of the assay in the absence of repressor, and the open circles are the results in the presence of repressor bound to the operator.



16.7 What is the value of the forward rate constant (kf) for closed to open complex formation under the two different conditions?

16.8 What is the value of the equilibrium constant (KB) for binding of the RNA polymerase to the promoter under the 2 conditions?

SECTION OVERVIEW

17: Transcriptional regulation of bacteriophage lambda

Topic hierarchy

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

18: Transcriptional regulation after initiation

Although regulation of the initiation of transcription appears to be a dominant factor in control of expression of many genes, the importance of regulation after initiation is becoming better appreciated in an increasing number and variety of systems. The classic systems in which these issues have been explored are antitermination in bacteriophage l and in attenuation of transcription in bacterial biosynthetic operons, in particular the *trp* operon in *E. coli*. Although some of the mechanistic details may be peculiar to bacteria, especially the need for coupled transcription and translation in the *trp*attenuation system, the phenomenon of regulation after initiation is seen in a wide variety of organisms, ranging from bacteria to humans. Some of this work was discussed in the sections on elongation of transcription in Chapter II of Part Three.

Introduction

Both systems discussed in this chapter control the frequency of termination of transcription. Antitermination in bacteriophage l can prevent RNA polymerase from stopping at r-dependent terminators, thus leading to transcription of downstream genes. Attenuation in the trp operon also controls the frequency at which RNA polymerase stops at an early terminator in the operon, hence regulating the transcription of downstream genes. In contrast to the system in l, attenuation in trp regulates termination at a r-independentterminator.

Antitermination in Bacteriophage l

Just to quickly review one of the points in Chapter III, antitermination occurs at two different times in the llife cycle. The N protein allows read-through transcription in the shift from immediate-early to early transcription, and the Q protein allows read-through transcription of the late genes.

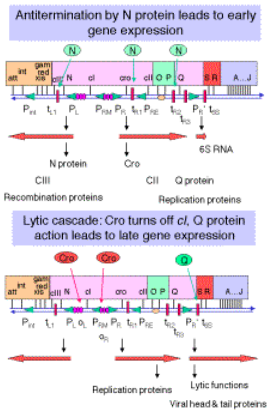


Figure 4.4.1

Recall from Part Three of the text that r-dependent terminators do not have a well-conserved sequence or secondary structure. Also, the protein r tracks along protein-free regions of the RNA until it hits a paused transcriptional complex at a r-dependent site, at which point its RNA helicase activity can cause termination and dissociation of the polymerase and transcript from the template DNA.

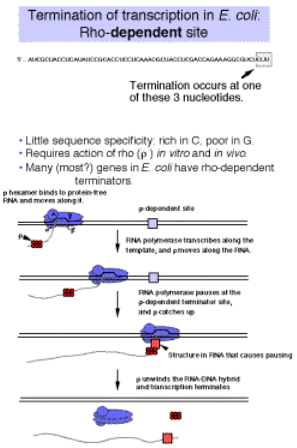


Figure 4.4.2. Action of rprotein at terminators of transcription

Sites on the DNA needed for antitermination in bacteriophage l. *nut*sites (N utilization sites) for pN, *qut*site for pQ and the *nut*sites are within the transcription unit, not at the promoter and not at the terminator.

- *nutL*is in the 5' untranslated region of the *N*gene, and *nutR*is in the 3' untranslated region of the *cro*gene.
- In both cases, the *nut*site precedes the terminator at which pN will act.

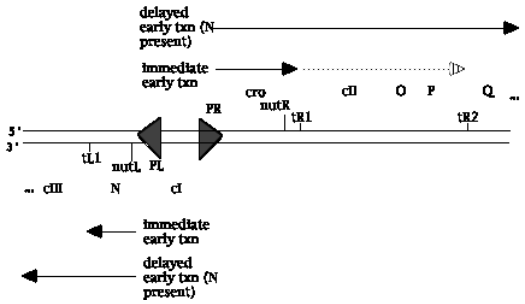


Figure 4.4.3. nut sites are located within transcription units

Both *nutL*and *nutR* are 17 bp sequences with a dyad symmetry.

5' AGCCCTGAARAAGGGCA

TCGGGACTTYTTCCCGT 5'

The protein pN recognizes the *nut*site and binds to RNA polymerase as it transcribes through the site. The complex of pN with the RNA polymerases is highly processive and overrides the efforts of r at the terminator.

E. coli (host) proteins needed for action of pN. These were isolated as host functions that when mutated prevented action of pN. NusA (encoded by *nusA*, for N utilization substance, complementation group A) is the best characterized.

1. Can form part of the transcription complex
2. Has been proposed to bind to the core RNA polymerase after s dissociates.
3. Can also bind pN.
4. Model:

NusA binds the core polymerase after s dissociates. As this complex transcribes through a *nut*site, pN binds also. The complex a2bb'-NusA-pN prevents r-dependent termination at *tR1*, *tR2*, and *tL1*.

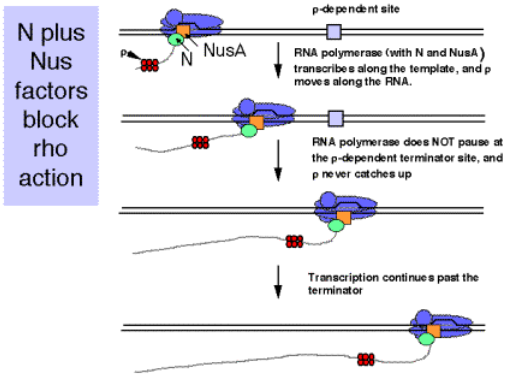


Figure 4.4.4.

Several other *nus*genes have been identified. NusG is the bacterial homolog of a family of conserved proteins involved in elongation. It is homologous to the large subunit of DSIF, which is an elongation factor in mammals. DSIF is the DRB-sensitivity inducing factor. Current studies implicate it in both negative and positive effects on elongation. It has two subunits, one of 160 kDa that is homologous to the yeast transcriptional regulatory protein Spt5, and one of 14 kDa that is homologous to the yeast Spt4 protein. Another *nus*gene encodes a ribosomal protein. Much more needs to be learned about both termination and antitermination. The *nus*phenotype of mutations in a gene encoding a ribosomal protein suggests that translation is also coupled to this process.

Components of the *E. coli trp* operon

The *trp*operon encodes the enzymes required for biosynthesis of tryptophan. More specifically, its five genes (*trpEDCBA*) encode five subunits of proteins that in total catalyze five enzymatic steps, converting chorismic acid to tryptophan. However, there is not a 1:1 correspondence between a cistron and an enzyme. For example, *trpB*and *trpA*encode, respectively, the b and a subunits of tryptophan synthase, which catalyzes the replacement of glycerol-3-phosphate from indole-3-glycerol-phosphate with serine to form tryptophan, with glyceraldehyde-3-phosphate as the other product

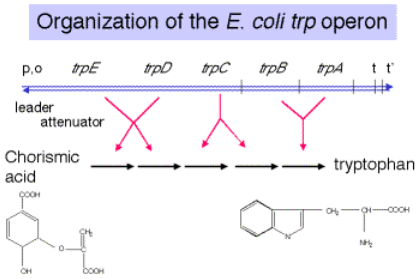


Figure 4.4.5.

A leader sequence separates the promoter and operator from the first structural gene of the operon, *trpE*. An attenuator of transcription follows the leader. As we will see in more detail below, the efficiency of "premature" termination at this attenuator is determined by the extent of translation of the leader, which in turn is determined by the availability of Trp-tRNAtrp. This is an important part of the regulation of the operon. Two terminators of transcription follow the structural genes, one dependent on r and one independent of r.

Modes of regulation: turn operon off in presence of Trp

Repressor-operator: requires a protein binding to a specific site in the presence of Trp to decrease the efficiency of initiation of transcription. Attenuation: the elongation (and termination) of transcription by RNA polymerase is linked to the progress of translation by a ribosome. In the presence of Trp, the translation by the ribosome causes transcription of the subsequent genes in the operon to terminate.

Repressor: apo-repressor and co-repressor (Trp)

The apo-repressor is encoded by *trpR*at a distant locus. The apo-repressor is a homo-tetramer. It has a high affinity for the operator only when it is bound by the amino acid Trp, which serves as a co-repressor. Thus the active repressor is a tetramer of (formerly

apo-) repressor in complex with Trp. The active repressor binds to the operator to prevent initiation of transcription. The operator overlaps the promoter, including the -10 region of the promoter. It has a dyad axis of symmetry.

