GENES ARE DNA

1.1.1 Introduction

Key Terms

The **genome** is the complete set of sequences in the genetic material of an organism. It includes the sequence of each chromosome plus any DNA in organelles.

Nucleic acids are molecules that encode genetic information. They consist of a series of nitrogenous bases connected to ribose molecules that are linked by phosphodiester bonds. DNA is deoxyribonucleic acid, and RNA is ribonucleic acid.

A **gene (cistron)** is the segment of DNA specifying production of a polypeptide chain; it includes regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).

An **allele** is one of several alternative forms of a gene occupying a given locus on a chromosome.

A **locus** is the position on a chromosome at which the gene for a particular trait resides; a locus may be occupied by any one of the alleles for the gene.

Linkage describes the tendency of genes to be inherited together as a result of their location on the same chromosome; measured by percent recombination between loci.

The hereditary nature of every living organism is defined by its **genome**, which consists of a long sequence of **nucleic acid** that provides the *information* need to construct the organism. We use the term "information" because the genome does not itself perform any active role in building the organism; rather it is the sequence of the individual subunits (bases) of the nucleic acid that determines hereditary features. By a complex series of interactions, this sequence is used to produce all the proteins of the organism in the appropriate time and place. The proteins either form part of the structure of the organism, or have the capacity to build the structures or to perform the metabolic reactions necessary for life.

The genome contains the complete set of hereditary information for any organism. Physically the genome may be divided into a number of different nucleic acid molecules. Functionally it may be divided into **genes**. Each gene is a sequence within the nucleic acid that represents a single protein. Each of the discrete nucleic acid molecules comprising the genome may contain a large number of genes. Genomes for living organisms may contain as few as <500 genes (for a mycoplasma, a type of bacterium) to as many as >40,000 for Man.

In this Chapter, we analyze the properties of the gene in terms of its basic molecular construction. **Figure 1.1** summarizes the stages in the transition from the historical concept of the gene to the modern definition of the genome.

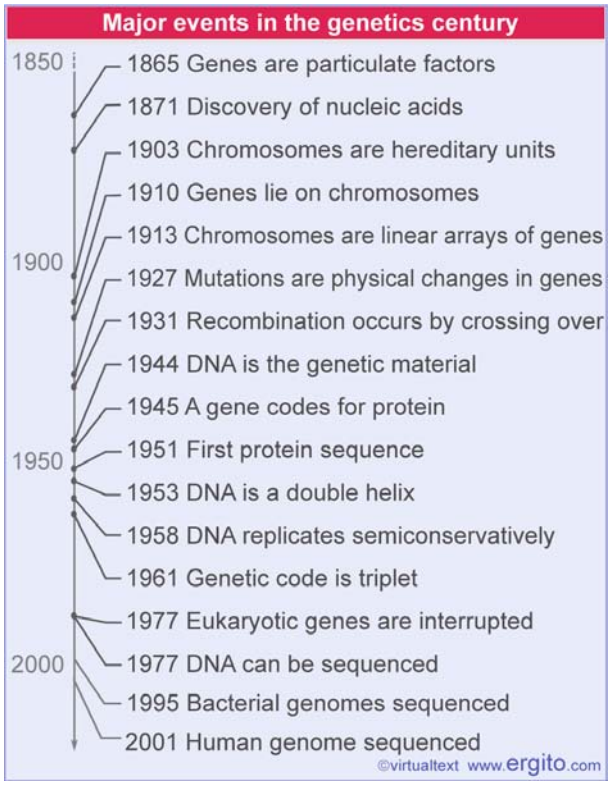

Figure 1.1 A brief history of genetics.

The basic behavior of the gene was defined by Mendel more than a century ago. Summarized in his two laws, the gene was recognized as a "particulate factor" that passes unchanged from parent to progeny. A gene may exist in alternative forms. These forms are called **alleles**.

In diploid organisms, which have two sets of chromosomes, one copy of each chromosome is inherited from each parent. This is the same behavior that is displayed by genes. One of the two copies of each gene is the paternal allele (inherited from the father), the other is the maternal allele (inherited from the mother). The equivalence led to the discovery that chromosomes in fact carry the genes.

Each chromosome consists of a linear array of genes. Each gene resides at a particular location on the chromosome. This is more formally called a genetic **locus**. We can then define the alleles of this gene as the different forms that are found at this locus.

The key to understanding the organization of genes into chromosomes was the discovery of genetic **linkage**. This describes the observation that alleles on the same chromosome tend to remain together in the progeny instead of assorting independently as predicted by Mendel's laws (see *Molecular Biology Supplement 32.3 Linkage and mapping*). Once the unit of recombination (reassortment) was introduced as the measure of linkage, the construction of genetic maps became possible.

On the genetic maps of higher organisms established during the first half of this century, the genes are arranged like beads on a string. They occur in a fixed order, and genetic recombination involves transfer of corresponding portions of the string between homologous chromosomes. The gene is to all intents and purposes a mysterious object (the bead), whose relationship to its surroundings (the string) is unclear.

The resolution of the recombination map of a higher eukaryote is restricted by the small number of progeny that can be obtained from each mating. Recombination occurs so infrequently between nearby points that it is rarely observed between different mutations in the same gene. By moving to a microbial system in which a very large number of progeny can be obtained from each genetic cross, it became possible to demonstrate that recombination occurs within genes. It follows the same rules that were previously deduced for recombination between genes.

Mutations within a gene can be arranged into a linear order, showing that the gene itself has the same linear construction as the array of genes on a chromosome. So the genetic map is linear within as well as between loci: it consists of an unbroken sequence within which the genes reside. This conclusion leads naturally into the modern view that the genetic material of a chromosome consists of an uninterrupted length of DNA representing many genes.

A genome consists of the entire set of chromosomes for any particular organism. It therefore comprises a series of DNA molecules (one for each chromosome), each of which contains many genes. The ultimate definition of a genome is to determine the sequence of the DNA of each chromosome.

The first definition of the gene as a functional unit followed from the discovery that individual genes are responsible for the production of specific proteins. The difference in chemical nature between the DNA of the gene and its protein product led to the concept that a gene *codes* for a protein. This in turn led to the discovery of the complex apparatus that allows the DNA sequence of gene to generate the amino acid sequence of a protein.

Understanding the process by which a gene is expressed allows us to make a more rigorous definition of its nature. **Figure 1.2** shows the basic theme of this book. A gene is a sequence of DNA that produces another nucleic acid, RNA. The DNA has two strands of nucleic acid, and the RNA has only one strand. The sequence of the RNA is determined by the sequence of the DNA (in fact, it is identical to one of the DNA strands). In many, but not in all cases, the RNA is in turn used to direct production of a protein. *Thus a gene is a sequence of DNA that codes for an RNA; in protein-coding genes, the RNA in turn codes for a protein.*

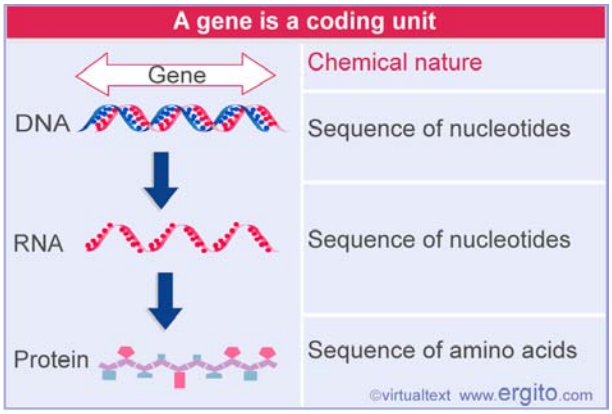

Figure 1.2 A gene codes for an RNA, which may code for protein.

From the demonstration that a gene consists of DNA, and that a chromosome consists of a long stretch of DNA representing many genes, we move to the overall organization of the genome in terms of its DNA sequence. In *Molecular Biology 1.2 The interrupted gene* we take up in more detail the organization of the gene and its representation in proteins. In *Molecular Biology 1.3 The content of the genome* we consider the total number of genes, and in *Molecular Biology 1.4 Clusters and repeats* we discuss other components of the genome and the maintenance of its organization (for review see 1; 2; 5).

Last updated on July 18, 2002

Reviews

1. Cairns, J., Stent, G., and Watson, J. D. (1966). *Phage and the origins of molecular biology*. Cold Spring Harbor Symp. Quant. Biol..

2. Olby, R. (1974). . The Path to the Double Helix.

5. Judson, H. (1978). . The Eighth Day of Creation.

GENES ARE DNA

1.1.2 DNA is the genetic material of bacteria

Key Terms

Transformation of bacteria is the acquisition of new genetic material by incorporation of added DNA.

Avirulent mutants of a bacterium or virus have lost the capacity to infect a host productively, that is, to make more bacterium or virus.

The **transforming principle** is DNA that is taken up by a bacterium and whose expression then changes the properties of the recipient cell.

Deoxyribonucleic acid (DNA) is a nucleic acid molecule consisting of long chains of polymerized (deoxyribo)nucleotides. In double-stranded DNA the two strands are held together by hydrogen bonds between complementary nucleotide base pairs.

Key Concepts

- Bacterial transformation provided the first proof that DNA is the genetic material. Genetic properties can be transferred from one bacterial strain to another by extracting DNA from the first strain and adding it to the second strain.

The idea that genetic material is nucleic acid had its roots in the discovery of **transformation** in 1928. The bacterium *Pneumococcus* kills mice by causing pneumonia. The virulence of the bacterium is determined by its *capsular polysaccharide*. This is a component of the surface that allows the bacterium to escape destruction by the host. Several types (I, II, III) of *Pneumococcus* have different capsular polysaccharides. They have a smooth (S) appearance.

Each of the smooth *Pneumococcal* types can give rise to variants that fail to produce the capsular polysaccharide. These bacteria have a *rough* (R) surface (consisting of the material that was beneath the capsular polysaccharide). They are **avirulent**. They do not kill the mice, because the absence of the polysaccharide allows the animal to destroy the bacteria.

When smooth bacteria are killed by heat treatment, they lose their ability to harm the animal. But inactive heat-killed S bacteria and the ineffectual variant R bacteria together have a quite different effect from either bacterium by itself. **Figure 1.3** shows that when they are jointly injected into an animal, the mouse dies as the result of a *Pneumococcal* infection. Virulent S bacteria can be recovered from the mouse postmortem.

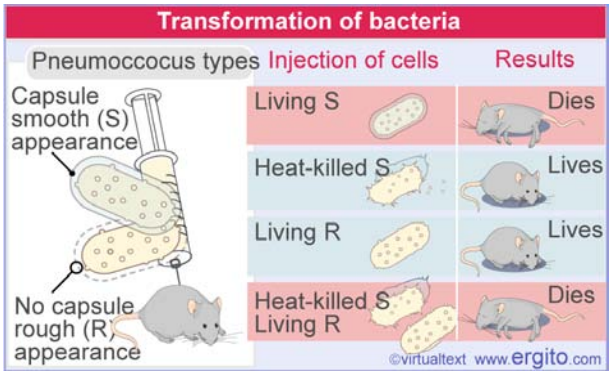

Figure 1.3 Neither heat-killed S-type nor live R-type bacteria can kill mice, but simultaneous injection of both can kill mice just as effectively as the live S-type.

In this experiment, the dead S bacteria were of type III. The live R bacteria had been derived from type II. The virulent bacteria recovered from the mixed infection had the smooth coat of type III. So some property of the dead type III S bacteria can *transform* the live R bacteria so that they make the type III capsular polysaccharide, and as a result become virulent (371).

Figure 1.4 shows the identification of the component of the dead bacteria responsible for transformation. This was called the **transforming principle**. It was purified by developing a cell-free system, in which extracts of the dead S bacteria could be added to the live R bacteria before injection into the animal. Purification of the transforming principle in 1944 showed that it is **deoxyribonucleic acid** (**DNA**) (372).

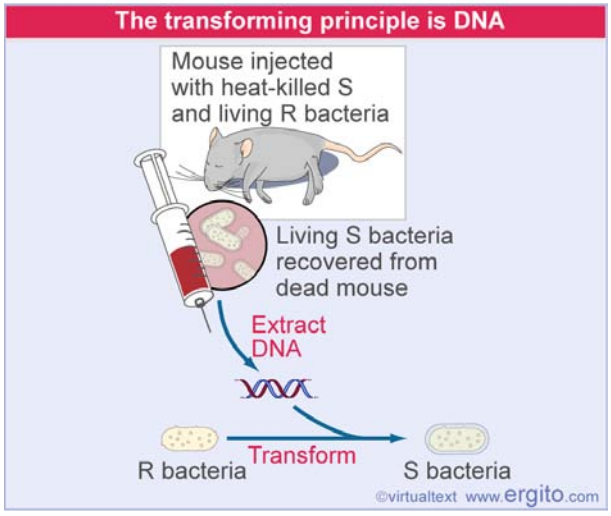

Figure 1.4 The DNA of S-type bacteria can transform R-type bacteria into the same S-type.

References

371. Griffith, F. (1928). *The significance of pneumococcal types*. J. Hyg. 27, 113-159.

372. Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). *Studies on the chemical nature of the substance inducing transformation of pneumococcal types*. J. Exp. Med. 98, 451-460.

GENES ARE DNA

1.1.3 DNA is the genetic material of viruses

Key Concepts

- Phage infection proved that DNA is the genetic material of viruses. When the DNA and protein components of bacteriophages are labeled with different radioactive isotopes, only the DNA is transmitted to the progeny phages produced by infecting bacteria.

Having shown that DNA is the genetic material of bacteria, the next step was to demonstrate that DNA provides the genetic material in a quite different system. Phage T2 is a virus that infects the bacterium *E. coli.* When phage particles are added to bacteria, they adsorb to the outside surface, some material enters the bacterium, and then ~20 minutes later each bacterium bursts open (lyses) to release a large number of progeny phage.

Figure 1.5 illustrates the results of an experiment in 1952 in which bacteria were infected with T2 phages that had been radioactively labeled *either* in their DNA component (with ^{32}P) *or* in their protein component (with ^{35}S). The infected bacteria were agitated in a blender, and two fractions were separated by centrifugation. One contained the empty phage coats that were released from the surface of the bacteria. The other fraction consisted of the infected bacteria themselves.

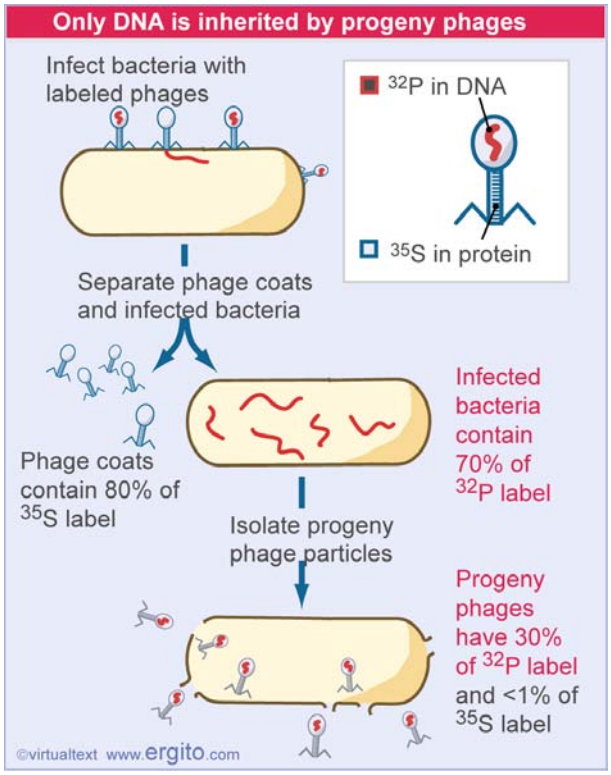

Figure 1.5 The genetic material of phage T2 is DNA.

Most of the ^{32}P label was present in the infected bacteria. The progeny phage particles produced by the infection contained ~30% of the original ^{32}P label. The progeny received very little – less than 1% – of the protein contained in the original phage population. The phage coats consist of protein and therefore carried the ^{35}S radioactive label. This experiment therefore showed directly that only the DNA of the parent phages enters the bacteria and then becomes part of the progeny phages, exactly the pattern of inheritance expected of genetic material (373).