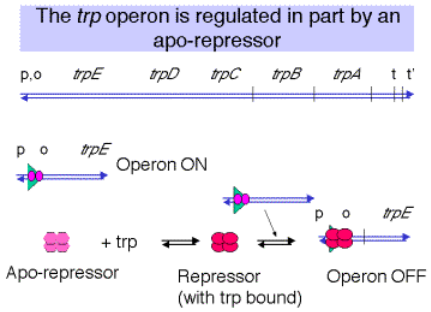


Figure 4.4.6.

Attenuation

The attenuator is a conditional transcriptional terminator used to regulate expression of biosynthetic operons in bacteria. It is upstream of the structural genes *trpEDCBA* and is a r-independent termination site. Its ability to terminate transcription is dependent on its ability to form the stem of duplex RNA that is characteristic of r-independent termination sites.

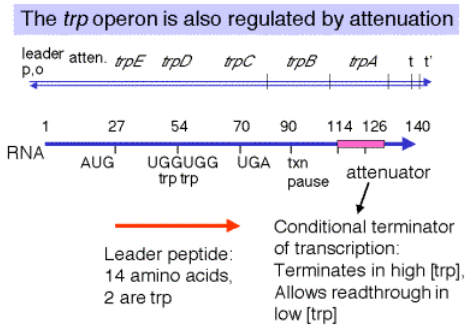


Figure 4.4.7

The fraction of transcripts that read through the attenuator is determined by the [Trp-tRNAtrp]. The concentration of charged tRNAs is a measure of the amount of Trp available for protein synthesis. If most tRNAtrp is charged, there is an abundance of Trp, and the cell does not need to make more. Low [Trp-tRNAtrp] allows read-through transcription through the attenuator, so that *trpEDCBA* is expressed and high [Trp-tRNAtrp] causes termination of transcription at the attenuator.

The [Trp-tRNAtrp] determines the progress of ribosomes as they translate a short leader peptide. The leader peptide is a short 14 amino acid polypeptide encoded by *trpL*. Two codons for Trp are in the leader, and the progress of ribosomes past these Trp codons will be determined by the availability of Trp-tRNAtrp. When the concentration of tryptophanyl-tRNA is high, translation of the *trp*leader will be completed, but when it is low, translation will stall at the tryptophan codons.

The extent of progress of the ribosomes determines the secondary structures formed in the leader RNA. When the [Trp-tRNAtrp] is high, the ribosomes translate past the Trp codons to complete the synthesis leader of the peptide. This allows the nascent RNA to form the structure for r-independent terminator. Thus transcription terminates before the RNA polymerase reaches *trpEDCBA*. When the [Trp-tRNAtrp] is low, the ribosomes stall at the Trp codons, which prevents formation of the secondary structures in the RNA necessary for termination at the attenuator. Thus read-through transcription continues through *trpEDCBA*and the operon is expressed, so that more Trp is made.

Table 4.4.1: The basic components of regulation at the attenuator of the E. coli trpoperon are tabulated below.

[trp-tRNA]	translation of trpL	secondary structures formed in RNA	Attenuator	Operon
High	complete	3-4 stem	terminate transcription	OFF
Low	stalls at Trp codons	2-3 stem	allow read-through transcription	ON

Alternative base-paired structures in leader RNA. Four regions of the leader RNA can be involved in secondary structure formation, in particular base-paired stems; these are referred to simply as regions 1, 2, 3, and 4. Potentially, 1 can pair with 2, 2 can pair with 3, and 3 can pair with 4.

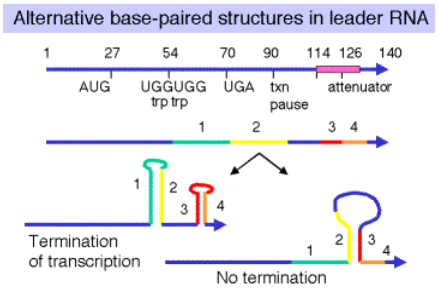


Figure 4.4.8

A stem formed by pairing between 3 and 4 makes a G+C rich stem followed by U's, which is sufficient for r-independent termination of transcription. When the [Trp-tRNAtrp] is high, the 3-4 base-paired structure forms, and transcription terminates at the attenuator. This turns the operon OFF. The formation of a base-paired stem between regions 2 and 3 precludes formation of the 3-4 terminator, and transcription will continue into the structural genes *trpEDCBA*. This turns the operon ON.

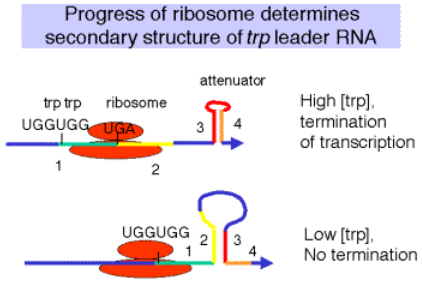


Figure 4.4.9

The choice between a 2-3 stem or a 3-4 stem is dictated by the progress of the ribosome. If the ribosome can translate past the Trp codons (when the [Trp-tRNAtrp] is high), then it will reach a natural translation termination codon. When the ribosome is in that position, region 2 of the leader RNA is covered by the ribosome, so the 2-3 stem cannot form but the 3-4 stem can. This generates the secondary structure needed for termination of transcription at the attenuator. In contrast, if the ribosome stalls at the Trp codons in the leader, because the [Trp-tRNAtrp] is low, then region 2 of the leader RNA is not covered by the ribosome. It can then base pair with region 3. This prevents formation of the 3-4 terminator, and RNA polymerase can continue elongation through *trpEDCBA*.

Mutational Analysis (selected examples)

- Translation of *trpL* is needed for regulation by attenuation. Mutation of the AUG for initiation of translation of the leader RNA prevents transcription past the attenuator. In the absence of translation, both the 1-2 and 3-4 stems can form. The latter 3-4 stem is the terminator.
- Charged tRNAtrp is required for regulation. Mutation of the genes for tRNAtrp or Trp-tRNAtrp synthetase leads to constitutive expression of *trpEDCBA*. In these mutants, translation will stall at Trp codons regardless of the intracellular [Trp], and no terminator will form at the attenuator.
- Specific secondary structures in the *trp* leader RNA are needed for regulation. E.g. mutations that decrease the number of base pairs between the 3 and 4 regions will decrease the amount of transcriptional termination (i.e. increase expression of the operon). Compensatory mutations that increase the number of base pairs between 3 and 4 will suppress the original mutations.

Attenuation requires coupled transcription and translation

Requires no regulatory proteins: charging of cognate tRNA is the regulatory signal. Need a transcriptional pause site at +90 to allow the ribosomes to catch up with the RNA polymerase and thereby affect the secondary structures in the nascent RNA.

Attenuation is a common mechanism for regulating biosynthetic operons

Many operons that encode the enzymes catalyzing biosynthesis of amino acids are regulated by attenuation. In each case, the leader polypeptide is rich in the amino acid that is the product of the pathway, e.g. *his*, *phe*, *leu*, *thr*, *ilv*.

Additional readings

- Friedman, D.I. and Count, D.L. (1995) Transcriptional antitermination: The lambda paradigm updated. Molecular Microbiology 18: 191-200.
- Henkin, T. (2000) Transcriptional termination in bacteria. Current Opinions in Microbiology 3: 149-153.
- Gusarov, I. and Nudler, E. (2001) Control of intrinsic transcriptional termination by N and NusA: The basic mechanism. Cell 107: 437-449.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

18.E: Transcriptional regulation after initiation (Exercises)

18.1 Which of the following statements concerning the action of N protein are true?

1. N action requires sequences on thel DNA called *nut*L+ and *nut*R+.
2. N activity requires a host function encoded by *nus*A+.
3. N protein acts to promote *rho*-dependent termination.
4. N protein can relieve the polarity of certain amber mutations.

18.2 Antitermination at tL1 ofl by N protein allows read-through transcription through *int*, which encodes the integrase enzyme. However, large amounts of the Int protein are not produced lytic infection, because these transcripts continue past the r-dependent terminator tint. This allows the formation of a secondary structure in the RNA that serves as a signal for RNases to degrade the transcripts from the 3' end. Why are large amounts of Int made during lysogeny?

18.3 Sketch the RNA secondary structures in the *trp* leader/attenuator region being translated by a ribosome under conditions of low and high concentrations of tryptophan.