A phage (virus) reproduces by commandeering the machinery of an infected host cell to manufacture more copies of itself. The phage possesses genetic material whose behavior is analogous to that of cellular genomes: its traits are faithfully reproduced, and they are subject to the same rules that govern inheritance. The case of T2 reinforces the general conclusion that the genetic material is DNA, whether part of the genome of a cell or virus.

References

373. Hershey, A. D. and Chase, M. (1952). *Independent functions of viral protein and nucleic acid in growth of bacteriophage*. J. Gen. Physiol. 36, 39-56.

GENES ARE DNA

1.1.4 DNA is the genetic material of animal cells

Key Terms

Transfection of eukaryotic cells is the acquisition of new genetic markers by incorporation of added DNA.

Key Concepts

- DNA can be used to introduce new genetic features into animal cells or whole animals.
- In some viruses, the genetic material is RNA.

When DNA is added to populations of single eukaryotic cells growing in culture, the nucleic acid enters the cells, and in some of them results in the production of new proteins. When a purified DNA is used, its incorporation leads to the production of a particular protein (2486). **Figure 1.6** depicts one of the standard systems.

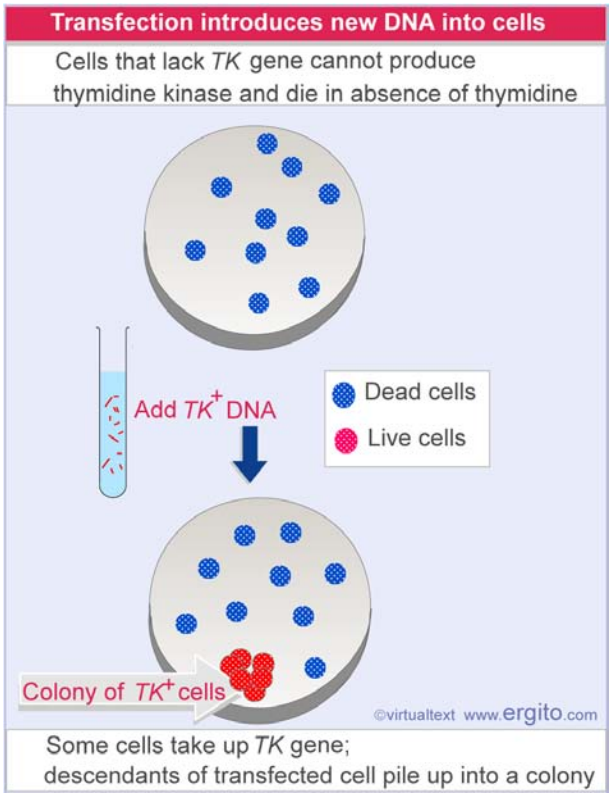

Figure 1.6 Eukaryotic cells can acquire a new phenotype as the result of transfection by added DNA.

Although for historical reasons these experiments are described as **transfection**

when performed with eukaryotic cells, they are a direct counterpart to bacterial transformation. The DNA that is introduced into the recipient cell becomes part of its genetic material, and is inherited in the same way as any other part. Its expression confers a new trait upon the cells (synthesis of thymidine kinase in the example of the figure). At first, these experiments were successful only with individual cells adapted to grow in a culture medium. Since then, however, DNA has been introduced into mouse eggs by microinjection; and it may become a stable part of the genetic material of the mouse (see *Molecular Biology 4.18.18 Genes can be injected into animal eggs*).

Such experiments show directly not only that DNA is the genetic material in eukaryotes, but also that *it can be transferred between different species and yet remain functional*.

The genetic material of all known organisms and many viruses is DNA. However, some viruses use an alternative type of nucleic acid, *ribonucleic acid (RNA)*, as the genetic material. The general principle of the nature of the genetic material, then, is that it is always nucleic acid; in fact, it is DNA except in the RNA viruses.

References

2486. Pellicer, A., Wigler, M., Axel, R., and Silverstein, S. (1978). *The transfer and stable integration of the HSV thymidine kinase gene into mouse cells.* Cell 14, 133-141.

GENES ARE DNA

1.1.5 Polynucleotide chains have nitrogenous bases linked to a sugar-phosphate backbone

Key Concepts

- A nucleoside consists of a purine or pyrimidine base linked to position 1 of a pentose sugar.
- Positions on the ribose ring are described with a prime (′) to distinguish them.
- The difference between DNA and RNA is in the group at the 2 ′ position of the sugar. DNA has a deoxyribose sugar (2 ′ –H); RNA has a ribose sugar (2 ′ –OH).
- A nucleotide consists of a nucleoside linked to a phosphate group on either the 5 ′ or 3 ′ position of the (deoxy)ribose.
- Successive (deoxy)ribose residues of a polynucleotide chain are joined by a phosphate group between the 3 ′ position of one sugar and the 5 ′ position of the next sugar.
- One end of the chain (conventionally the left) has a free 5 ′ end and the other end has a free 3 ′ end.
- DNA contains the four bases adenine, guanine, cytosine, and thymine; RNA has uracil instead of thymine.

The basic building block of nucleic acids is the nucleotide. This has three components:

- a nitrogenous base;
- a sugar;
- and a phosphate.

The nitrogenous base is a purine or pyrimidine ring. The base is linked to position 1 on a pentose sugar by a glycosidic bond from N_1 of pyrimidines or N_9 of purines. To avoid ambiguity between the numbering systems of the heterocyclic rings and the sugar, positions on the pentose are given a prime (′).

Nucleic acids are named for the type of sugar; DNA has 2 ′ –deoxyribose, whereas RNA has ribose. The difference is that the sugar in RNA has an OH group at the 2 ′ position of the pentose ring. The sugar can be linked by its 5 ′ or 3 ′ position to a phosphate group.

A nucleic acid consists of a long chain of nucleotides. **Figure 1.7** shows that the backbone of the polynucleotide chain consists of an alternating series of pentose

(sugar) and phosphate residues. This is constructed by linking the 5′ position of one pentose ring to the 3′ position of the next pentose ring via a phosphate group. So the sugar-phosphate backbone is said to consist of 5′–3′ phosphodiester linkages. The nitrogenous bases "stick out" from the backbone.

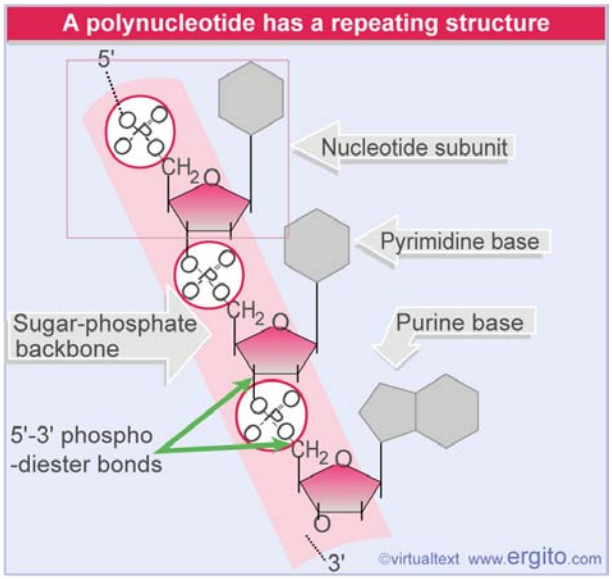

Figure 1.7 A polynucleotide chain consists of a series of 5′-3′ sugar-phosphate links that form a backbone from which the bases protrude.

Each nucleic acid contains 4 types of base. The same two purines, adenine and guanine, are present in both DNA and RNA. The two pyrimidines in DNA are cytosine and thymine; in RNA uracil is found instead of thymine. The only difference between uracil and thymine is the presence of a methyl substituent at position C_5. The bases are usually referred to by their initial letters. DNA contains A, G, C, T, while RNA contains A, G, C, U.

The terminal nucleotide at one end of the chain has a free 5′ group; the terminal nucleotide at the other end has a free 3′ group. It is conventional to write nucleic acid sequences in the 5′ → 3′ direction – that is, from the 5′ terminus at the left to the 3′ terminus at the right.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.1.5

GENES ARE DNA

1.1.6 DNA is a double helix

Key Terms

Base pairing describes the specific (complementary) interactions of adenine with thymine or of guanine with cytosine in a DNA double helix (thymine is replaced by uracil in double helical RNA).

Complementary base pairs are defined by the pairing reactions in double helical nucleic acids (A with T in DNA or with U in RNA, and C with G).

Antiparallel strands of the double helix are organized in opposite orientation, so that the 5′ end of one strand is aligned with the 3′ end of the other strand.

The **minor groove** of DNA is 12Å across.

The **major groove** of DNA is 22Å across.

A helix is said to be **right-handed** if the turns runs clockwise along the helical axis.

B-form DNA is a right-handed double helix with 10 base pairs per complete turn (360°) of the helix. This is the form found under physiological conditions whose structure was proposed by Crick and Watson.

A stretch of **overwound** DNA has more base pairs per turn than the usual average (10 bp = 1 turn). This means that the two strands of DNA are more tightly wound around each other, creating tension.

A stretch of **underwound** DNA has fewer base pairs per turn than the usual average (10 bp = 1 turn). This means that the two strands of DNA are less tightly wound around each other; ultimately this can lead to strand separation.

Key Concepts

- The B-form of DNA is a double helix consisting of two polynucleotide chains that run antiparallel.
- The nitrogenous bases of each chain are flat purine or pyrimidine rings that face inwards and pair with one another by hydrogen bonding to form A-T or G-C pairs only.
- The diameter of the double helix is 20 Å, and there is a complete turn every 34 Å, with 10 base pairs per turn.
- The double helix forms a major (wide) groove and a minor (narrow) groove.

The observation that the bases are present in different amounts in the DNAs of different species led to the concept that the *sequence of bases is the form in which genetic information is carried.* By the 1950s, the concept of genetic information was common: the twin problems it posed were working out the structure of the nucleic acid, and explaining how a sequence of bases in DNA could represent the sequence of amino acids in a protein.

Three notions converged in the construction of the double helix model for DNA by Watson and Crick in 1953:

- X-ray diffraction data showed that DNA has the form of a regular helix, making a complete turn every 34 Å (3.4 nm), with a diameter of ~20 Å (2 nm). Since the distance between adjacent nucleotides is 3.4 Å, there must be 10 nucleotides per turn.
- The density of DNA suggests that the helix must contain two polynucleotide chains. The constant diameter of the helix can be explained if the bases in each chain face inward and are restricted so that a purine is always opposite a pyrimidine, avoiding partnerships of purine-purine (too wide) or pyrimidine-pyrimidine (too narrow).
- Irrespective of the absolute amounts of each base, the proportion of G is always the same as the proportion of C in DNA, and the proportion of A is always the same as that of T. So the composition of any DNA can be described by the proportion of its bases that is G + C. This ranges from 26% to 74% for different species.

Watson and Crick proposed that the two polynucleotide chains in the double helix associate by *hydrogen bonding between the nitrogenous bases.* G can hydrogen bond specifically only with C, while A can bond specifically only with T. These reactions are described as **base pairing**, and the paired bases (G with C, or A with T) are said to be **complementary**.

The model proposed that the two polynucleotide chains to run in opposite directions (**antiparallel**), as illustrated in **Figure 1.8**. Looking along the helix, one strand runs in the $5' \rightarrow 3'$ direction, while its partner runs $3' \rightarrow 5'$ (374; 376; 375).

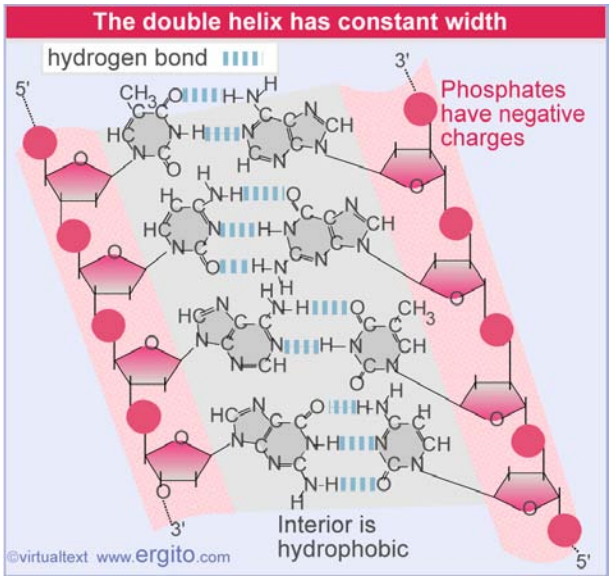

Figure 1.8 The double helix maintains a constant width because purines always face pyrimidines in the complementary A-T and G-C base pairs. The sequence in the figure is T-A, C-G, A-T, G-C.

The sugar-phosphate backbone is on the outside and carries negative charges on the phosphate groups. When DNA is in solution *in vitro*, the charges are neutralized by the binding of metal ions, typically by Na^+. In the cell, positively charged proteins provide some of the neutralizing force. These proteins play an important role in determining the organization of DNA in the cell.

The bases lie on the inside. They are flat structures, lying in pairs perpendicular to the axis of the helix. Consider the double helix in terms of a spiral staircase: the base pairs form the treads, as illustrated schematically in **Figure 1.9**. Proceeding along the helix, bases are stacked above one another, in a sense like a pile of plates.

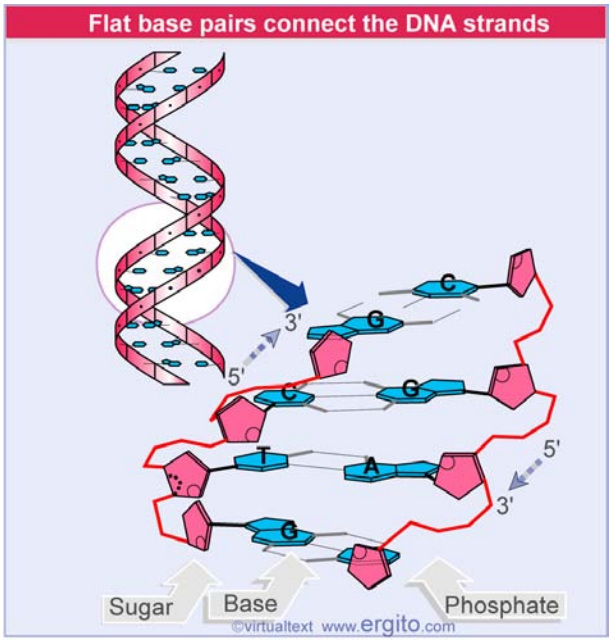

Figure 1.9 Flat base pairs lie perpendicular to the sugar-phosphate backbone.

Each base pair is rotated ~36° around the axis of the helix relative to the next base pair. So ~10 base pairs make a complete turn of 360°. The twisting of the two strands around one another forms a double helix with a **minor groove** (~12 Å across) and a **major groove** (~22 Å across), as can be seen from the scale model of **Figure 1.10**. The double helix is **right-handed**; the turns run clockwise looking along the helical axis. These features represent the accepted model for what is known as the **B-form** of DNA.

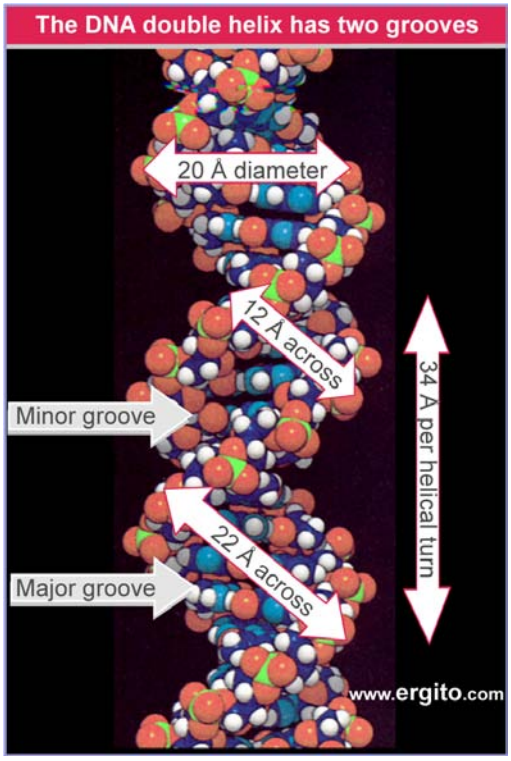

Figure 1.10 The two strands of DNA form a double helix.

It is important to realize that the B-form represents an *average*, not a precisely specified structure. DNA structure can change locally. If it has more base pairs per turn it is said to be **overwound**; if it has fewer base pairs per turn it is **underwound**. Local winding can be affected by the overall conformation of the DNA double helix in space or by the binding of proteins to specific sites.

Last updated on February 9, 2004

References

374. Watson, J. D., and Crick, F. H. C. (1953). *A structure for DNA.* Nature 171, 737-738.

375. Watson, J. D., and Crick, F. H. C. (1953). *Genetic implications of the structure of DNA.* Nature 171, 964-967.

376. Wilkins, M. F. H., Stokes, A. R., and Wilson, H. R. (1953). *Molecular structure of DNA.* Nature 171, 738-740.

GENES ARE DNA

1.1.7 DNA replication is semiconservative

Key Terms

A **parental** strand or duplex of DNA refers to the DNA that will be replicated.

The **antisense strand (Template strand)** of DNA is complementary to the sense strand, and is the one that acts as the template for synthesis of mRNA.

A **daughter** strand or duplex of DNA refers to the newly synthesized DNA.

Semiconservative replication is accomplished by separation of the strands of a parental duplex, each then acting as a template for synthesis of a complementary strand.

Key Concepts

- The Meselson-Stahl experiment used density labeling to prove that the single polynucleotide strand is the unit of DNA that is conserved during replication.
- Each strand of a DNA duplex acts as a template to synthesize a daughter strand.
- The sequences of the daughter strands are determined by complementary base pairing with the separated parental strands.

It is crucial that the genetic material is reproduced accurately. Because the two polynucleotide strands are joined only by hydrogen bonds, they are able to separate without requiring breakage of covalent bonds. The specificity of base pairing suggests that each of the separated **parental** strands could act as a **template strand** for the synthesis of a complementary **daughter** strand. **Figure 1.11** shows the principle that a new daughter strand is assembled on each parental strand. The sequence of the daughter strand is dictated by the parental strand; an *A* in the parental strand causes a *T* to be placed in the daughter strand, a parental *G* directs incorporation of a daughter *C*, and so on.

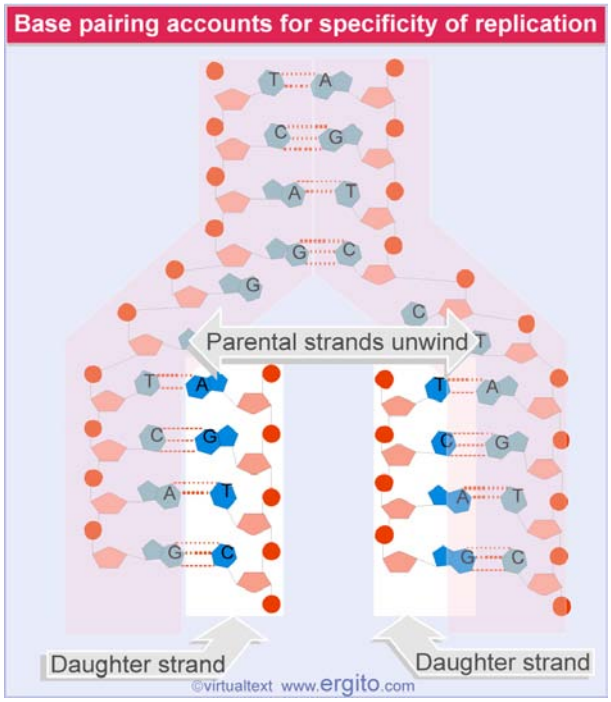

Figure 1.11 Base pairing provides the mechanism for replicating DNA.

The top part of the figure shows a parental (unreplicated) duplex that consists of the original two parental strands. The lower part shows the two daughter duplexes that are being produced by complementary base pairing. Each of the daughter duplexes is identical in sequence with the original parent, and contains one parental strand and one newly synthesized strand. *The structure of DNA carries the information needed to perpetuate its sequence.*

The consequences of this mode of replication are illustrated in **Figure 1.12**. The parental duplex is replicated to form two daughter duplexes, each of which consists of one parental strand and one (newly synthesized) daughter strand. *The unit conserved from one generation to the next is one of the two individual strands comprising the parental duplex.* This behavior is called **semiconservative replication**.

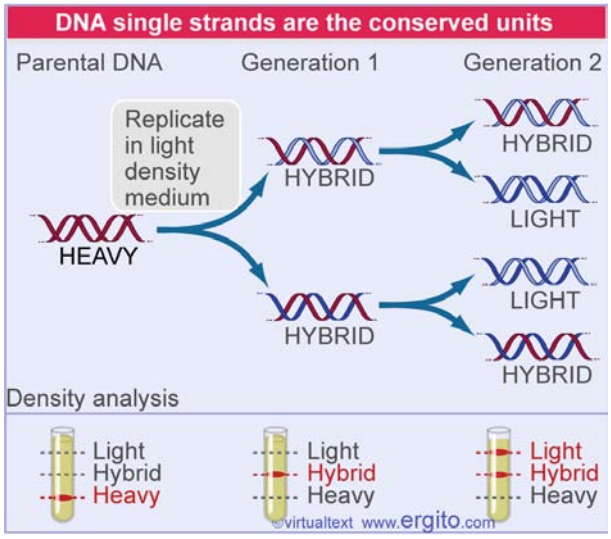

Figure 1.12 Replication of DNA is semiconservative.

The figure illustrates a prediction of this model. If the parental DNA carries a "heavy" density label because the organism has been grown in medium containing a suitable isotope (such as ^{15}N), its strands can be distinguished from those that are synthesized when the organism is transferred to a medium containing normal "light" isotopes.

The parental DNA consists of a duplex of two heavy strands (red). After one generation of growth in light medium, the duplex DNA is "hybrid" in density – it consists of one heavy parental strand (red) and one light daughter strand (blue). After a second generation, the two strands of each hybrid duplex have separated; each gains a light partner, so that now half of the duplex DNA remains hybrid while half is entirely light (both strands are blue).

The individual strands of these duplexes are entirely heavy or entirely light. This pattern was confirmed experimentally in the Meselson-Stahl experiment of 1958, which followed the semiconservative replication of DNA through three generations of growth of *E. coli.* When DNA was extracted from bacteria and its density measured by centrifugation, the DNA formed bands corresponding to its density – heavy for parental, hybrid for the first generation, and half hybrid and half light in the second generation (377; for review see 2524).

Reviews

2524. Holmes, F. (2001). . Meselson, Stahl, and the Replication of DNA: A History of The Most Beautiful Experiment in Biology.

References

377. Meselson, M. and Stahl, F. W. (1958). *The replication of DNA in E. coli*. Proc. Natl. Acad. Sci. USA 44, 671-682.

GENES ARE DNA

1.1.8 DNA strands separate at the replication fork

Key Terms

A **replication fork (Growing point)** is the point at which strands of parental duplex DNA are separated so that replication can proceed. A complex of proteins including DNA polymerase is found at the fork.

A **DNA polymerase** is an enzyme that synthesizes a daughter strand(s) of DNA (under direction from a DNA template). Any particular enzyme may be involved in repair or replication (or both).

RNA polymerases are enzymes that synthesize RNA using a DNA template (formally described as DNA-dependent RNA polymerases).

A **deoxyribonuclease (DNAase)** is an enzyme that attacks bonds in DNA. It may cut only one strand or both strands.

Ribonucleases (RNAase) are enzymes that cleave RNA. They may be specific for single-stranded or for double-stranded RNA, and may be either endonucleases or exonucleases.

Exonucleases cleave nucleotides one at a time from the end of a polynucleotide chain; they may be specific for either the 5′ or 3′ end of DNA or RNA.

Endonucleases cleave bonds within a nucleic acid chain; they may be specific for RNA or for single-stranded or double-stranded DNA.

Key Concepts

- Replication of DNA is undertaken by a complex of enzymes that separate the parental strands and synthesize the daughter strands.
- The replication fork is the point at which the parental strands are separated.
- The enzymes that synthesize DNA are called DNA polymerases; the enzymes that synthesize RNA are RNA polymerases.
- Nucleases are enzymes that degrade nucleic acids; they include DNAases and RNAases, and can be divided into endonucleases and exonucleases.

Replication requires the two strands of the parental duplex to separate. However, the disruption of structure is only transient and is reversed as the daughter duplex is formed. Only a small stretch of the duplex DNA is separated into single strands at any moment.

The helical structure of a molecule of DNA engaged in replication is illustrated in **Figure 1.10**. The nonreplicated region consists of the parental duplex, opening into the replicated region where the two daughter duplexes have formed. The double helical structure is disrupted at the junction between the two regions, which is called the **replication fork**. Replication involves movement of the replication fork along the parental DNA, so there is a continuous unwinding of the parental strands and

rewinding into daughter duplexes.

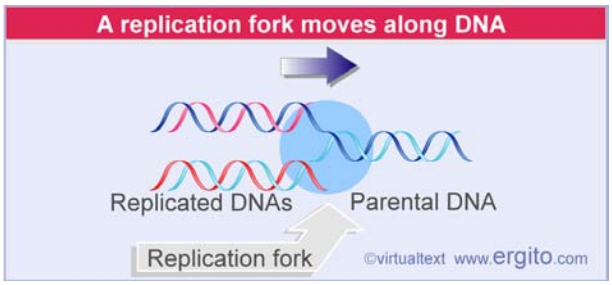

Figure 1.10 The replication fork is the region of DNA in which there is a transition from the unwound parental duplex to the newly replicated daughter duplexes.

The synthesis of nucleic acids is catalyzed by specific enzymes, which recognize the template and undertake the task of catalyzing the addition of subunits to the polynucleotide chain that is being synthesized. The enzymes are named according to the type of chain that is synthesized: **DNA polymerases** synthesize DNA, and **RNA polymerases** synthesize RNA.

Degradation of nucleic acids also requires specific enzymes: **deoxyribonucleases** (**DNAases**) degrade DNA, and **ribonucleases** (**RNAases**) degrade RNA. The nucleases fall into the general classes of **exonucleases** and **endonucleases**:

- Endonucleases cut individual bonds *within* RNA or DNA molecules, generating discrete fragments. Some DNAases cleave both strands of a duplex DNA at the target site, while others cleave only one of the two strands. Endonucleases are involved in cutting reactions, as shown in **Figure 1.11**.
- Exonucleases remove residues one at a time from the end of the molecule, generating mononucleotides. They always function on a single nucleic acid strand, and each exonuclease proceeds in a specific direction, that is, starting at either a 5′ or at a 3′ end and proceeding toward the other end. They are involved in trimming reactions, as shown in **Figure 1.12**.

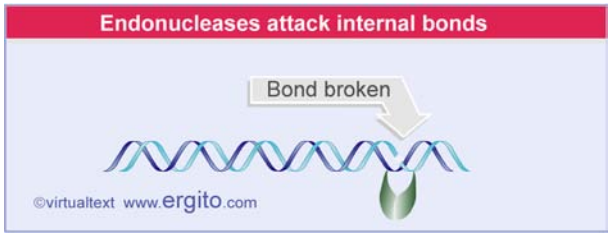

Figure 1.11 An endonuclease cleaves a bond within a nucleic acid. This example shows an enzyme that attacks one strand of a DNA duplex.

Figure 1.12 An exonuclease removes bases one at a time by cleaving the last bond in a polynucleotide chain.

Last updated on March 15, 2004

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.1.8

GENES ARE DNA

1.1.9 Nucleic acids hybridize by base pairing

Key Terms

Denaturation of protein describes its conversion from the physiological conformation to some other (inactive) conformation.

Renaturation describes the reassociation of denatured complementary single strands of a DNA double helix.

Annealing of DNA describes the renaturation of a duplex structure from single strands that were obtained by denaturing duplex DNA.

Hybridization describes the pairing of complementary RNA and DNA strands to give an RNA-DNA hybrid.

Key Concepts

- Heating causes the two strands of a DNA duplex to separate.
- The T_m is the midpoint of the temperature range for denaturation.
- Complementary single strands can renature when the temperature is reduced.
- Denaturation and renaturation/hybridization can occur with DNA-DNA, DNA-RNA, or RNA-RNA combinations, and can be intermolecular or intramolecular.
- The ability of two single-stranded nucleic acid preparations to hybridize is a measure of their complementarity.

A crucial property of the double helix is the ability to separate the two strands without disrupting covalent bonds. This makes it possible for the strands to separate and reform under physiological conditions at the (very rapid) rates needed to sustain genetic functions. The specificity of the process is determined by complementary base pairing.

*The concept of base pairing is central to all processes involving nucleic acids. Disruption of the base pairs is a crucial aspect of the function of a double-stranded molecule, while the ability to form base pairs is essential for the activity of a single-stranded nucleic acid.***Figure 1.16** shows that base pairing enables complementary single-stranded nucleic acids to form a duplex structure.

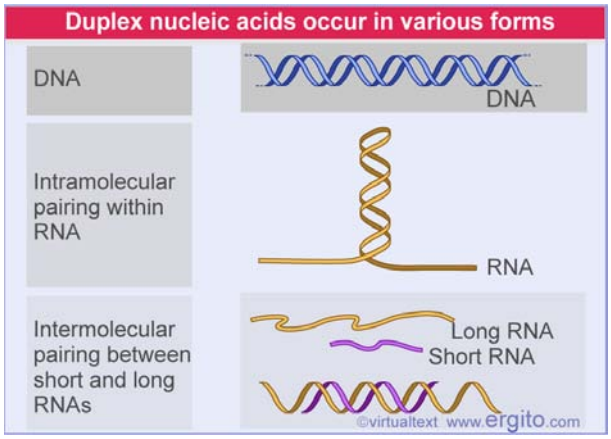

Figure 1.16 Base pairing occurs in duplex DNA and also in intra- and inter-molecular interactions in single-stranded RNA (or DNA).

- An intramolecular duplex region can form by base pairing between two complementary sequences that are part of a single-stranded molecule.
- A single-stranded molecule may base pair with an independent, complementary single-stranded molecule to form an intermolecular duplex.

Formation of duplex regions from single-stranded nucleic acids is most important for RNA, but single-stranded DNA also exists (in the form of viral genomes). Base pairing between independent complementary single strands is not restricted to DNA-DNA or RNA-RNA, but can also occur between a DNA molecule and an RNA molecule.

The lack of covalent links between complementary strands makes it possible to manipulate DNA *in vitro*. The noncovalent forces that stabilize the double helix are disrupted by heating or by exposure to low salt concentration. The two strands of a double helix separate entirely when all the hydrogen bonds between them are broken.

The process of strand separation is called **denaturation** or (more colloquially) *melting*. ("Denaturation" is also used to describe loss of authentic protein structure; it is a general term implying that the natural conformation of a macromolecule has been converted to some other form.)

Denaturation of DNA occurs over a narrow temperature range and results in striking changes in many of its physical properties. The midpoint of the temperature range over which the strands of DNA separate is called the *melting temperature* (T_m). It depends on the proportion of G·C base pairs. Because each G·C base pair has three hydrogen bonds, it is more stable than an A·T base pair, which has only two hydrogen bonds. The more G·C base pairs are contained in a DNA, the greater the energy that is needed to separate the two strands. In solution under physiological conditions, a DNA that is 40% G·C – a value typical of mammalian genomes – denatures with a T_m of about 87°C. So duplex DNA is stable at the temperature prevailing in the cell.

The denaturation of DNA is reversible under appropriate conditions. The ability of the two separated complementary strands to reform into a double helix is called **renaturation**. Renaturation depends on specific base pairing between the complementary strands. **Figure 1.17** shows that the reaction takes place in two stages. First, single strands of DNA in the solution encounter one another by chance; if their sequences are complementary, the two strands base pair to generate a short double-helical region. Then the region of base pairing extends along the molecule by a zipper-like effect to form a lengthy duplex molecule. Renaturation of the double helix restores the original properties that were lost when the DNA was denatured.