What determines the progress of the ribosome, and how does this affect *trp*expression?

18.4 Which of the following events occur when *E. coli* is starved for the amino acid tryptophan?

1. No tryptophanyl-tRNA is made.
2. The ribosome translates the leader peptide completely (to the UGA stop codon).
3. A G+C rich stem-loop structure forms in the nascent RNA (regions 3 and 4) at the attenuator site.
4. A step-loop structure forms in the nascent RNA (regions 2 and 3) that precludes formation of the G+C rich stem-loop at the attenuator site.
5. Transcription reads through the attenuator into *trp EDCBA*.

18.5 (POB) Transcription attenuation.

In the leader region of the *trp*mRNA, what would be the effect of:

1. Increasing the distance (number of bases) between the leader peptide gene and sequence 2?
2. Increasing the distance between sequences 2 and 3?
3. Removing sequence 4?

19: Transcriptional regulation in eukaryotes

Promoters

1. Eukaryotic genes differ in their state of expression

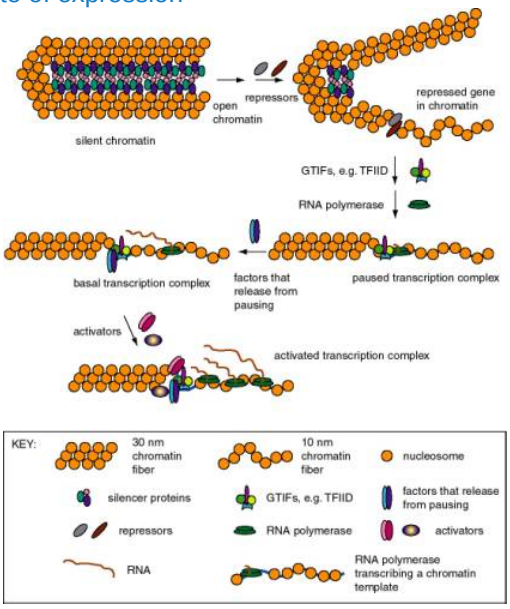


Figure 4.5.1. Expression states of promoters for RNA polymerase II. Each of these states has been described for particular genes, but it is not clear that all states are in one obligatory pathway. For instance, it possible that some gene activation events could go from silent chromatin to basal transcription without passing through open but repressed and paused transcription.

a. Basal transcription

1. Is frequently studied by *in vitro*transcription, using defined templates and either extracts from nuclei or purified components.
2. Requires RNA polymerase with general transcription factors (e.g. TFIID, TFIIA, TFIIB, TFIIE, TFIIF, and TFIIH for RNA polymerase II), as previously covered in Part Three.

b. Activated transcription

1. Occurs via transcriptional activators interacting directly or indirectly with the general transcription complex to increase the efficiency of initiation.
2. The transcriptional activators may bind to specific DNA sequences in the upstream promoter elements, or they may bind to enhancers (see Section B below).
3. The basic idea is to increase the local concentration of the general transcription factors so the initiation complex can be assembled more readily. The fact that the activators are bound to DNA that is close to the target (or becomes close because of looping of the DNA) means that the local concentration of that protein is high, and hence it can boost the local concentration of the interacting general transcription factors.

3. Stalled polymerases

RNA polymerase will transcribe about 20 to 40 nucleotides at the start of some genes and then stall at a pause site. The classic example are heat-shock genes in Drosophila, but other cases are also known. These genes are activated by release of stalled polymerases to elongate. In the case of the heat shock genes, this requires heat shock transcription factor (HSTF). The mechanism is still under study; some interesting *ideas* are:

1. Phosphorylation of the CTD of the large subunit of RNA polymerase II causes release to elongation ("promoter clearance"). One candidate (but not the only one) for the CTD kinase is TFIIH.
2. Addition of a processivity factor (analogous to *E.* coliNus A?), maybe TFIIS.

B. Silencers

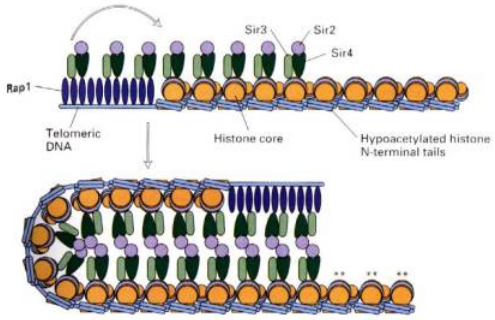


Figure 4.5.2. Transcriptionally silent chromatin, mediated by Rap1 and SIR proteins.

C. Enhancers

1. Enhancers are **cis-acting regulatory sequences that increase level of expression of a gene**, but they operate ***independently*****of position and orientation**. These last two operational criteria distinguish enhancers from promoters.
2. **Examples**

a. **SV40 control region**

1. SV40 (simian virus 40) infects monkey kidney cells, and it will also cause transformation of rodent cells. It has a double stranded DNA genome of about 5 kb. Because of its involvement in tumorigenesis, it has been a favorite subject of molecular virologists. The early region encodes tumor antigens (T-Ag and t-Ag) with many functions, including stimulating DNA replication of SV40 and blocking the action of endogenous tumor suppressors like p53 (the 1993 "Molecule of the Year"). The late region encodes three capsid proteins called VP1, VP2 and VP3 (viral protein n). A region between the early and late genes controls both replication and transcription of both classes of genes.
2. The control region has an origin of replication with binding sites for T-Ag.

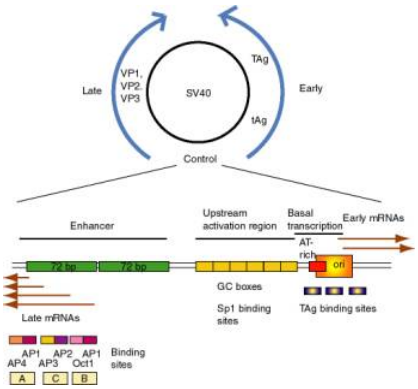


Figure 4.5.4

a. Wild type SV40 expresses T-Ag upon infection of monkey cells and lyses infected cells. However, a viral strain lacking the 72 bp repeats shows a highly reduced level of T-Ag and rarely lyses infected cells.
b. If the 72 bp repeats are added back to the mutant SV40 genome, except they are placed between the ends of the early and late genes (180° from their wild-type position), T-Ag is expressed at a high level and one obtains productive infections.
c. If the orientation of the 72 bp repeats is reversed, one still gets high level expression of viral genes and productive infection. In fact, it is needed for expression of the late genes in the wild-type, which are transcribed in the opposite direction from the early genes.
d. One concludes that the enhancer is needed for efficient transcription of the target promoters, but it can act in either orientation and at a variety of different positions and distances from the targets.
e. Work done virtually concurrently with that described above showed that the 72 bp repeats work on other "heterologous" genes, so that, for example b-globin genes could be expressed in nonerythroid cells. In fact this was one of the key observations in the discovery of the enhancer.
f. One copy of the 72 bp region will work as an enhancer, but two copies work better.

b. Immunoglobulin genes

1. This was the first enhancer of a cellular gene discovered. Researchers noted that a region of the intron was exceptionally well conserved among human, rabbit and mouse sequences, and subsequent deletion experiments showed that the intronic enhancer was required for expression.
2. After rearrangement of the immunoglobulin gene to fuse VDJ regions, one is left with a large intron between this combined variable region gene and the constant region. An enhancer is found in that intron, and another enhancer is found 3' to the polyA addition site.

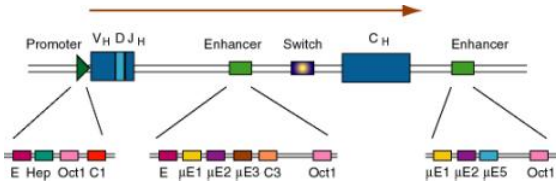


Figure 4.5.5. Enhancers in the intron and 3' flank of an immunoglobulin gene.

(3) The enhancers have multiple binding sites for transcriptional regulatory proteins

(a) Several of these sites are named for the enhancer they were discovered in. E.g. mE1, mE2, etc. are binding sites for enhancer proteins identified in the gene for the immunoglobulin heavy chain m (mu).

The protein YY1 (ying yang 1) binds to the mE1 site (CCAT is the core of the consensus) and bends DNA there.