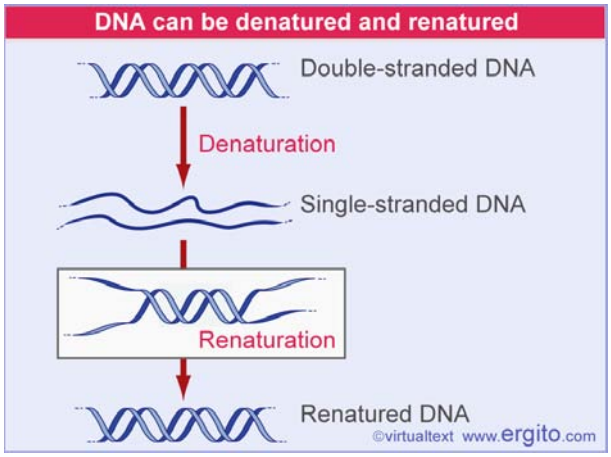

Figure 1.17 Denatured single strands of DNA can renature to give the duplex form.

Renaturation describes the reaction between two complementary sequences that were separated by denaturation. However, the technique can be extended to allow any two complementary nucleic acid sequences to react with each other to form a duplex structure. This is sometimes called **annealing**, but the reaction is more generally described as **hybridization** whenever nucleic acids of different sources are involved, as in the case when one preparation consists of DNA and the other consists of RNA. *The ability of two nucleic acid preparations to hybridize constitutes a precise test for their complementarity since* only *complementary sequences can form a duplex structure.*

The principle of the hybridization reaction is to expose two single-stranded nucleic acid preparations to each other and then to measure the amount of double-stranded material that forms. **Figure 1.18** illustrates a procedure in which a DNA preparation is denatured and the single strands are adsorbed to a filter. Then a second denatured DNA (or RNA) preparation is added. The filter is treated so that the second preparation can adsorb to it only if it is able to base pair with the DNA that was originally adsorbed. Usually the second preparation is radioactively labeled, so that the reaction can be measured as the amount of radioactive label retained by the filter.

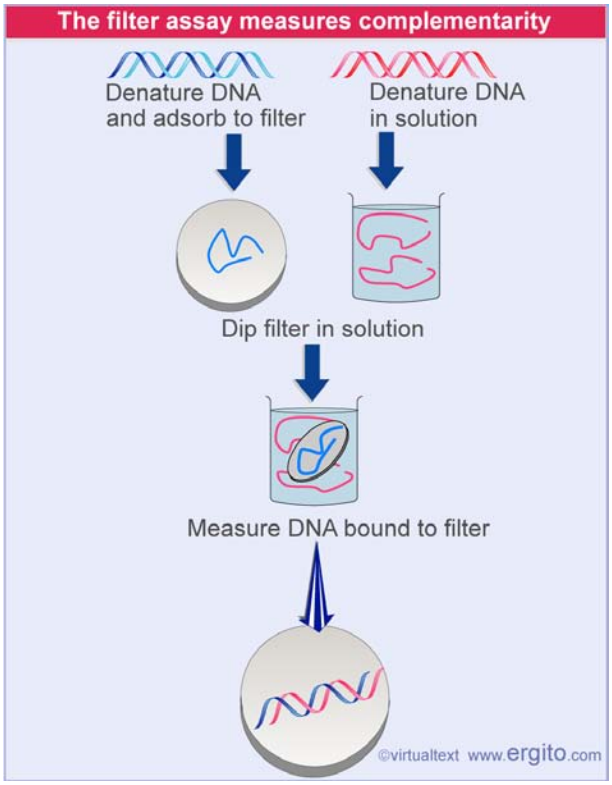

Figure 1.18 Filter hybridization establishes whether a solution of denatured DNA (or RNA) contains sequences complementary to the strands immobilized on the filter.

The extent of hybridization between two single-stranded nucleic acids is determined by their complementarity. Two sequences need not be *perfectly* complementary to hybridize. If they are closely related but not identical, an imperfect duplex is formed in which base pairing is interrupted at positions where the two single strands do not correspond.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.1.9

GENES ARE DNA

1.1.10 Mutations change the sequence of DNA

Key Terms

Spontaneous mutations occur in the absence of any added reagent to increase the mutation rate, as the result of errors in replication (or other events involved in the reproduction of DNA) or by environmental damage.

The **background level** of mutation describes the rate at which sequence changes accumulate in the genome of an organism. It reflects the balance between the occurrence of spontaneous mutations and their removal by repair systems, and is characteristic for any species.

Mutagens increase the rate of mutation by inducing changes in DNA sequence, directly or indirectly.

Induced mutations result from the action of a mutagen. The mutagen may act directly on the bases in DNA or it may act indirectly to trigger a pathway that leads to a change in DNA sequence.

Key Concepts

- All mutations consist of changes in the sequence of DNA.
- Mutations may occur spontaneously or may be induced by mutagens.

Mutations provide decisive evidence that DNA is the genetic material. When a change in the sequence of DNA causes an alteration in the sequence of a protein, we may conclude that the DNA codes for that protein. Furthermore, a change in the phenotype of the organism may allow us to identify the function of the protein. The existence of many mutations in a gene may allow many variant forms of a protein to be compared, and a detailed analysis can be used to identify regions of the protein responsible for individual enzymatic or other functions.

All organisms suffer a certain number of mutations as the result of normal cellular operations or random interactions with the environment. These are called **spontaneous mutations**; the rate at which they occur is characteristic for any particular organism and is sometimes called the **background level**. Mutations are rare events, and of course those that damage a gene are selected against during evolution. It is therefore difficult to obtain large numbers of spontaneous mutants to study from natural populations.

The occurrence of mutations can be increased by treatment with certain compounds. These are called **mutagens**, and the changes they cause are referred to as **induced mutations**. Most mutagens act directly by virtue of an ability either to modify a particular base of DNA or to become incorporated into the nucleic acid. The effectiveness of a mutagen is judged by how much it increases the rate of mutation above background. By using mutagens, it becomes possible to induce many changes in any gene (for review see 3).

Spontaneous mutations that inactivate gene function occur in bacteriophages and bacteria at a relatively constant rate of 3-4 × 10^{-3} per genome per generation (2221). Given the large variation in genome sizes between bacteriophages and bacteria, this corresponds to wide differences in the mutation rate per base pair. This suggests that the overall rate of mutation has been subject to selective forces that have balanced the deleterious effects of most mutations against the advantageous effects of some mutations. This conclusion is strengthened by the observation that an archaeal microbe that lives under harsh conditions of high temperature and acidity (which are expected to damage DNA) does not show an elevated mutation rate, but in fact has an overall mutation rate just below the average range (2203).

Figure 1.19 shows that in bacteria, the mutation rate corresponds to ~10^{-6} events per locus per generation or to an average rate of change per base pair of 10^{-9}-10^{-10} per generation. The rate at individual base pairs varies very widely, over a 10,000 fold range. We have no accurate measurement of the rate of mutation in eukaryotes, although usually it is thought to be somewhat similar to that of bacteria on a per-locus per-generation basis (2487).

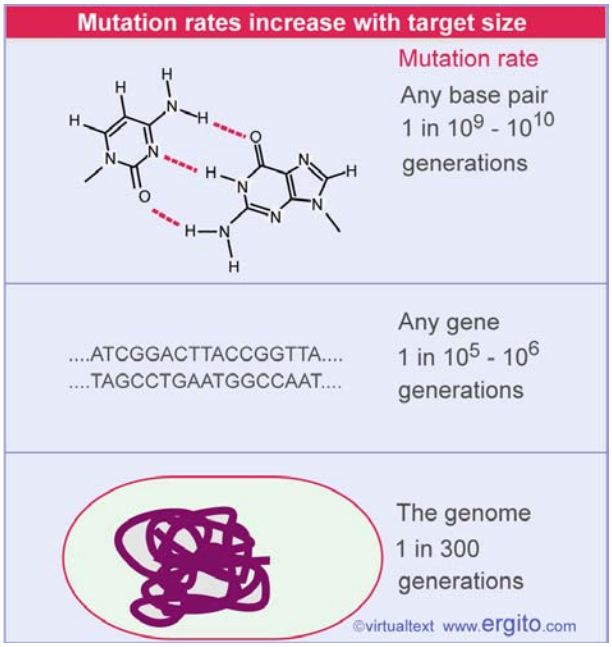

Figure 1.19 A base pair is mutated at a rate of 10^{-9} - 10^{-10} per generation, a gene of 1000 bp is mutated at ~10^{-6} per generation, and a bacterial genome is mutated at 3 × 10^{-3} per generation.

Last updated on 8-15-2002

Reviews

3. Drake, J. W. and Balz, R. H. (1976). *The biochemistry of mutagenesis.* Annu. Rev. Biochem. 45, 11-37.

2487. Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. (1998). *Rates of spontaneous mutation.* Genetics 148, 1667-1686.

References

2203. Grogan, D. W., Carver, G. T., and Drake, J. W. (2001). *Genetic fidelity under harsh conditions: analysis of spontaneous mutation in the thermoacidophilic archaeon Sulfolobus acidocaldarius.* Proc. Natl. Acad. Sci. USA 98, 7928-7933.

2221. Drake, J. W. (1991). *A constant rate of spontaneous mutation in DNA-based microbes.* Proc. Natl. Acad. Sci. USA 88, 7160-7164.

GENES ARE DNA

1.1.11 Mutations may affect single base pairs or longer sequences

Key Terms

A **point mutation** is a change in the sequence of DNA involving a single base pair.

A **transition** is a mutation in which one pyrimidine is replaced by the other and/or in which one purine is replaced by the other.

A **transversion** is a mutation in which a purine is replaced by a pyrimidine or vice versa.

Base mispairing is a coupling between two bases that does not conform to the Watson-Crick rule, e.g., adenine with cytosine, thymine with guanine.

An **insertion** is the addition of a stretch of base pairs in DNA. Duplications are a special class of insertions.

A **transposon (transposable element)** is a DNA sequence able to insert itself (or a copy of itself) at a new location in the genome, without having any sequence relationship with the target locus.

A **deletion** is the removal of a sequence of DNA, the regions on either side being joined together except in the case of a terminal deletion at the end of a chromosome.

Key Concepts

- A point mutation changes a single base pair.
- Point mutations can be caused by the chemical conversion of one base into another or by mistakes that occur during replication.
- A transition replaces a G·C base pair with an A·T base pair or vice-versa.
- A transversion replaces a purine with a pyrimidine, such as changing A·T to T·A.
- Insertions are the most common type of mutation, and result from the movement of transposable elements.

Any base pair of DNA can be mutated. A **point mutation** changes only a single base pair, and can be caused by either of two types of event (for review see 3238):

- Chemical modification of DNA directly changes one base into a different base.
- A malfunction during the replication of DNA causes the wrong base to be inserted into a polynucleotide chain during DNA synthesis.

Point mutations can be divided into two types, depending on the nature of the change when one base is substituted for another:

- The most common class is the **transition**, comprising the substitution of one pyrimidine by the other, or of one purine by the other. This replaces a G·C pair with an A·T pair or vice versa.

- The less common class is the **transversion**, in which a purine is replaced by a pyrimidine or vice versa, so that an A·T pair becomes a T·A or C·G pair.

The effects of nitrous acid provide a classic example of a transition caused by the chemical conversion of one base into another. **Figure 1.20** shows that nitrous acid performs an oxidative deamination that converts cytosine into uracil. In the replication cycle following the transition, the U pairs with an A, instead of with the G with which the original C would have paired. So the C·G pair is replaced by a T·A pair when the A pairs with the T in the next replication cycle. (Nitrous acid also deaminates adenine, causing the reverse transition from A·T to G·C.)

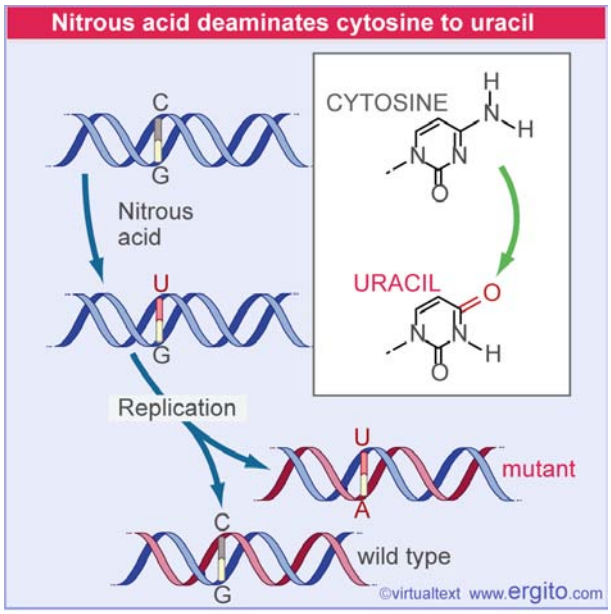

Figure 1.20 Mutations can be induced by chemical modification of a base.

Transitions are also caused by **base mispairing**, when unusual partners pair in defiance of the usual restriction to Watson-Crick pairs. Base mispairing usually occurs as an aberration resulting from the incorporation into DNA of an abnormal base that has ambiguous pairing properties. **Figure 1.21** shows the example of bromouracil (BrdU), an analog of thymine that contains a bromine atom in place of the methyl group of thymine. BrdU is incorporated into DNA in place of thymine. But it has ambiguous pairing properties, because the presence of the bromine atom allows a shift to occur in which the base changes structure from a keto (=O) form to an enol (–OH) form. The enol form can base pair with guanine, which leads to substitution of the original A·T pair by a G·C pair.

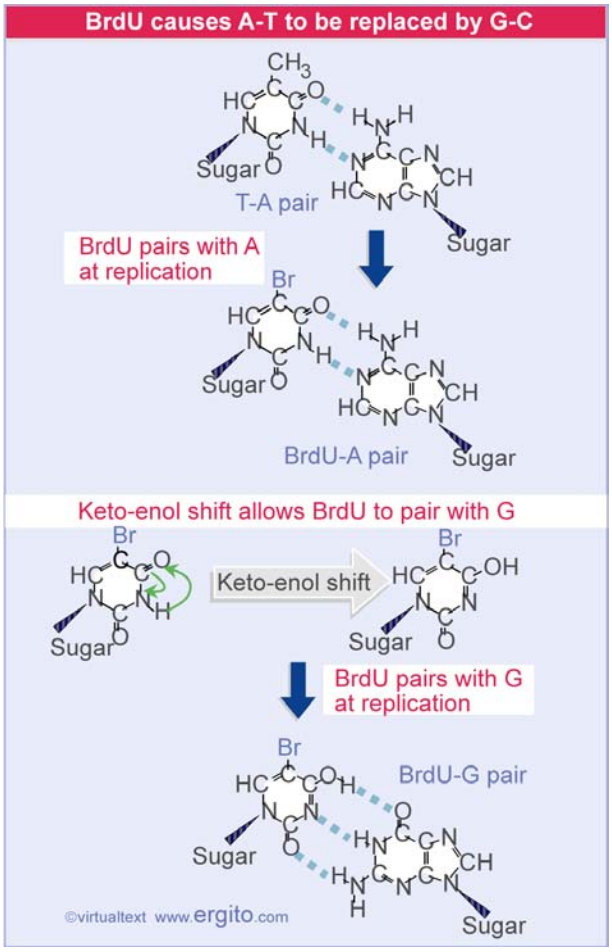

Figure 1.21 Mutations can be induced by the incorporation of base analogs into DNA.

The mistaken pairing can occur either during the original incorporation of the base or in a subsequent replication cycle. The transition is induced with a certain probability in each replication cycle, so the incorporation of BrdU has continuing effects on the sequence of DNA.

Point mutations were thought for a long time to be the principal means of change in individual genes. However, we now know that **insertions** of stretches of additional material are quite frequent. The source of the inserted material lies with **transposable elements**, sequences of DNA with the ability to move from one site to another (see *Molecular Biology 4.16 Transposons* and *Molecular Biology 4.17 Retroviruses and retroposons*.) An insertion usually abolishes the activity of a gene. Where such insertions have occurred, **deletions** of part or all of the inserted material, and sometimes of the adjacent regions, may subsequently occur.

A significant difference between point mutations and the insertions/deletions is that the frequency of point mutation can be increased by mutagens, whereas the occurrence of changes caused by transposable elements is not affected. However, insertions and deletions can also occur by other mechanisms – for example, involving mistakes made during replication or recombination – although probably

these are less common. And a class of mutagens called the acridines introduce (very small) insertions and deletions.

Reviews

3238. Maki, H. (2002). *Origins of Spontaneous Mutations: Specificity and Directionality of Base-Substitution, Frameshift, and Sequence-Substitution Mutageneses.* Annu. Rev. Genet. 36, 279-303.

GENES ARE DNA

1.1.12 The effects of mutations can be reversed

Key Terms

Revertants are derived by reversion of a mutant cell or organism to the wild-type phenotype.

Forward mutations inactivate a wild-type gene.

A **back mutation** reverses the effect of a mutation that had inactivated a gene; thus it restores wild type.

A **true reversion** is a mutation that restores the original sequence of the DNA.

Second-site reversion occurs when a second mutation suppresses the effect of a first mutation.

Suppression occurs when a second event eliminates the effects of a mutation without reversing the original change in DNA.

A **suppressor** is a second mutation that compensates for or alters the effects of a primary mutation.

Key Concepts

- Forward mutations inactivate a gene, and back mutations (or revertants) reverse their effects.
- Insertions can revert by deletion of the inserted material, but deletions cannot revert.
- Suppression occurs when a mutation in a second gene bypasses the effect of mutation in the first gene.

Figure 1.22 shows that the isolation of **revertants** is an important characteristic that distinguishes point mutations and insertions from deletions:

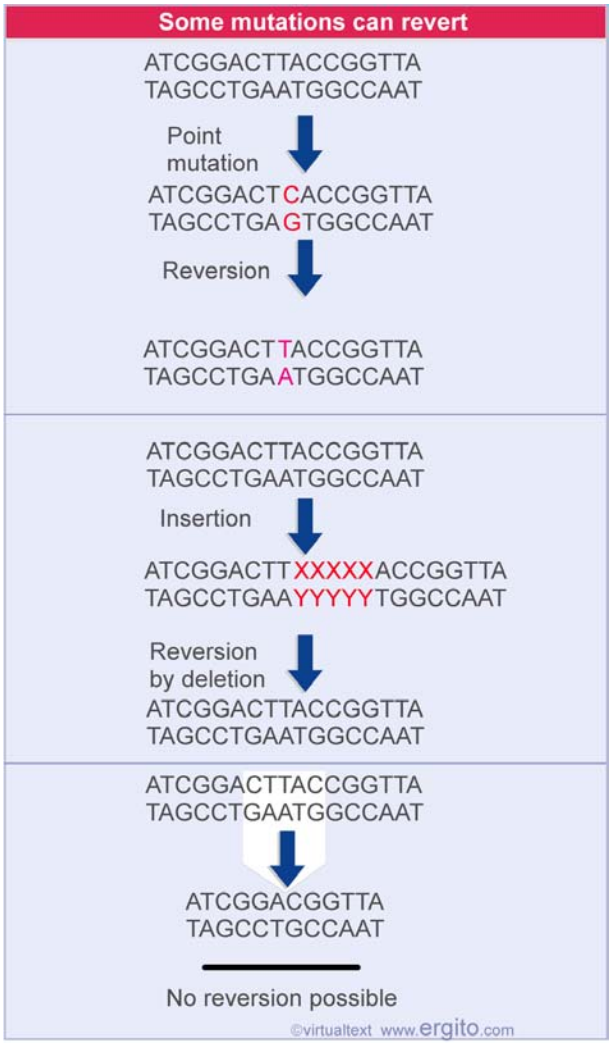

Figure 1.22 Point mutations and insertions can revert, but deletions cannot revert.

- A point mutation can revert by restoring the original sequence or by gaining a compensatory mutation elsewhere in the gene.
- An insertion of additional material can revert by deletion of the inserted material.
- A deletion of part of a gene cannot revert.

Mutations that inactivate a gene are called **forward mutations**. Their effects are reversed by **back mutations**, which are of two types.

An exact reversal of the original mutation is called **true reversion**. So if an A·T pair has been replaced by a G·C pair, another mutation to restore the A·T pair will exactly regenerate the wild-type sequence.

Alternatively, another mutation may occur elsewhere in the gene, and its effects

compensate for the first mutation. This is called **second-site reversion**. For example, one amino acid change in a protein may abolish gene function, but a second alteration may compensate for the first and restore protein activity.

A forward mutation results from any change that inactivates a gene, whereas a back mutation must restore function to a protein damaged by a particular forward mutation. So the demands for back mutation are much more specific than those for forward mutation. The rate of back mutation is correspondingly lower than that of forward mutation, typically by a factor of ~10.

Mutations can also occur in other genes to circumvent the effects of mutation in the original gene. This effect is called **suppression**. A locus in which a mutation suppresses the effect of a mutation in another locus is called a **suppressor**.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.1.12

GENES ARE DNA

1.1.13 Mutations are concentrated at hotspots

Key Terms

A **hotspot** is a site in the genome at which the frequency of mutation (or recombination) is very much increased, usually by at least an order of magnitude relative to neighboring sites.

Key Concepts

- The frequency of mutation at any particular base pair is determined by statistical fluctuation, except for hotspots, where the frequency is increased by at least an order of magnitude.

So far we have dealt with mutations in terms of individual changes in the sequence of DNA that influence the activity of the genetic unit in which they occur. When we consider mutations in terms of the inactivation of the gene, most genes within a species show more or less similar rates of mutation relative to their size. This suggests that the gene can be regarded as a target for mutation, and that damage to any part of it can abolish its function. As a result, susceptibility to mutation is roughly proportional to the size of the gene. But consider the sites of mutation within the sequence of DNA; are all base pairs in a gene equally susceptible or are some more likely to be mutated than others?

What happens when we isolate a large number of independent mutations in the same gene? Many mutants are obtained. Each is the result of an individual mutational event. Then the site of each mutation is determined. Most mutations will lie at different sites, but some will lie at the same position. Two independently isolated mutations at the same site may constitute exactly the same change in DNA (in which case the same mutational event has happened on more than one occasion), or they may constitute different changes (three different point mutations are possible at each base pair).

The histogram of **Figure 1.23** shows the frequency with which mutations are found at each base pair in the *lacI* gene of *E. coli*. The statistical probability that more than one mutation occurs at a particular site is given by random-hit kinetics (as seen in the Poisson distribution). So some sites will gain one, two, or three mutations, while others will not gain any. But some sites gain far more than the number of mutations expected from a random distribution; they may have 10× or even 100× more mutations than predicted by random hits. These sites are called **hotspots**. Spontaneous mutations may occur at hotspots; and different mutagens may have different hotspots.

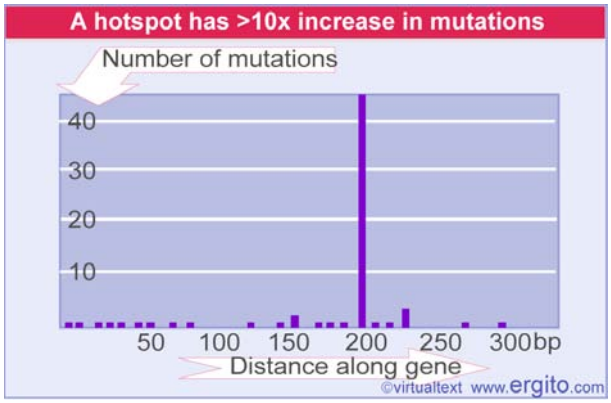

Figure 1.23 Spontaneous mutations occur throughout the *lacI* gene of *E. coli*, but are concentrated at a hotspot.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.1.13

GENES ARE DNA

1.1.14 Many hotspots result from modified bases

Key Terms

Modified bases are all those except the usual four from which DNA (T, C, A, G) or RNA (U, C, A, G) are synthesized; they result from postsynthetic changes in the nucleic acid.

A **mismatch** describes a site in DNA where the pair of bases does not conform to the usual G-C or A-T pairs. It may be caused by incorporation of the wrong base during replication or by mutation of a base.

Key Concepts

- A common cause of hotspots is the modified base 5-methylcytosine, which is spontaneously deaminated to thymine.

A major cause of spontaneous mutation results from the presence of an unusual base in the DNA. In addition to the four bases that are inserted into DNA when it is synthesized, **modified bases** are sometimes found. The name reflects their origin; they are produced by chemically modifying one of the four bases already present in DNA. The most common modified base is 5-methylcytosine, generated by a methylase enzyme that adds a methyl group to certain cytosine residues at specific sites in the DNA.

Sites containing 5-methylcytosine provide hotspots for spontaneous point mutation in *E. coli*. In each case, the mutation takes the form of a G·C to A·T transition. The hotspots are not found in strains of *E. coli* that cannot methylate cytosine.

The reason for the existence of the hotspots is that cytosine bases suffer spontaneous deamination at an appreciable frequency. In this reaction, the amino group is replaced by a keto group. Recall that deamination of cytosine generates uracil (see **Figure 1.20**). **Figure 1.24** compares this reaction with the deamination of 5-methylcytosine where deamination generates thymine. The effect in DNA is to generate the base pairs G·U and G·T, respectively, where there is a **mismatch** between the partners.

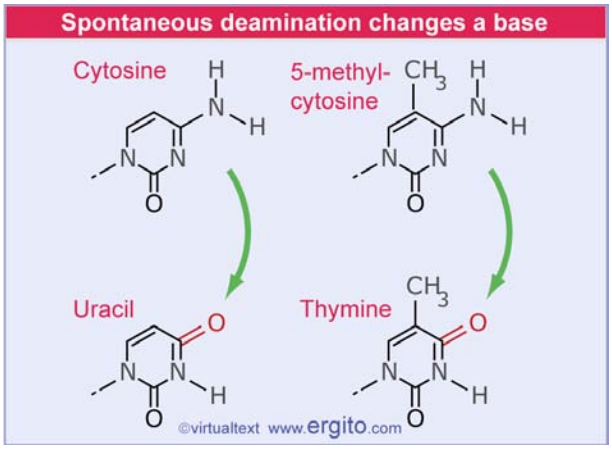

Figure 1.24 Deamination of cytosine produces uracil, whereas deamination of 5-methyl-cytosine produces thymine.

All organisms have repair systems that correct mismatched base pairs by removing and replacing one of the bases. The operation of these systems determines whether mismatched pairs such as G·U and G·T result in mutations.

Figure 1.25 shows that the consequences of deamination are different for 5-methylcytosine and cytosine. Deaminating the (rare) 5-methylcytosine causes a mutation, whereas deamination of the more common cytosine does not have this effect (382). This happens because the repair systems are much more effective in recognizing G·U than G·T.

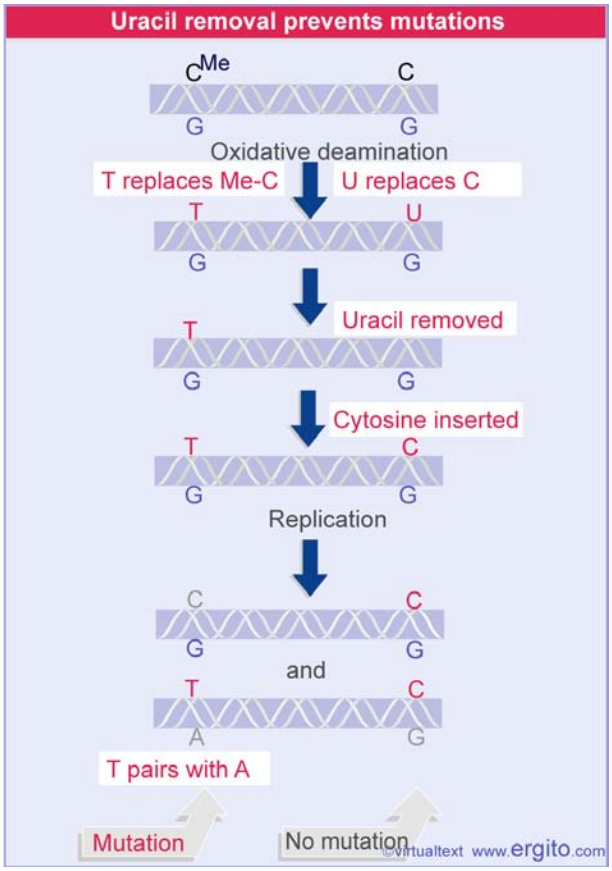

Figure 1.25 The deamination of 5-methylcytosine produces thymine (by C·G to T·A transitions), while the deamination of cytosine produces uracil (which usually is removed and then replaced by cytosine).

E. coli contains an enzyme, uracil-DNA-glycosidase, that removes uracil residues from DNA (see *Molecular Biology 4.15.22 Base flipping is used by methylases and glycosylases*). This action leaves an unpaired G residue, and a "repair system" then inserts a C base to partner it. The net result of these reactions is to restore the original sequence of the DNA. This system protects DNA against the consequences of spontaneous deamination of cytosine (although it is not active enough to prevent the effects of the increased level of deamination caused by nitrous acid; see **Figure 1.20**).

But the deamination of 5-methylcytosine leaves thymine. This creates a mismatched base pair, G·T. If the mismatch is not corrected before the next replication cycle, a mutation results. At the next replication, the bases in the mispaired G·T partnership separate, and then they pair with new partners to produce one wild-type G·C pair and one mutant A·T pair.

Deamination of 5-methylcytosine is the most common cause of production of G·T mismatched pairs in DNA. Repair systems that act on G·T mismatches have a bias toward replacing the T with a C (rather than the alternative of replacing the G with an A), which helps to reduce the rate of mutation (see *Molecular Biology 4.15.24 Controlling the direction of mismatch repair*). However, these

systems are not as effective as the removal of U from G·U mismatches. As a result, deamination of 5-methylcytosine leads to mutation much more often than does deamination of cytosine.

5-methylcytosine also creates hotspots in eukaryotic DNA. It is common at CpG dinucleotides that are concentrated in regions called CpG islands (see *Molecular Biology 5.21.19 CpG islands are regulatory targets*). Although 5-methylcytosine accounts for ~1% of the bases in human DNA, sites containing the modified base account for ~30% of all point mutations. This makes the state of 5-methylcytosine a particularly important determinant of mutation in animal cells.

The importance of repair systems in reducing the rate of mutation is emphasized by the effects of eliminating the mouse enzyme MBD4, a glycosylase that can remove T (or U) from mismatches with G. The result is to increase the mutation rate at CpG sites by a factor of 3× (2845). (The reason the effect is not greater is that MBD4 is only one of several systems that act on G·T mismatches; we can imagine that elimination of all the systems would increase the mutation rate much more.)

The operation of these systems casts an interesting light on the use of T in DNA compared with U in RNA. Perhaps it relates to the need of DNA for stability of sequence; the use of T means that any deaminations of C are immediately recognized, because they generate a base (U) not usually present in the DNA. This greatly increases the efficiency with which repair systems can function (compared with the situation when they have to recognize G·T mismatches, which can be produced also by situations where removing the T would not be the appropriate response). Also, the phosphodiester bond of the backbone is more labile when the base is U.

Last updated on 8-15-2002

References

382. Coulondre, C. et al. (1978). *Molecular basis of base substitution hotspots in E. coli*. Nature 274, 775-780.

2845. Millar, C. B., Guy, J., Sansom, O. J., Selfridge, J., MacDougall, E., Hendrich, B., Keightley, P. D., Bishop, S. M., Clarke, A. R., and Bird, A. (2002). *Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice.* Science 297, 403-405.

GENES ARE DNA

1.1.15 A gene codes for a single polypeptide

Key Terms

A **homomultimer** is a protein composed of identical subunits.

A **heteromultimer** is a protein that is composed of nonidentical subunits (coded by different genes).

Key Concepts

- The one gene : one enzyme hypothesis summarizes the basis of modern genetics: that a gene is a stretch of DNA coding for a single polypeptide chain.
- Most mutations damage gene function.

The first systematic attempt to associate genes with enzymes showed that each stage in a metabolic pathway is catalyzed by a single enzyme and can be blocked by mutation in a different gene. This led to the *one gene : one enzyme hypothesis*. Each metabolic step is catalyzed by a particular enzyme, whose production is the responsibility of a single gene. A mutation in the gene alters the activity of the protein for which it is responsible.