The octamer site (ATTTGCAT) is bound by two related proteins. Oct1 is found in all tissues examined, whereas Oct2 is lymphoid specific - the first example of a tissue-specific transcription factor. Transcriptional activators that do not have their own DNA binding sequence, like VP16 from Herpes virus, will bind to Oct proteins, which bind to DNA, and the complex can activate transcription.

(b) Some proteins will bind to sites both in the promoter and the enhancer, e.g. Oct proteins. Remember Oct1 also acts at the SV40 enhancer.

c. Summary

1. The position of the enhancer can be virtually anywhere relative to the gene, but the promoter is always at the 5' end.
2. Examples are known of enhancers 5' to the gene (upstream), adjacent to the promoter (like in SV40), downstream from the gene (some globin genes), within the gene (immunoglobulins) or far upstream within a locus control region (globin genes, see Chapter 20.)

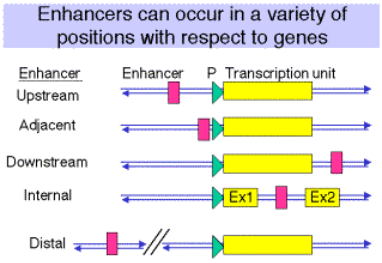


Figure 4.5.6.

3. **Multiple binding sites for transcriptional activators**

a. All enhancers characterized thus far have multiple binding sites for activator proteins.

b. Multiples of binding sites are ***needed*** for function of the enhancer.

1. In experiments with the SV40 enhancer, it was noted that some mutations that decreased the infectivity of the virus caused a mutation of one of the domains of the enhancer, e.g. domain A. When these mutants were then selected for pseudo-revertants to wild-type, with infectivity largely restored, it was found that the pseudo-revertants had duplicated one of the remaining domains. Subsequently, multimers of the various protein-binding sites were shown to be active, but monomers had little activity.

2. The domain (e.g. A, C and B in the SV40 enhancer) with at least two binding sites is called an **enhanson**. Multiple enhansons make up an enhancer.

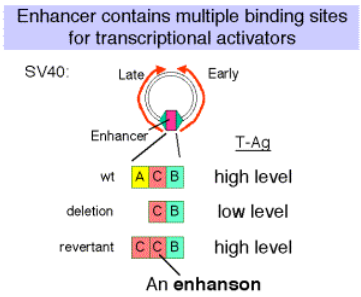


Figure 4.5.11. Two-hybrid screen for interacting proteins.

The two-hybrid screening method is a rapid and sensitive way to test a large group of proteins for their ability to interact *in* vivowith a particular protein. For example, one component of a regulatory complex may be characterized and a cDNA available. This cDNA for the “bait” protein is fused to a DNA segments encoding a well-known DNA binding domain, such as that of LexA, which binds to *lex o*. When introduced into yeast cells with the *lacZ*gene (encoding beta-galactosidase) under control of *lex o*, the *lacZ*gene is not expressed because the hybrid bait protein has no activation domain. A library of cDNAs to be tested are fused to the DNA encoding the activation domain of GAL2. When these are transformed into yeast cells carrying the hybrid LexA_DBD-bait and the *lex o - lacZ* reporter, only the hybrid proteins that interact with the bait will stimulate expression of lacZ. Transformed cells that are positive in this assay are carrying a plasmid with a hybrid gene with the cDNA encoding a protein (the “trap”) that interacts with the protein of interest (bait).

D. DNA binding domains

Computer-assisted three-dimensional views of several transcription factors, illustrating many of the domains described here, can be viewed as Chime tutorials at

- www.bmb.psu.edu/pugh/514/mdls
- **www.clunet.edu/BioDev/OMM/cro/cromast.htm**

1. Helix-turn-helix, homeodomain

1. The sequence of the "homeodomain" forms three helices separated by tight turns.
2. Helix three occupies the major groove at the binding site on the DNA. It is the recognition helix, forming specific interactions (H-bonds and hydrophobic interactions) with the edges of the base pairs in the major groove.
3. Helices one and two are perpendicular to and above helix three, providing alignment with the phosphodiester backbone. The N-terminal tail of helix interacts with the minor groove of the DNA on the opposite face of the DNA.
4. Helix two + helix three is comparable to the helix-turn-helix motif first identified in the l Cro and repressor system.

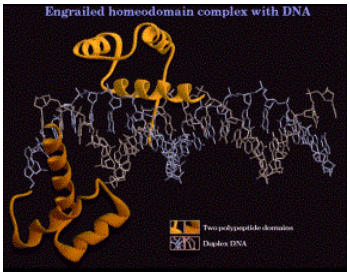


Figure 4.5.12. Helix-turn-helix in the “homeodomain”

(5) Examples

(a) Homeotic genes and their relatives.

All these are involved in regulating early developmental events in *Drosophila*. They are transcription factors (regulating the genes that determine the next developmental fate), and they have this same protein motif for their DNA binding domains.

Some specific examples are the products of these genes:

- the pair-rule gene *eve= even skipped*
- the segment polarity gene *en= engrailed*
- the homeotic gene *Antp= antennapedia*

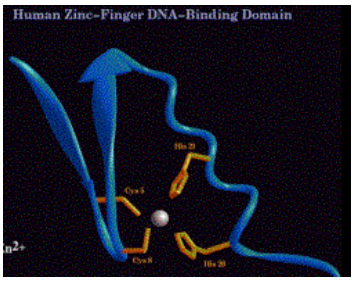


Figure 4.5.13. C_2H_2 Zn finger

(d) In a protein with 3 adjacent Zn fingers, e.g. Sp1 (remember this protein from the SV40 early promoter), each finger binds in the major groove to contact three adjacent base pairs. For the high affinity binding site, one finger contacts GGG, the next finger contacts GCG, and the remaining finger contacts GGG. So the three fingers curve along to contact the major groove for most of one turn of the helix.

(e) Members of this class of Zn finger proteins have multiple fingers, usually in a tandem array. Examples include TFIIIA (the motif was discovered in this protein) with 9 fingers, a CAC-binding protein (related to some extent to Sp1) with 3 fingers, and Drosophila ADR1 with 2 fingers.

(2)**Cys2Cys2**

(a) Consensus sequence:

Cys-X2-Cys-X1-3-Cys-X2-Cys

(b)Forms a distinctly different structure from the Cys2His2 Zn fingers.

1. Note that the number of amino acids between the 2 "halves" of the finger (1 to 3 in this case) is much less than the 12 that separate the two halves of a Cys2His2 Zn finger.
2. The Cys2Cys2 fingers are not interchangable with Cys2His2 Zn fingers in domain swap experiments.
3. The proteins do not have extensive repetitions of the motif, in contrast to proteins with Cys2His2 Zn fingers.

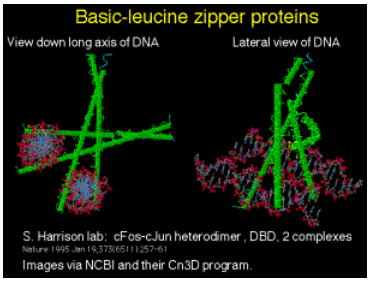


Figure 4.5.15. Basic helix-loop-helix proteins

(3) Examples include heterodimers that can exchange partners

(a) MyoD is a key protein in committment of mesodermal tissues to muscle differentiation. Other relatives, such as myogenin and myf5, are equally important and provide redundant functions. All are muscle-specific and have a similar binding domain. MyoD is active when it has E12 or E47 as its heterodimeric partner; when active it will stimulate transcription of muscle specific genes such as the one encoding creatine kinase. E12 and E47 were initially discovered as proteins that bound to enhancers of immunoglobulin genes, but are found in virtually all cell-types. Another protein, called Id, can also bind to E12 or E47 by its HLH domain. However, Id lacks a basic domain, so heterodimers with Id are not active. So the activity of bHLH proteins can be regulated by exchange of partners.

(b)A developing theme is that one of partners of a bHLH heterodimer is ubiquitous (e.g. E12, E47 in mammals, da = daughterless in Drosophila) and the other is tissue-specific (MyoD or AC-S = achaete-scute, a regulator of neurogenesis in Drosophila). The ubiquitous components may be involved in regulating a variety of other tissue-specific proteins with bHLH domains.

(c) Myc, one of many regulators of the cell cycle, is a bHLH protein. It forms partners with Max, and it is possible that this is important in regulation of the cell cycle.