A modification in the hypothesis is needed to accommodate proteins that consist of more than one subunit. If the subunits are all the same, the protein is a **homomultimer**, represented by a single gene. If the subunits are different, the protein is a **heteromultimer**. Stated as a more general rule applicable to any heteromultimeric protein, the one gene : one enzyme hypothesis becomes more precisely expressed as *one gene : one polypeptide chain*.

Identifying which protein represents a particular gene can be a protracted task. The mutation responsible for creating Mendel's wrinkled-pea mutant was identified only in 1990 as an alteration that inactivates the gene for a starch branching enzyme!

It is important to remember that a gene does not directly generate a protein. As shown previously in **Figure 1.2**, a gene codes for an RNA, which may in turn code for a protein. Most genes code for proteins, but some genes code for RNAs that do not give rise to proteins. These RNAs may be structural components of the apparatus responsible for synthesizing proteins or may have roles in regulating gene expression. The basic principle is that the gene is a sequence of DNA that specifies the sequence of an independent product. The process of gene expression may terminate in a product that is either RNA or protein.

A mutation is a random event with regard to the structure of the gene, so the greatest probability is that it will damage or even abolish gene function. Most mutations that affect gene function are recessive: *they represent an absence of function, because the mutant gene has been prevented from producing its usual protein.***Figure 1.26** illustrates the relationship between recessive and wild-type alleles. When a

heterozygote contains one wild-type allele and one mutant allele, the wild-type allele is able to direct production of the enzyme. The wild-type allele is therefore dominant. (This assumes that an adequate *amount* of protein is made by the single wild-type allele. When this is not true, the smaller amount made by one allele as compared to two alleles results in the intermediate phenotype of a partially dominant allele in a heterozygote.)

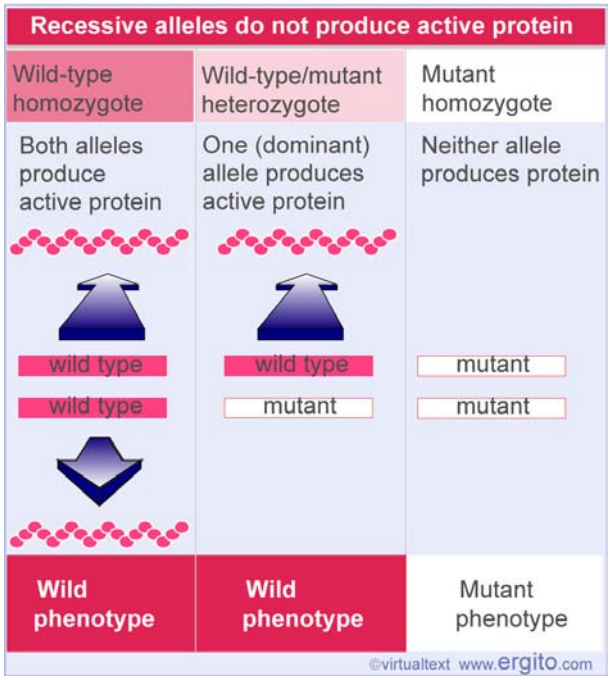

Figure 1.26 Genes code for proteins; dominance is explained by the properties of mutant proteins. A recessive allele does not contribute to the phenotype because it produces no protein (or protein that is nonfunctional).

Last updated on 7-18-2002

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.1.15

GENES ARE DNA

1.1.16 Mutations in the same gene cannot complement

Key Terms

A **complementation test** determines whether two mutations are alleles of the same gene. It is accomplished by crossing two different recessive mutations that have the same phenotype and determining whether the wild-type phenotype can be produced. If so, the mutations are said to complement each other and are probably not mutations in the same gene.

Two mutants are said to **complement** each other when a diploid that is heterozygous for each mutation produces the wild type phenotype.

A **complementation group** is a series of mutations unable to complement when tested in pairwise combinations in *trans;* defines a genetic unit (the cistron).

A **cistron** is the genetic unit defined by the complementation test; it is equivalent to the gene.

A **gene (cistron)** is the segment of DNA specifying production of a polypeptide chain; it includes regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).

Key Concepts

- A mutation in a gene affects only the protein coded by the mutant copy of the gene, and does not affect the protein coded by any other allele.
- Failure of two mutations to complement (produce wild-phenotype) when they are present in *trans* configuration in a heterozygote means that they are part of the same gene.

How do we determine whether two mutations that cause a similar phenotype lie in the same gene? If they map close together, they may be alleles. However, they could also represent mutations in two *different* genes whose proteins are involved in the same function. The **complementation test** is used to determine whether two mutations lie in the same gene or in different genes. The test consists of making a heterozygote for the two mutations (by mating parents homozygous for each mutation).

If the mutations lie in the same gene, the parental genotypes can be represented as:

$$\frac{m_1}{m_1} \text{ and } \frac{m_2}{m_2}$$

The first parent provides an m_1 mutant allele and the second parent provides an m_2

allele, so that the heterozygote has the constitution:

$$\frac{m_1}{m_2}$$

No wild-type gene is present, so the heterozygote has mutant phenotype.

If the mutations lie in different genes, the parental genotypes can be represented as:

$$\frac{m_1 +}{m_1 +} \text{ and } \frac{+ m_2}{+ m_2}$$

Each chromosome has a wild-type copy of one gene (represented by the plus sign) and a mutant copy of the other. Then the heterozygote has the constitution:

$$\frac{m_1 +}{+ m_2}$$

in which the two parents between them have provided a wild-type copy of each gene. The heterozygote has wild phenotype; the two genes are said to **complement**.

The complementation test is shown in more detail in **Figure 1.27**. The basic test consists of the comparison shown in the top part of the figure. If two mutations lie in the same gene, we see a difference in the phenotypes of the *trans* configuration and the *cis* configuration. The *trans* configuration is mutant, because each allele has a (different) mutation. But the *cis* configuration is wild-type, because one allele has two mutations but the other allele has no mutations. The lower part of the figure shows that if the two mutations lie in different genes, we always see a wild phenotype. There is always one wild-type and one mutant allele of each gene, and the configuration is irrelevant. The basic test and some exceptions to it are discussed in *Molecular Biology Supplement 32.9 Complementation*.

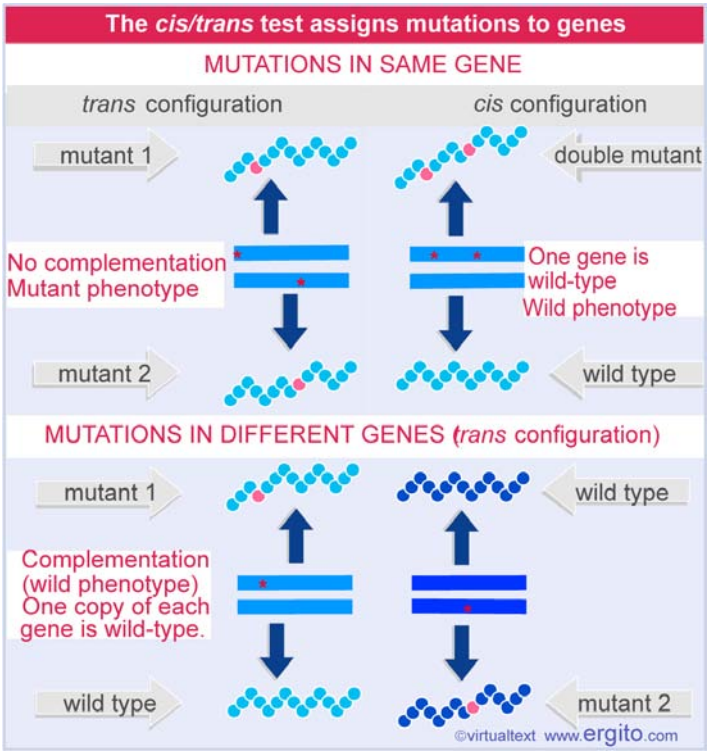

Figure 1.27 The cistron is defined by the complementation test. Genes are represented by bars; red stars identify sites of mutation.

Failure to complement means that two mutations are part of the *same* genetic unit. Mutations that do not complement one another are said to comprise part of the same **complementation group**. Another term that is used to describe the unit defined by the complementation test is the **cistron**. This is the same as the **gene**. Basically these three terms all describe a stretch of DNA that functions as a unit to give rise to an RNA or protein product. The properties of the gene with regards to complementation are explained by the fact that this product is a single molecule that behaves as a functional unit.

Last updated on 7-18-2002

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.1.16

GENES ARE DNA

1.1.17 Mutations may cause loss-of-function or gain-of-function

Key Terms

A **null mutation** completely eliminates the function of a gene.

Leaky mutations leave some residual function, for instance when the mutant protein is partially active (in the case of a missense mutation), or when read-through produces a small amount of wild-type protein (in the case of a nonsense mutation).

A **loss-of-function** mutation eliminates or reduces the activity of a gene. It is often, but not always, recessive.

A **gain-of-function** mutation usually refers to a mutation that causes an increase in the normal gene activity. It sometimes represents acquisition of certain abnormal properties. It is often, but not always, dominant.

Silent mutations do not change the sequence of a protein because they produce synonymous codons.

Neutral substitutions in a protein cause changes in amino acids that do not affect activity.

Key Concepts

- Recessive mutations are due to loss-of-function by the protein product.
- Dominant mutations result from a gain-of-function.
- Testing whether a gene is essential requires a null mutation (one that completely eliminates its function).
- Silent mutations have no effect, either because the base change does not change the sequence or amount of protein, or because the change in protein sequence has no effect.
- Leaky mutations do affect the function of the gene product, but are not revealed in the phenotype because sufficient activity remains.

The various possible effects of mutation in a gene are summarized in **Figure 1.28**.

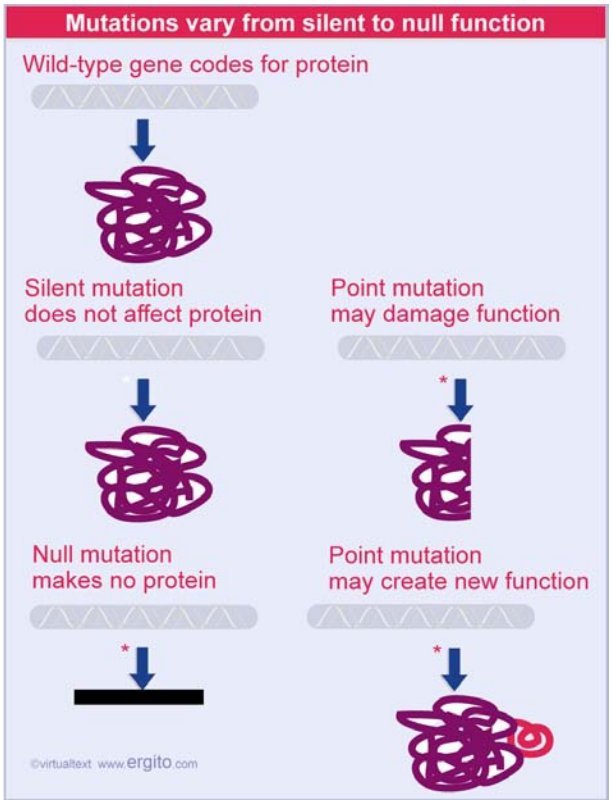

Figure 1.28 Mutations that do not affect protein sequence or function are silent. Mutations that abolish all protein activity are null. Point mutations that cause loss-of-function are recessive; those that cause gain-of-function are dominant.

When a gene has been identified, insight into its function in principle can be gained by generating a mutant organism that entirely lacks the gene. A mutation that completely eliminates gene function, usually because the gene has been deleted, is called a **null mutation**. If a gene is essential, a null mutation is lethal.

To determine what effect a gene has upon the phenotype, it is essential to characterize a null mutant. When a mutation fails to affect the phenotype, it is always possible that this is because it is a **leaky mutation** – enough active product is made to fulfill its function, even though the activity is quantitatively reduced or qualitatively different from the wild type. But if a null mutant fails to affect a phenotype, we may safely conclude that the gene function is not necessary.

Null mutations, or other mutations that impede gene function (but do not necessarily abolish it entirely) are called **loss-of-function** mutations. A loss-of-function mutation is recessive (as in the example of **Figure 1.26**). Sometimes a mutation has the opposite effect and causes a protein to acquire a new function; such a change is called a **gain-of-function** mutation. A gain-of-function mutation is dominant.

Not all mutations in DNA lead to a detectable change in the phenotype. Mutations without apparent effect are called **silent mutations**. They fall into two types. Some involve base changes in DNA that do not cause any change in the amino acid present in the corresponding protein. Others change the amino acid, but the replacement in

the protein does not affect its activity; these are called **neutral substitutions**.

Last updated on 10-2-2003

GENES ARE DNA

1.1.18 A locus may have many different mutant alleles

Key Terms

A locus is said to have **multiple alleles** when more than two allelic forms have been found. Each allele may cause a different phenotype.

Key Concepts

- The existence of multiple alleles allows heterozygotes to occur representing any pairwise combination of alleles.

If a recessive mutation is produced by every change in a gene that prevents the production of an active protein, there should be a large number of such mutations in any one gene. Many amino acid replacements may change the structure of the protein sufficiently to impede its function.

Different variants of the same gene are called **multiple alleles**, and their existence makes it possible to create a heterozygote between mutant alleles. The relationship between these multiple alleles takes various forms.

In the simplest case, a wild-type gene codes for a protein product that is functional. Mutant allele(s) code for proteins that are nonfunctional.

But there are often cases in which a series of mutant alleles have different phenotypes. For example, wild-type function of the *white* locus of *D. melanogaster* is required for development of the normal red color of the eye. The locus is named for the effect of extreme (null) mutations, which cause the fly to have a white eye in mutant homozygotes.

To describe wild-type and mutant alleles, wild genotype is indicated by a plus superscript after the name of the locus (w^+ is the wild-type allele for [red] eye color in *D. melanogaster*). Sometimes + is used by itself to describe the wild-type allele, and only the mutant alleles are indicated by the name of the locus.

An entirely defective form of the gene (or absence of phenotype) may be indicated by a minus superscript. To distinguish among a variety of mutant alleles with different effects, other superscripts may be introduced, such as w^i or w^a.

The w^+ allele is dominant over any other allele in heterozygotes. There are many different mutant alleles. **Figure 1.29** shows a (small) sample. Although some alleles have no eye color, many alleles produce some color. Each of these mutant alleles must therefore represent a different mutation of the gene, which does not eliminate its function entirely, but leaves a residual activity that produces a characteristic phenotype. These alleles are named for the color of the eye in a homozygote. (Most

w alleles affect the quantity of pigment in the eye, and the examples in the Figure are arranged in [roughly] declining amount of color, but others, such as w^{sp}, affect the pattern in which it is deposited.)

Figure 1.29 The *w* locus has an extensive series of alleles, whose phenotypes extend from wild-type (red) color to complete lack of pigment.

When multiple alleles exist, an animal may be a heterozygote that carries two different mutant alleles. The phenotype of such a heterozygote depends on the nature of the residual activity of each allele. The relationship between two mutant alleles is in principle no different from that between wild-type and mutant alleles: one allele may be dominant, there may be partial dominance, or there may be codominance.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.1.18

GENES ARE DNA

1.1.19 A locus may have more than one wild-type allele

Key Terms

Polymorphism (more fully genetic polymorphism) refers to the simultaneous occurrence in the population of genomes showing variations at a given position. The original definition applied to alleles producing different phenotypes. Now it is also used to describe changes in DNA affecting the restriction pattern or even the sequence. For practical purposes, to be considered as an example of a polymorphism, an allele should be found at a frequency > 1% in the population.

Key Concepts

- A locus may have a polymorphic distribution of alleles, with no individual allele that can be considered to be the sole wild-type.

There is not necessarily a unique wild-type allele at any particular locus. Control of the human blood group system provides an example. Lack of function is represented by the null type, *O* group. But the functional alleles *A* and *B* provide activities that are codominant with one another and dominant over *O* group. The basis for this relationship is illustrated in **Figure 1.30**.