E. Transcriptional activation domains

1. Acidic

This domain has been postulated to be an "acid blob" or an amphipathic helix with acidic residues on one face. Recent physico-chemical studies of GAL4 have shown b-sheet structure. At this point no single structure has been established. Examples:

GAL4 protein, VP16, GCN4, glucocorticoid hormone receptor, AP1, and the l repressor (activation of PRM).

2. Gln-rich

This domain is rich in glutamine, as its name implies. Examples of proteins containing the domain are Sp1, Antp, Oct1 and Oct2

3. Pro-rich

Again, the domain is rich in proline. Examples include CTF/NF1 (involved in regulation of replication as nuclear factor 1, and proposed to be one of many proteins binding to CCAAT motifs).

4. Work so far has not established well-defined secondary or tertiary structures for these domains.

One possibility is that the activation domains assume their proper structure after binding to its target, i.e. an induced fit model.

Table 4.5.1. Selected eukarytoic transcription factors and their properties

Name	System	Binding site (top strand)	Quaternary structure	DNA binding domain	Activation domain	Other comments
Engrailed	early development			homeodomain		
Sp1	SV40, cellular housekeeping genes	GGGGCGGGG	monomer	3 Zn fingers Cys2His2	Gln-rich	phosphoprotein
AP1	SV40, cellular enhancers	TGASTCA	heterodimer, Jun-Fos, Jun2, others	basic region + Leu zipper	acidic	regulated by phosphorylation
Oct1	lymphoid and other genes	ATTTGCAT	monomer, but can bind VP16	POU domain + homeodomain (HTH)	Gln-rich, also binds VP16	Oct1 is ubiquitous, Oct2 is lymphoid specific
GAL4	yeast galactose regulon	CGGASGACWGT CSTCCG	homodimer	Zn2Cys6, binuclear cluster	acidic	
Glucocorticoid receptor	glucocorticoid responsive genes	TGGTACAAATGT TCT	cytoplasm: with "heat shock" proteins; nucleus: homodimer	2 Zn fingers, Cys2Cys2	close to Zn finger	binding of hormone ligand changes conformation, move to nucleus and activate genes
MyoD	determination of myogenesis	CAGCTG	heterodimer with E12/E47: active; heterodimer with ID: inactive	basic-helix-loop-helix		switch partners to activate or inactivate
HMG(I)Y	interferon gene and others	minor groove	monomer (?)			bends DNA to provide favorable interactions of other proteins

VP16	Herpes simplex virus	not bind tightly to DNA	binds to proteins like Oct1		acidic activation domain; very potent	binds to other proteins that themselves bind specifically to DNA

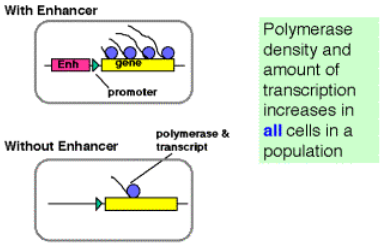


Figure 4.5.18.

a. In looping models, the activators bound to the enhancer are brought in close proximity to their targets at the promoter by forming loops in the DNA.

1. The activators can make direct contact with their target (perhaps the pre-initiation complex), or they may operate through an intermediary called a *co-activator*or *mediator*.
2. If a loop is formed, in principle it does not matter how large the loop is or if the activator binding site is 5' or 3' to the target. This could explain the ability of enhancers to operate independently of position.

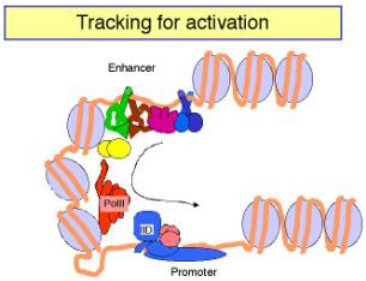


Figure 4.5.19.

c. The looping model is favored at this time. However, it has been difficult to design experiments that definitely rule out tracking. Several observations show that DNA can form loops *in vitro*, allowing contact between proteins at the enhancer and those at the promoter. For instance:

1. Using electron microscopy, one can visualize loops of DNA held together by interactions between enhancer-bound activator proteins and proteins bound to the promoter.
2. The biochemical approaches show that the activation domains of transcription factors *can*bind to components of the pre-initiation complex, such as TFIID (see Section H).

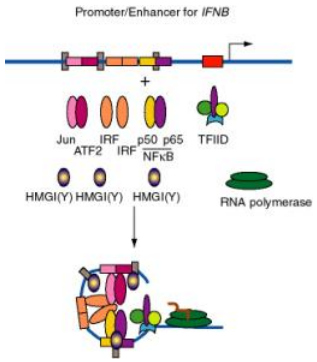


Figure 4.5.20. DNA bending by HMGI(Y) in formation of enhanceosome at **IFNBpromoter.**

b. E.g. the enhancer for the interferon-bgene, which is located just upstream from the promoter, has binding sites for three dimeric "conventional" transcription factors: NFKB (p50 + p65), IRF, and a heterodimer of ATF2 + Jun (a relative of AP1). In addition, there are three specific binding sites for HMGI(Y).

1. HMGI(Y) is a member of the "high mobility group" of nonhistone chromosomal proteins. Most HMG proteins are abundant in the nucleus, albeit not as abundant as histones.
2. HMGI(Y) binds in the minor groove of DNA and bends the DNA.
3. It also makes specific protein-protein contacts with IRF, ATF2 and NFkB, even in the absence of DNA.
4. By bending the DNA at precise positions by a defined amount, and by aiding the binding of other proteins, HMGI(Y) seems to play a critical role in assembly of the enhancer complex in juxtaposition with the promoter.
5. In general, proteins that bend DNA can be the agents that cause the looping to bring the enhancer-binding proteins in proximity to their targets.

c. Other proteins that bend DNA

cAMP-CAP (recall this from catabolite repression in E. coli), IHF = integration host factor (required for integration of l DNA to form a prophage, via a large complex called an intasome), and YY1 (ying yang 1) which has either negative or positive effects on a large variety of genes in mammals.

Contributors and Attributions

- Ross C. Hardison, T. Ming Chu Professor of Biochemistry and Molecular Biology (The Pennsylvania State University)

19.E: Transcriptional regulation in eukaryotes (Exercises)

19.1 (POB) Specific DNA binding by regulatory proteins.

A typical prokaryotic repressor protein discriminates between its specific DNA binding site (operator) and nonspecific DNA by a factor of 105 to 106. About ten molecules of the repressor per cell are sufficient to ensure a high level of repression. Assume that a very similar repressor existed in a human cell and had a similar specificity for its binding site. How many copies of the repressor would be required per cell to elicit a level of repression similar to that seen in the prokaryotic cell? (Hint: The *E. coli*genome contains about 4.7 million base pairs and the human haploid genome contains about 2.4 billion base pairs).

Use the following information for the next 3 problems. Let's imagine that part of the regulation of expression of the OB gene is mediated by a protein we will call OBF1. There is one binding site for OBF1 in the OB gene, and let's assume that is the only specific binding site in the haploid genome, or 2 specific sites in a diploid genome. The haploid human genome has about 3 x 109 bp, or 6 x 109 bp in a diploid genome. If we assume that about 33.3% of the nuclear DNA is in an accessible chromatin conformation, that means that about 2 x 109 bp of DNA are available to bind OBF1 nonspecifically.

19.2 The diameter of a mammalian nucleus is about 10 mm. If you model a nucleus as a sphere, what is its volume? What is the molar concentration of specific and nonspecific binding sites in the nucleus?

Binding of OBF1 to a specific site and to nonspecific sites is described by the following equations.

Let P = OBF1

Ds = a specific binding site in DNA

Dns = a nonspecific binding site in the genomic DNA

P + Ds

Ks = = 1011 M-1 (eqn 2)

Kns = = 105 M-1 (eqn 3)

19.3 What fraction of the OBF1 (or P in the equations) is not bound to either specific or nonspecific sites in the DNA?

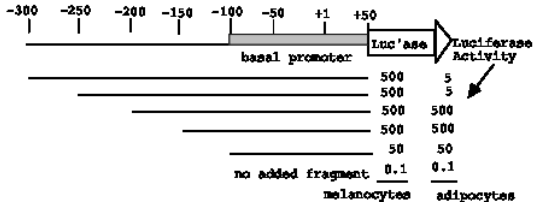
19.4 How many molecules of OBF1 are needed per nucleus to maintain 90% occupancy of the specific sites? This condition means

= 9

Use the following information for the next seven questions.

The *agouti*gene in mice controls the amount and distribution of pigments within coat hairs. Some mutations of this gene also lead to adult-onset obesity, a mild diabetes-like syndrome, tumor susceptibility and recessive embryonic lethality. The gene encodes a predicted protein of 131 amino acids that has the structural features of a secreted protein, but no striking homology to other known proteins has been recognized. This protein is likely to be a regulator of melanin pigment synthesis, and it may also be a more general metabolic regulator.

Let's suppose that you are investigating the regulation of the *agouti*gene, and have the capacity to transfect a melanocyte cell line, which transcribes the wild-type *agouti* gene, and an adipocyte cell line, which transcribes the wild-type *agouti*gene only at a very low level. Further, you already know that the basal promoter is in a DNA segment located between -100 and +50. You make progressive 5' deletions of a fragment that includes -300 to +50, link it to a luciferase reporter gene, and transfect the constructs into melanocyte and adipocyte cells, with the following results.



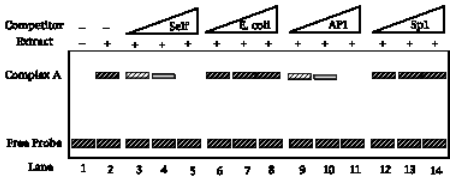
19.5. What do you conclude about the region between -250 and -200?

19.6. What do you conclude about the region between -200 and -150?

19.7. What do you conclude about the region between -150 and -100?

You also investigate the binding of nuclear proteins to these DNA segments located upstream of the *agouti*gene. Extracts containing nuclear proteins from melanocytes were tested for the ability to bind to the fragments delineated in the deletion series above.

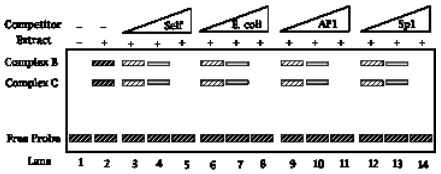
The fragment from -150 to -100 was used as the labeled probe in a mobility shift assay. The mobility of the free probe is shown in lane 1, and the pattern after binding to melanocyte nuclear extract is shown in lane 2. Lanes 3-14 show the mobility shifts after addition of the competitors to the binding reaction; the triangle above the lanes indicates that an increasing amount of competitor is used in successive lanes. "Self" is the same -150 to -100 fragment that is used as a probe, but it is unlabeled and present in an excess over the labeled probe (lanes 3-5). A completely different DNA (sheared E. coli DNA) was used as a nonspecific competitor (lanes 6-8). Two different duplex oligonucleotides, one containing the binding site for AP1 (lanes 9-11) and the other containing the binding site for Sp1 (lanes 12-14) were also tested. Thinner, less densely filled boxes denote bands of less intensity than the darker, thicker bands. Use these results to answer the next two questions.



19.8. What do you conclude from these data?

19.9. What sequence within the -150 to -100 segment might you expect to be bound in melanocyte nuclei?

19.10. The fragment from -200 to -150 was also used as a labeled probe in a mobility shift assay similar to that described for the -150 to -100 segment, as shown below.



What do you conclude from these data?

19.11. Some mutant alleles of the *agouti* gene are expressed ectopically (i.e. in the wrong tissue). Just using the information on the 5' deletions above, what region is a likely candidate for the position of a loss-of-function mutation that leads to ectopic expression in adipose tissue?

19.12 (POB) Functional domains in regulatory proteins.

A biochemist replaces the DNA-binding domain of the yeast GAL4 protein with the DNA-binding domain from the lambda repressor (CI) and finds that the engineered protein no longer functions as a transcriptional activator (it no longer regulates transcription of the *GAL*operon in yeast). What might be done to the GAL4 binding site in the DNA to make the engineered protein functional in activating *GAL*operon transcription?

19.13 What is the DNA-binding domain of the transcription factor Sp1?

19.14 What is the dimerization domain of the transcription factor AP1?

19.15 (ASC) Describe three mechanisms for regulating the activity of transcription factors.

19.16 (ASC) You have constructed a plasmid set containing a series of nucleotide insertions spaced along the length of the glucocorticoid-receptor gene. Each insertion encodes three or four amino acids. The map positions of the various insertions in the coding sequence of the receptor gene is as follows:

0 Glucocorticoid-receptor coding sequence 783

| |

| |

| |

Insertion: A B C D E F G H I J K L M N O P Q R S

The plasmids containing the receptor gene can be functionally expressed in CV-1 and COS cells, which contain a steroid-responsive gene. Using these cells, you determine the effect of each of these insertions in the receptor on the induction of the steroid-responsive gene and on binding of the synthetic steroid dexamethasone. The results of these analyses are summarized in the tablebelow.

Insertion Induction Dexamethasone binding

A ++++ ++++

B ++++ ++++

C ++++ ++++

D 0 ++++

E 0 ++++

F 0 ++++

G ++++ ++++

H ++++ ++++

I + ++++

J ++++ ++++

K 0 ++++

L 0 ++++

M 0 ++++

N + ++++

O ++++ ++++

P ++++ ++++

Q 0 0

R 0 0

S 0 0

wild-type ++++ ++++

a) From this analysis, how many different functional domains does the glucocorticoid receptor have? Indicate the position of these domains relative to the insertion map.

b) Which domain is the steroid-binding domain?

c) How could you determine which of the domains is the DNA-binding domain?

20: Transcriptional regulation via chromatin alterations

REGULATION BY CHANGES IN CHROMATIN STRUCTURE

Review of nucleosome and chromatin structure

Nucleosome composition

- Nucleosomes are the repeating subunit of chromatin.
- Nucleosomes are composed of a nucleosome core, histone H1 (in higher eukaryotes) and variable length linker DNA (0-50bp).
- The nucleosome core contains an octamer of 2 each of the core histones (H2A, H2B, H3 and H4) and 146 bp of DNA wrapped 1.75 turns (Figure 4.6.1).

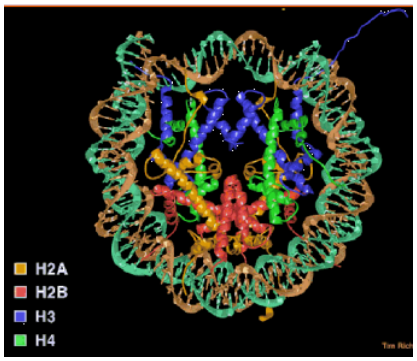


Figure 4.6.1. Nucleosome core particle. A “top” view derived from the three-dimensional structure deduced in T. Richmond’s laboratory.

Histone interactions in the nucleosome

- Core histones dimerize through their histone fold motifs generating H3/H4 dimers and H2A H2B dimers (Figure 4.6.2.).
- Two H3/H4 dimers associate to form a tetramer, which binds DNA.
- Two H2A/H2B dimers associate with the tetramer to form the histone octamer.
- At physiological salt the octamer is not stable unless bound to DNA and dissociates into the H3/H4 tetramer and two H2A/H2B dimers.

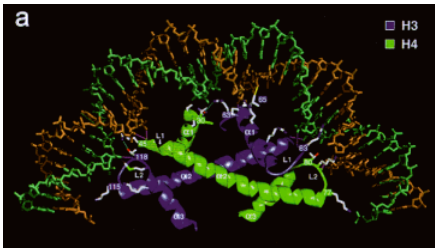


Figure 4.6.2. An H3-H4 dimer bound to DNA.

Chromatin higher order structure

- Arrays of nucleosomes condense into higher order chromatin fibers (Figure 4.6.3.).
- Despite over 2 decades of investigation the structure of the “30nm” chromatin fiber is not known.
- This may be due to irregularity or instability of the structure.
- This level of structure has been implicated in mechanisms of chromatin repression; thus, the lack of structural information at this level is particularly troublesome

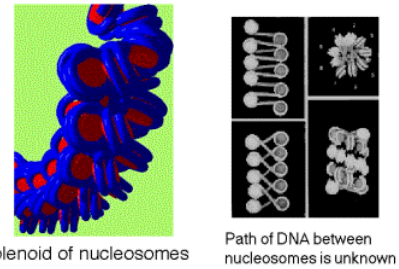


Figure 4.6.7.More open chromatin can be transcriptionally active

Biochemical investigation of different states of chromatin and gene activity in cells

Sensitivity of chromatin to nucleases

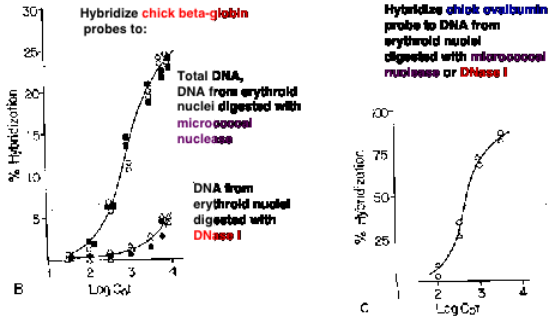


Figure 4.6.18.Domain opening is associated with movement to non-heterochromatic regions.