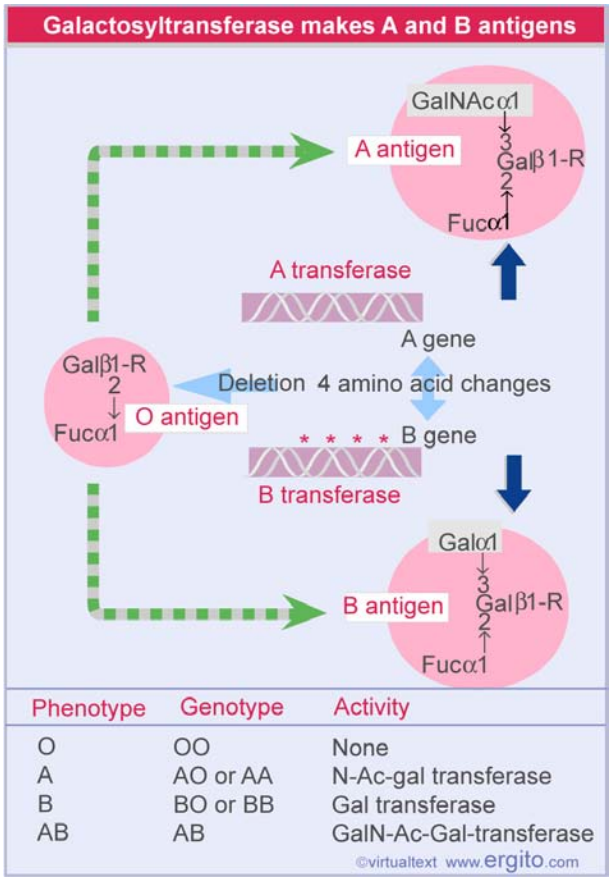

Figure 1.30 The ABO blood group locus codes for a galactosyltransferase whose specificity determines the blood group.

The O (or H) antigen is generated in all individuals, and consists of a particular carbohydrate group that is added to proteins. The *ABO* locus codes for a galactosyltransferase enzyme that adds a further sugar group to the O antigen. The specificity of this enzyme determines the blood group. The *A* allele produces an enzyme that uses the cofactor UDP-N-acetylgalactose, creating the A antigen. The *B* allele produces an enzyme that uses the cofactor UDP-galactose, creating the B antigen. The A and B versions of the transferase protein differ in 4 amino acids that presumably affect its recognition of the type of cofactor. The *O* allele has a mutation (a small deletion) that eliminates activity, so no modification of the O antigen occurs.

This explains why *A* and *B* alleles are dominant in the *AO* and *BO* heterozygotes: the corresponding transferase activity creates the A or B antigen. The *A* and *B* alleles are codominant in *AB* heterozygotes, because both transferase activities are expressed. The *OO* homozygote is a null that has neither activity, and therefore lacks both antigens.

Neither *A* nor *B* can be regarded as uniquely wild type, since they represent alternative activities rather than loss or gain of function. A situation such as this, in which there are multiple functional alleles in a population, is described as a **polymorphism** (see *Molecular Biology 1.3.3 Individual genomes show extensive*

variation).

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.1.19

GENES ARE DNA

1.1.20 Recombination occurs by physical exchange of DNA

Key Terms

Crossing-over describes the reciprocal exchange of material between chromosomes that occurs during prophase I of meiosis and is responsible for genetic recombination.

A **bivalent** is the structure containing all four chromatids (two representing each homologue) at the start of meiosis.

Chromatids are the copies of a chromosome produced by replication. The name is usually used to describe the copies in the period before they separate at the subsequent cell division.

A **chiasma** (*pl.* chiasmata) is a site at which two homologous chromosomes appear to have exchanged material during meiosis.

Breakage and reunion describes the mode of genetic recombination, in which two DNA duplex molecules are broken at corresponding points and then rejoined crosswise (involving formation of a length of heteroduplex DNA around the site of joining).

Heteroduplex DNA (Hybrid DNA) is generated by base pairing between complementary single strands derived from the different parental duplex molecules; it occurs during genetic recombination.

Key Concepts

- Recombination is the result of crossing-over that occurs at chiasmata and involves two of the four chromatids.
- Recombination occurs by a breakage and reunion that proceeds via an intermediate of hybrid DNA.

Genetic recombination describes the generation of new combinations of alleles that occurs at each generation in diploid organisms. The two copies of each chromosome may have different alleles at some loci. By exchanging corresponding parts between the chromosomes, recombinant chromosomes can be generated that are different from the parental chromosomes.

Recombination results from a physical exchange of chromosomal material. This is visible in the form of the **crossing-over** that occurs during meiosis (the specialized division that produces haploid germ cells). Meiosis starts with a cell that has duplicated its chromosomes, so that it has four copies of each chromosome. Early in meiosis, all four copies are closely associated (synapsed) in a structure called a **bivalent**. Each individual chromosomal unit is called a **chromatid** at this stage. Pairwise exchanges of material occur between the chromatids.

The visible result of a crossing-over event is called a **chiasma**, and is illustrated diagrammatically in **Figure 1.31**. A chiasma represents a site at which two of the chromatids in a bivalent have been broken at corresponding points. The broken ends have been rejoined crosswise, generating new chromatids. Each new chromatid consists of material derived from one chromatid on one side of the junction point, with material from the other chromatid on the opposite side. The two recombinant chromatids have reciprocal structures. The event is described as a **breakage and reunion**. Its nature explains why a single recombination event can produce only 50% recombinants: each individual recombination event involves only two of the four associated chromatids.

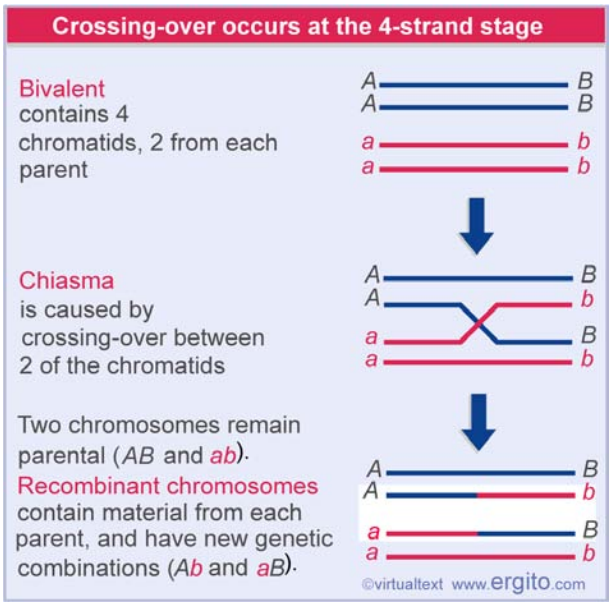

Figure 1.31 Chiasma formation is responsible for generating recombinants.

The complementarity of the two strands of DNA is essential for the recombination process. Each of the chromatids shown in **Figure 1.31** consists of a very long duplex of DNA. For them to be broken and reconnected without any loss of material requires a mechanism to recognize exactly corresponding positions. This is provided by complementary base pairing.

Recombination involves a process in which the single strands in the region of the crossover exchange their partners. **Figure 1.32** shows that this creates a stretch of **hybrid DNA** in which the single strand of one duplex is paired with its complement from the other duplex. The mechanism of course involves other stages (strands must be broken and resealed), and we discuss this in more detail in *Molecular Biology 4.15 Recombination and repair*, but the crucial feature that makes precise recombination possible is the complementarity of DNA strands. The figure shows only some stages of the reaction, but we see that a stretch of hybrid DNA forms in the recombination intermediate when a single strand crosses over from one duplex to the other. Each recombinant consists of one parental duplex DNA at the left, connected by a stretch of hybrid DNA to the other parental duplex at the right. Each duplex DNA corresponds to one of the chromatids involved in recombination in **Figure 1.31**.

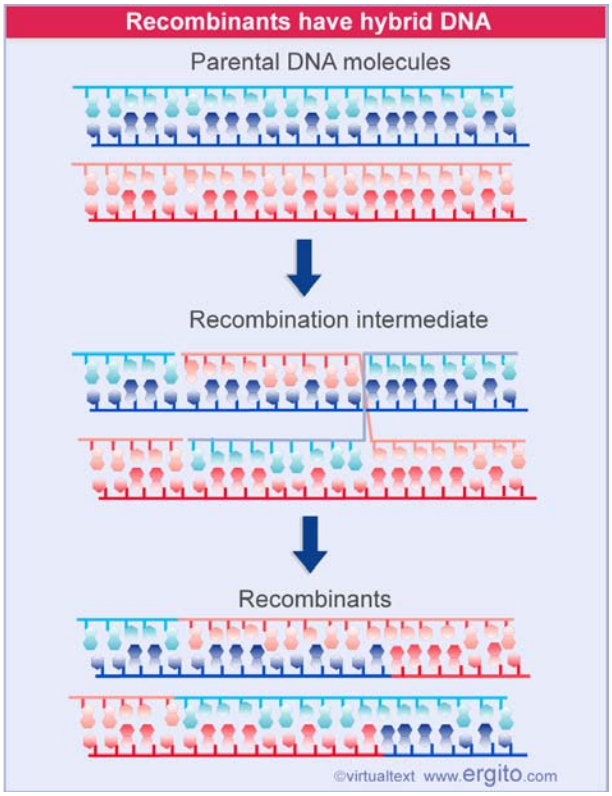

Figure 1.32 Recombination involves pairing between complementary strands of the two parental duplex DNAs.

The formation of hybrid DNA requires the sequences of the two recombining duplexes to be close enough to allow pairing between the complementary strands. If there are no differences between the two parental genomes in this region, formation of hybrid DNA will be perfect. But the reaction can be tolerated even when there are small differences. In this case, the hybrid DNA has points of mismatch, at which a base in one strand faces a base in the other strand that is not complementary to it. The correction of such mismatches is another feature of genetic recombination (see *Molecular Biology 4.15 Recombination and repair*).

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.1.20

GENES ARE DNA

1.1.21 The genetic code is triplet

Key Terms

The **genetic code** is the correspondence between triplets in DNA (or RNA) and amino acids in protein.

A **codon** is a triplet of nucleotides that represents an amino acid or a termination signal.

Frameshift mutations arise by deletions or insertions that are not a multiple of 3 base pairs and change the frame in which triplets are translated into protein. The term is inappropriate outside of coding sequences.

Acridines are mutagens that act on DNA to cause the insertion or deletion of a single base pair. They were useful in defining the triplet nature of the genetic code.

A **suppressor** is a second mutation that compensates for or alters the effects of a primary mutation.

A frameshift **suppressor** is an insertion or deletion of a base that restores the original reading frame in a gene that has had a base deletion or insertion.

Key Concepts

- The genetic code is read in triplet nucleotides called codons.
- The triplets are nonoverlapping and are read from a fixed starting point.
- Mutations that insert or delete individual bases cause a shift in the triplet sets after the site of mutation.
- Combinations of mutations that together insert or delete 3 bases (or multiples of three) insert or delete amino acids but do not change the reading of the triplets beyond the last site of mutation.

Each gene represents a particular protein chain. The concept that each protein consists of a particular series of amino acids dates from Sanger's characterization of insulin in the 1950s. The discovery that a gene consists of DNA faces us with the issue of how a sequence of nucleotides in DNA represents a sequence of amino acids in protein.

A crucial feature of the general structure of DNA is that *it is independent of the particular sequence of its component nucleotides.* The sequence of nucleotides in DNA is important not because of its structure *per se,* but because it *codes* for the sequence of amino acids that constitutes the corresponding polypeptide. The relationship between a sequence of DNA and the sequence of the corresponding protein is called the **genetic code**.

The structure and/or enzymatic activity of each protein follows from its primary sequence of amino acids. By determining the sequence of amino acids in each

protein, the gene is able to carry all the information needed to specify an active polypeptide chain. In this way, a single type of structure – the gene – is able to represent itself in innumerable polypeptide forms.

Together the various protein products of a cell undertake the catalytic and structural activities that are responsible for establishing its phenotype. Of course, in addition to sequences that code for proteins, DNA also contains certain sequences whose function is to be recognized by regulator molecules, usually proteins. Here the function of the DNA is determined by its sequence directly, not via any intermediary code. Both types of region, genes expressed as proteins and sequences recognized as such, constitute genetic information.

The genetic code is deciphered by a complex apparatus that interprets the nucleic acid sequence. This apparatus is essential if the information carried in DNA is to have meaning. In any given region, only one of the two strands of DNA codes for protein, so we write the genetic code as a sequence of bases (rather than base pairs).

The genetic code is read in groups of three nucleotides, each group representing one amino acid. Each trinucleotide sequence is called a **codon**. A gene includes a series of codons that is read sequentially from a starting point at one end to a termination point at the other end. Written in the conventional $5' \rightarrow 3'$ direction, the nucleotide sequence of the DNA strand that codes for protein corresponds to the amino acid sequence of the protein written in the direction from N-terminus to C-terminus.

The genetic code is read in *nonoverlapping triplets from a fixed starting point*:

- Nonoverlapping implies that each codon consists of three nucleotides and that successive codons are represented by successive trinucleotides.
- The use of a *fixed starting point* means that assembly of a protein must start at one end and work to the other, so that different parts of the coding sequence cannot be read independently.

The nature of the code predicts that two types of mutations will have different effects. If a particular sequence is read sequentially, such as:

UUU AAA GGG CCC (codons)

aa1 aa2 aa3 aa4 (amino acids)

then a point mutation will affect only one amino acid. For example, the substitution of an A by some other base (X) causes aa2 to be replaced by aa5:

UUU AAX GGG CCC

aa1 aa5 aa3 aa4

because only the second codon has been changed.

But a mutation that inserts or deletes a single base will change the triplet sets for the entire subsequent sequence. A change of this sort is called a **frameshift**. An insertion might take the form:

UUU AAX AGG GCC C

aa1 aa5 aa6 aa7

Because the new sequence of triplets is completely different from the old one, the entire amino acid sequence of the protein is altered beyond the site of mutation. So the function of the protein is likely to be lost completely.

Frameshift mutations are induced by the **acridines**, compounds that bind to DNA and distort the structure of the double helix, causing additional bases to be incorporated or omitted during replication. Each mutagenic event sponsored by an acridine results in the addition or removal of a single base pair (for review see 4).

If an acridine mutant is produced by, say, addition of a nucleotide, it should revert to wild type by deletion of the nucleotide. But reversion can also be caused by deletion of a different base, at a site close to the first. Combinations of such mutations provided revealing evidence about the nature of the genetic code.

Figure 1.33 illustrates the properties of frameshift mutations. An insertion or a deletion changes the entire protein sequence following the site of mutation. But the combination of an insertion and a deletion causes the code to be read incorrectly only between the two sites of mutation; correct reading resumes after the second site.

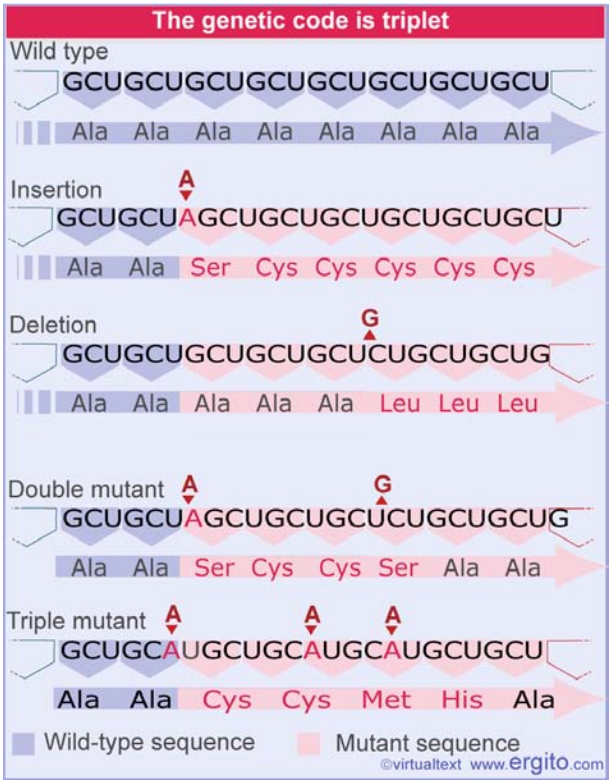

Figure 1.33 Frameshift mutations show that the genetic code is read in triplets from a fixed starting point.