Proposed sequence for gene activation

1. Open a chromatin domain

    Relocate away from pericentromeric heterochromatin

    Establish a locus-wide open chromatin configuration

    General histone hyperacetylation

    DNase I sensitivity

2. **Activate transcription**

    Local hyperacetylation of histone H3

    Promoter activation to initiate and elongate transcription

Summary of cis-regulatory elements that act in chromatin

Generate an open, accessible chromatin structure

Can extend over about hundreds of kb

Can be tissue specific

Enhance expression of individual genes

Can be tissue specific

Can function at specific stages of development.

Insulate genes from position effects.

Enhancer blocking assay

How is the structure of chromatin modified in cells to change transcriptional activity?

Competition vs. Replacement models for how transcription factors occupy their binding sites on a chromatin template.

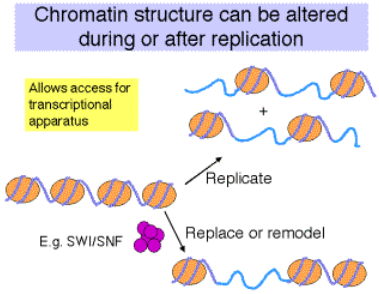


Figure 4.6.27. Model for HATs as co-activators.

The HAT complexes could be involved in other processes, or can affect them indirectly through their effects on transcription. For instance, one component of the SAGA HAT complex is Tra1, the yeast homolog of a human protein involved in cellular transformation. It may be a direct target of activator proteins.

Multiple nuclear HATs are found in yeast and in other species (Table 4.6.2). They are all large with many subunits. By comparison, their substrate, which is the nucleosome, is 0.2 MDa in mass. They have different substrate specificities. Some acetylated H3 preferentially, others acetylate H4. The reason for the diversity of HATs is a matter of current study.

Table 4.6.2. The four major nuclear HAT complexes in yeast

Complex	Mass (MDa=megadaltons)	
SAGA	1.8	
NuA4	1.4	
ADA	0.8	
NuA3	0.5	

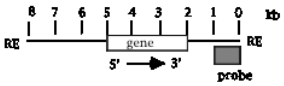
Contributors and Attributions

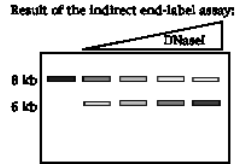
- {{template.ContribHardison()}}

20.E: Transcriptional regulation via chromatin alterations (Exercises)

Use the following information to answer the next two questions.

DNase hypersensitive sites around a gene were mapped by treating nuclei from cells that express that gene with increasing amounts of DNaseI. The partially digested DNA was isolated, cut to completion with a restriction enzyme, and analyzed by Southern blot-hybridization using a radioactive probe that is located 3' to the gene. Cleavage of genomic DNA with the restriction enzyme generates an 8 kb fragment that contains the gene, and the probe for the blot hybridization is located at the right end of the fragment (left to right defined as the direction of transcription of the gene). The results of this indirect end-labeling assay shows a gradual fade-out of the 8 kb fragment with increasing [DNaseI], and the appearance of a new band at 6 kb with DNaseI treatment.





20.1 Where is the DNase I hypersensitive site?

20.2 If the start site for transcription is 5 kb from the right end of the restriction fragment, what is a likely possibility for the function of the region mapped by the DNase hypersensitive site?

For the next three questions, consider the following information about a protein called Gcn5p. [This problem is based on Brownell et al. (1996) Cell 84: 843-851.]

[1] Gcn5p is needed for activation of some, but not all, genes in yeast.

[2] Gcn5p does not bind with high affinity to any particular site on DNA.

[3] Gcn5p will interact with acidic transcriptional activators.

[4] When incubated with histones and the following substrates, Gcn5p will have the designated effects. A + in the column under "Effect" means that the histones move slower than unmodified histones on a polyacrylamide gel that separates on the basis of charge, with the histones moving toward the negatively charged electrode. A - means that the treatment has no effect on the histones. S-adenosylmethionine is a substrate for some methyl transfer reactions, and NADH is the substrate for ADPribosyl-transferases.

Mixture	Effect
Gcn5p + histones	-
Gcn5p + histones + ATP	-
Gcn5p + histones + S-adenosylmethionine	-
Gcn5p + histones + acetyl-coenzyme A	+
Gcn5p + histones + NADH	-

20.3 What conclusion is consistent with these observations?

20.4 What enzymatic activity is associated with Gcn5p?

20.5 Which step in the gene expression pathway is likely to be regulated by Gcn5p?

20.6 What functions have been ascribed to the locus control region of mammalian beta-globin genes?

20.7 Use the following information to answer the next 6 parts (a-f) of this question. The regulatory scheme is imaginary but illustrative of some of the models we have discussed.

The protein surfactin is produced in the lung to provide surface area for efficient gas exchange in the alveoli. Let's suppose that expression of the surfactin gene is induced in lung cells by a new polypeptide hormone called pulmonin. Induction by pulmonin requires a particular DNA sequence upstream of the surfactin gene; this is called PRE for *p*ulmonin *r*esponse *e*lement. Proteins that bind specifically to that site were isolated, and the most highly purified fraction that bound to the PRE contained two polypeptides. A cDNA clone was isolated that encoded one of the polypeptides called NFL2. Antisera that specifically recognizes NFL2 is available.

The mechanism of the induction by pulmonin was investigated by testing various cell fractions (nuclear or cytoplasmic) from uninduced or pulmonin-induced lung cells in two assays. The presence or absence of NFL2 polypeptide was determined by reacting with the anti-NFL2 antisera, and the ability to bind to the PRE DNA sequence was tested by an electrophoretic mobility shift assay. In a further series of experiments, the NFL2 polypeptide was synthesized *in vitro*by transcribing the cDNA clone and translating that artificial mRNA. The product has the same amino acid sequence as the native polypeptide and is referred to below as "expressed cDNA." The expressed cDNA (which is the polypeptide synthesized *in vitro*) was tested in the same assays, before and after treatment with the cytoplasmic and nuclear extracts and also with a protein kinase that will phosphorylate the expressed cDNA on a specific serine.

Line	Source of protein and Type of treatment	React with anti-NFL2	Bind to PRE DNA
1	Uninduced cell cytoplasmic extract = unind. CE	+	-
2	Uninduced cell nuclear extract = unind. NE	-	-
3	Induced cell cytoplasmic extract = ind. CE	-	-
4	Induced cell nuclear extract = ind. NE	+	+
5	Induced cell nuclear extract + phosphatase	+	-
6	Expressed cDNA	+	-
7	Expressed cDNA + ind. CE	+	-
8	Expressed cDNA + unind. NE	+	-
9	Expressed cDNA + ind. CE + unind. NE	+	+
10	Expressed cDNA + unind. CE + unind. NE	+	-
11	Expressed cDNA + protein kinase + ATP	+	-
12	Expressed cDNA + protein kinase + ATP + unind. NE	+	+
13	Expressed cDNA + protein kinase + ATP + ind. CE	+	-

Based on these data, an affinity column was made with the expressed NFL2 cDNA as the ligand and used to test binding of proteins from nuclear extracts. When the column was pretreated with protein kinase + ATP (so that NFL2 was phosphorylated), a ubiquitous nuclear protein called UBF3 was bound from nuclear extracts from both induced and uninduced cells. If the NFL2 ligand was not phosphorylated, no binding of nuclear proteins was observed.

To confirm that NFL2 really was part of the protein complex on PRE, antibodies against NFL2 were shown to react with this protein-DNA complex. Furthermore, antibodies against phosphoserine, but not antibodies against phosphotyrosine, reacted with the specific PRE-protein complex.

Answer questions a to f based on the above observations.

a) Where is the NFL2 polypeptide? (Use data in lines 1-5.)

b) Where is the activity that will bind to the PRE site in DNA? (Use data in lines 1-5.)

c) From the data in lines 6-13, what must happen to the *in vitro*synthesized NFL2 (the expressed cDNA) in order to bind to the PRE site?

d) What proteins and covalent modifications of them are required to bind to the PRE site?

e) Which cell compartment has the protein kinase that acts on NFL2?

f) What model for pulmonin induction of the surfactin gene best fits the data given?

Index

O

operons
- 15: Positive and negative control of gene expression

origin recognition complex
- Replication in Yeast

P

Palindromes
- Control of initiation at oriC by methylation

Paralogous genes
- 4.7: Comparative Genome Analysis

PCR
- 3.5: Polymerase Chain Reaction (PCR)

phagemids
- 3.4: Introducting Recombinant DNA into Host Cells

plasmid universal cloning
- 3: Isolating and Analyzing Genes

plasmids
- 3.3: Introduction of recombinant DNA into cell and replication: Vectors

plectonemic coil
- 2.5: B-Form, A-Form, and Z-Form of DNA

polymerase chain reaction (PCR)
- 3: Isolating and Analyzing Genes

Polymerases
- Polymerases

positive gene expression
- 15: Positive and negative control of gene expression

pre-rRNA
- 12.1: Cutting and Trimming RNA

primase
- DNA Primers for Synthesis

primosome
- DNA Primers for Synthesis

proflavin
- 7.2: Reaction with Mutagens

prophage
- 3: Isolating and Analyzing Genes

pseudopalindrome
- Control of initiation at oriC by methylation

pyrimidine dimers
- 7.3: Ionizing Radiation

R

RecA
- 8.10: Synapsis and Invasion of Single Strands

RecBCD
- 8.9: Generation of Single Strands

recombinant DNA
- 3: Isolating and Analyzing Genes

recombination
- 8.2: Detecting Recombination

recombinational repair
- 7.4: Repair Mechanisms

Renaturation curves
- 4.1: Reassociation kinetics measure sequence complexity
- 4.2: Analysis of Renaturation curves with Multiple Components

replication
- Basic Mechanisms of Replication

Replication in Yeast
- Replication in Yeast

replicative recombination
- 8.1: Types and Examples of Recombination

replicator
- 6.1: The Replicon

replicon
- 6.1: The Replicon

replicores
- 4.6: Large Scale Genome Organization

restriction endonuclease
- 7.4: Repair Mechanisms

retrotransposition
- Mechanism of Retrotransposition

retrotransposons
- 4.2: Analysis of Renaturation curves with Multiple Components

reversal of damage
- 7.4: Repair Mechanisms

Reverse transcriptase
- Mechanism of Retrotransposition

RIBOZYMES
- 12.5: RNAs Can Function as Enzymes

RNA abundance
- 4.3: RNA Abundance

RuvA
- 8.11: Branch Migration

RuvB
- 8.11: Branch Migration

RuvC
- 8.12: Resolution

S

Self-splicing
- 12.4: Self-splicing by group I introns (pre-rRNA of Tetrahymena)

sense strand
- Transcription and mRNA structure

Shuttle vectors
- 3.4: Introducting Recombinant DNA into Host Cells

Slippage errors
- 7.1: Mutations and Mutagens

Southern blot
- 2: Structures of Nucleic Acids

spliceosome
- 12.6: Splicing of introns in pre-mRNAs

split genes
- 3.8: Eukaryotic Gene Structure

strand assimilation
- 8.10: Synapsis and Invasion of Single Strands

strand exchange
- 8.10: Synapsis and Invasion of Single Strands

strand invasion
- 8.6: Holliday Model for General Recombination - Single Strand Invasion

substitution mutation
- 7.1: Mutations and Mutagens

Sulfur Mustard
- Sulfur Mustard

superhelical density
- 2: Structures of Nucleic Acids

T

Taq polymerase
- 3.5: Polymerase Chain Reaction (PCR)

telomerase
- Linear Templates

telomere
- Linear Templates

tetrad analysis
- 8.5: Evidence for Heteroduplexes from Recombination in Fungi

topoisomerases
- 2: Structures of Nucleic Acids

Trans-splicing
- 12.6: Splicing of introns in pre-mRNAs

transcription
- Transcription and mRNA structure

Transducing phage
- 3: Isolating and Analyzing Genes

transformation
- 3.4: Introducting Recombinant DNA into Host Cells

Transposable Elements
- 9.1: Transposable Elements (Transposons)
- 9.2: Are Transposons Parasites or Symbionts?

transposase
- 9.6: Classes of Transposable Elements

transposition
- 9. Transposition of DNA

transposons
- 9. Transposition of DNA
- 9.1: Transposable Elements (Transposons)
- 9.6: Classes of Transposable Elements
- Mechanism of DNA-mediated transposition

transversion
- 7.1: Mutations and Mutagens

U

Unstable Alleles
- Unstable Alleles

V

vectors
- 3.3: Introduction of recombinant DNA into cell and replication: Vectors

Y

Yeast artificial chromosomes (YACs)
- 3.4: Introducting Recombinant DNA into Host Cells

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Book: Working with Molecular Genetics (Hardison)

Webpages: 145

All licenses found:

- Undeclared: 100% (145 pages)

By Page

- Book: Working with Molecular Genetics (Hardison) - *Undeclared*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - Unit I: Genes, Nucleic Acids, Genomes and Chromosomes - *Undeclared*
    - 1: Fundamental Properties of Genes - *Undeclared*
      - 1.1: Introduction to Genes - *Undeclared*
      - 1.2: Genes are the Units of Heredity: Mendel's Laws - *Undeclared*
      - 1.3: Properties of Genes - *Undeclared*
      - 1.4: Complementation and Recombination - *Undeclared*
      - 1.E: Fundamental Properties of Genes (Exercises) - *Undeclared*
      - 1.S: Additional Readings (Summary) - *Undeclared*
      - Central Dogma: DNA to RNA to protein - *Undeclared*
        - Transcription and mRNA structure - *Undeclared*
      - Finding the Functions of Genes - *Undeclared*
      - Genetic Methods in Microorganisms - *Undeclared*
        - Bacteriophage - *Undeclared*
        - Conjugation in Bacteria - *Undeclared*
        - Gene mapping by conjugal transfer - *Undeclared*
        - Genetic Methods Introduction - *Undeclared*
    - 2: Structures of Nucleic Acids - *Undeclared*
      - 2.5: B-Form, A-Form, and Z-Form of DNA - *Undeclared*
      - 2.8: Intro - *Undeclared*
      - 2:E: Structures of nucleic acids (Exercises) - *Undeclared*
    - 3: Isolating and Analyzing Genes - *Undeclared*
      - 3.1: Recombinant DNA, Polymerase Chain Reaction and Applications to Eukaryotic Gene Structure and Function - *Undeclared*
      - 3.2: Overview of Recombinant DNA Technology - *Undeclared*
      - 3.3: Introduction of recombinant DNA into cell and replication: Vectors - *Undeclared*
      - 3.4: Introducting Recombinant DNA into Host Cells - *Undeclared*
      - 3.5: Polymerase Chain Reaction (PCR) - *Undeclared*
      - 3.6: cDNA - *Undeclared*
      - 3.7: Genomic DNA clones - *Undeclared*
      - 3.8: Eukaryotic Gene Structure - *Undeclared*
      - 3.9: Introns and Exons - *Undeclared*
      - 3.10: Functional analysis of isolated genes - *Undeclared*
      - 3.E: Isolating and Analyzing Genes (Exercises) - *Undeclared*
    - 4: Genomes and Chromosomes - *Undeclared*
      - 4.1: Reassociation kinetics measure sequence complexity - *Undeclared*
      - 4.2: Analysis of Renaturation curves with Multiple Components - *Undeclared*
      - 4.3: RNA Abundance - *Undeclared*
      - 4.4: Genome Analysis by Large Scale Sequencing - *Undeclared*
      - 4.5: Sizes of genomes - The C-value paradox - *Undeclared*
      - 4.6: Large Scale Genome Organization - *Undeclared*
      - 4.7: Comparative Genome Analysis - *Undeclared*
      - 4.E: Genomes and Chromosomes (Exercises) - *Undeclared*
      - 4.S: Genomes and Chromosomes (Summary) - *Undeclared*
  - Unit II: Replication, Maintenance and Alteration of the Genetic Material - *Undeclared*
    - 5. DNA replication I: Enzymes and mechanism - *Undeclared*