Genetic analysis of acridine mutations in the *rII* region of the phage T6 in 1961 showed that all the mutations could be classified into one of two sets, described as (+) and (–). Either type of mutation by itself causes a frameshift, the (+) type by virtue of a base addition, the (–) type by virtue of a base deletion. Double mutant combinations of the types (+ +) and (—) continue to show mutant behavior. But combinations of the types (+ –) or (– +) suppress one another, giving rise to a description in which one mutation is described as a **supressor** of the other. (In the context of this work, "suppressor" is used in an unusual sense, because the second mutation is in the same gene as the first.)

These results show that the genetic code must be read as a sequence that is fixed by the starting point, so additions or deletions compensate for each other, whereas double additions or double deletions remain mutant. But this does not reveal how many nucleotides make up each codon.

When triple mutants are constructed, only (+ + +) and (——) combinations show the wild phenotype, while other combinations remain mutant. If we take three additions or three deletions to correspond respectively to the addition or omission overall of a single amino acid, this implies that the code is read in triplets. An incorrect amino acid sequence is found between the two outside sites of mutation, and the sequence on either side remains wild type, as indicated in **Figure 1.33** (378; 379).

Last updated on January 27, 2004

Reviews

4. Roth, J. R. (1974). *Frameshift mutations.* Annu. Rev. Genet. 8, 319-346.

References

378. Benzer, S. and Champe, S. P. (1961). *Ambivalent rII mutants of phage T4*. Proc. Natl. Acad. Sci. USA 47, 403-416.

379. Crick, F. H. C., Barnett, L., Brenner, S., and Watts-Tobin, R. J. (1961). *General nature of the genetic code for proteins*. Nature 192, 1227-1232.

GENES ARE DNA

1.1.22 Every sequence has three possible reading frames

Key Terms

A **reading frame** is one of the three possible ways of reading a nucleotide sequence. Each reading frame divides the sequence into a series of successive triplets. There are three possible reading frames in any sequence, depending on the starting point. If the first frame starts at position 1, the second frame starts at position 2, and the third frame starts at position 3.

An **open reading frame (ORF)** is a sequence of DNA consisting of triplets that can be translated into amino acids starting with an initiation codon and ending with a termination codon.

The **initiation codon** is a special codon (usually AUG) used to start synthesis of a protein.

A **stop codon (Termination codon)** is one of three triplets (UAG, UAA, UGA) that causes protein synthesis to terminate. They are also known historically as *nonsense codons*. The UAA codon is called ochre, and the UAA codon is called amber, after the names of the nonsense mutations by which they were originally identified.

A **blocked** reading frame cannot be translated into protein because of the occurrence of termination codons.

Key Concepts

- Usually only one reading frame is translated and the other two are blocked by frequent termination signals.

If the genetic code is read in nonoverlapping triplets, there are three possible ways of translating any nucleotide sequence into protein, depending on the starting point. These called **reading frames**. For the sequence

A C G A C G A C G A C G A C G A C G

the three possible reading frames are

ACG ACG ACG ACG ACG ACG ACG

CGA CGA CGA CGA CGA CGA CGA

GAC GAC GAC GAC GAC GAC GAC

A reading frame that consists exclusively of triplets representing amino acids is called an **open reading frame** or **ORF**. A sequence that is translated into protein has

a reading frame that starts with a special **initiation codon** (AUG) and that extends through a series of triplets representing amino acids until it ends at one of three types of **termination codon** (see *Molecular Biology 2.5 Messenger RNA*).

A reading frame that cannot be read into protein because termination codons occur frequently is said to be **blocked**. If a sequence is blocked in all three reading frames, it cannot have the function of coding for protein.

When the sequence of a DNA region of unknown function is obtained, each possible reading frame is analyzed to determine whether it is open or blocked. Usually no more than one of the three possible frames of reading is open in any single stretch of DNA. **Figure 1.34** shows an example of a sequence that can be read in only one reading frame, because the alternative reading frames are blocked by frequent termination codons. A long open reading frame is unlikely to exist by chance; if it were not translated into protein, there would have been no selective pressure to prevent the accumulation of termination codons. So the identification of a lengthy open reading frame is taken to be *prima facie* evidence that the sequence is translated into protein in that frame. An open reading frame (ORF) for which no protein product has been identified is sometimes called an unidentified reading frame (URF).

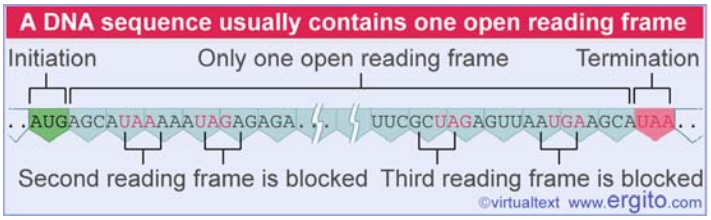

Figure 1.34 An open reading frame starts with AUG and continues in triplets to a termination codon. Blocked reading frames may be interrupted frequently by termination codons.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.1.22

GENES ARE DNA

1.1.23 Prokaryotic genes are colinear with their proteins

Key Terms

A **colinear** relationship describes the 1:1 representation of a sequence of triplet nucleotides in a sequence of amino acids.

Key Concepts

- A prokaryotic gene consists of a continuous length of 3N nucleotides that codes for N amino acids.
- The gene, mRNA, and protein are all colinear.

By comparing the nucleotide sequence of a gene with the amino acid sequence of a protein, we can determine directly whether the gene and the protein are **colinear**: whether the sequence of nucleotides in the gene corresponds exactly with the sequence of amino acids in the protein. In bacteria and their viruses, there is an exact equivalence. Each gene contains a continuous stretch of DNA whose length is directly related to the number of amino acids in the protein that it represents. A gene of 3*N* bp is required to code for a protein of *N* amino acids, according to the genetic code.

The equivalence of the bacterial gene and its product means that a physical map of DNA will exactly match an amino acid map of the protein. How well do these maps fit with the recombination map?

The colinearity of gene and protein was originally investigated in the tryptophan synthetase gene of *E. coli* (see *Great Experiments 1.2 Gene-protein colinearity*). Genetic distance was measured by the percent recombination between mutations; protein distance was measured by the number of amino acids separating sites of replacement. **Figure 1.35** compares the two maps. The order of seven sites of mutation is the same as the order of the corresponding sites of amino acid replacement. And the recombination distances are relatively similar to the actual distances in the protein. The recombination map expands the distances between some mutations, but otherwise there is little distortion of the recombination map relative to the physical map (380; 1225).

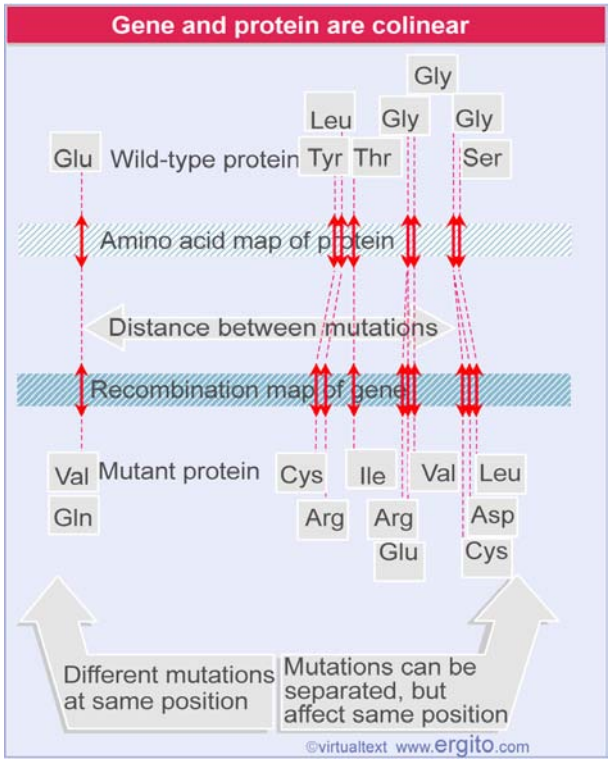

Figure 1.35 The recombination map of the tryptophan synthetase gene corresponds with the amino acid sequence of the protein.

The recombination map makes two further general points about the organization of the gene. Different mutations may cause a wild-type amino acid to be replaced with different substituents. If two such mutations cannot recombine, they must involve different point mutations at the same position in DNA. If the mutations can be separated on the genetic map, but affect the same amino acid on the upper map (the connecting lines converge in the figure), they must involve point mutations at different positions that affect the same amino acid. This happens because the unit of genetic recombination (actually 1 bp) is smaller than the unit coding for the amino acid (actually 3 bp).

References

380. Yanofsky, C. et al. (1964). *On the colinearity of gene structure and protein structure*. Proc. Natl. Acad. Sci. USA 51, 266-272.

1225. Yanofsky, C., Drapeau, G. R., Guest, J. R., and Carlton, B. C. (1967). *The complete amino acid sequence of the tryptophan synthetase A protein (μ subunit) and its colinear relationship with the genetic map of the A gene*. Proc. Natl. Acad. Sci. USA 57, 2966-2968.

GENES ARE DNA

1.1.24 Several processes are required to express the protein product of a gene

Key Terms

Messenger RNA (mRNA) is the intermediate that represents one strand of a gene coding for protein. Its coding region is related to the protein sequence by the triplet genetic code.

Transcription describes synthesis of RNA on a DNA template.

Translation is synthesis of protein on the mRNA template.

A **coding region** is a part of the gene that represents a protein sequence.

The **leader** of a protein is a short N-terminal sequence responsible for initiating passage into or through a membrane.

A **trailer (3′ UTR)** is a nontranslated sequence at the 3′ end of an mRNA following the termination codon.

Pre-mRNA is used to describe the nuclear transcript that is processed by modification and splicing to give an mRNA.

Processing of RNA describes changes that occur after its transcription, including modification of the 5′ and 3′ ends, internal methylation, splicing, or cleavage.

RNA splicing is the process of excising the sequences in RNA that correspond to introns, so that the sequences corresponding to exons are connected into a continuous mRNA.

Key Concepts

- A prokaryotic gene is expressed by transcription into mRNA and then by translation of the mRNA into protein.
- In eukaryotes, a gene may contain internal regions that are not represented in protein.
- Internal regions are removed from the RNA transcript by RNA splicing to give an mRNA that is colinear with the protein product.
- Each mRNA consists of a nontranslated 5′ leader, a coding region, and a nontranslated 3′ trailer.

In comparing gene and protein, we are restricted to dealing with the sequence of DNA stretching between the points corresponding to the ends of the protein. However, a gene is not directly translated into protein, but is expressed via the production of a **messenger RNA** (abbreviated to **mRNA**), a nucleic acid intermediate actually used to synthesize a protein (as we see in detail in *Molecular Biology 2.5 Messenger RNA*).

Messenger RNA is synthesized by the same process of complementary base pairing used to replicate DNA, with the important difference that it corresponds to only one strand of the DNA double helix. **Figure 1.36** shows that the sequence of messenger RNA is complementary with the sequence of one strand of DNA and is identical (apart from the replacement of T with U) with the other strand of DNA. The convention for writing DNA sequences is that the top strand runs 5′ → 3′, with the sequence that is the same as RNA.

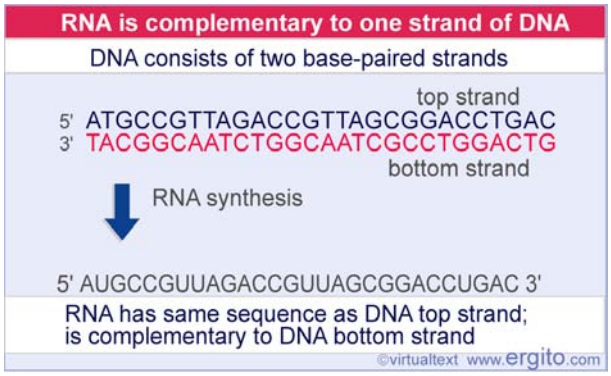

Figure 1.36 RNA is synthesized by using one strand of DNA as a template for complementary base pairing.

The process by which a gene gives rise to a protein is called gene expression. In bacteria, it consists of two stages. The first stage is **transcription**, when an mRNA copy of one strand of the DNA is produced. The second stage is **translation** of the mRNA into protein. This is the process by which the sequence of an mRNA is read in triplets to give the series of amino acids that make the corresponding protein.

A messenger RNA includes a sequence of nucleotides that corresponds with the sequence of amino acids in the protein. This part of the nucleic acid is called the **coding region**. But the messenger RNA includes additional sequences on either end; these sequences do not directly represent protein. The 5′ nontranslated region is called the **leader**, and the 3′ nontranslated region is called the **trailer**.

The *gene* includes the entire sequence represented in messenger RNA. Sometimes mutations impeding gene function are found in the additional, noncoding regions, confirming the view that these comprise a legitimate part of the genetic unit.

Figure 1.37 illustrates this situation, in which the gene is considered to comprise a continuous stretch of DNA, needed to produce a particular protein. It includes the sequence coding for that protein, but also includes sequences on either side of the coding region.

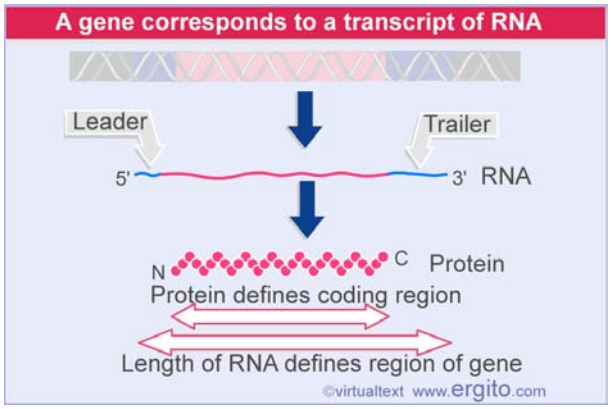

Figure 1.37 The gene may be longer than the sequence coding for protein.

A bacterium consists of only a single compartment, so transcription and translation occur in the same place, as illustrated in **Figure 1.38**.

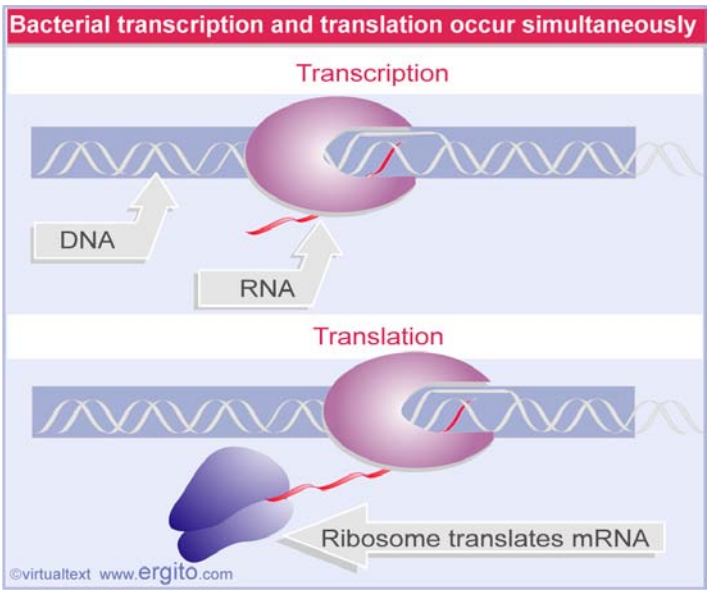

Figure 1.38 Transcription and translation take place in the same compartment in bacteria.

In eukaryotes transcription occurs in the nucleus, but the RNA product must be *transported* to the cytoplasm in order to be translated. For the simplest eukaryotic genes (just like in bacteria) the transcript RNA is in fact the mRNA. But for more complex genes, the immediate transcript of the gene is a **pre-mRNA** that requires **processing** to generate the mature mRNA. The basic stages of gene expression in a eukaryote are outlined in **Figure 1.39**. This results in a spatial separation between transcription (in the nucleus) and translation (in the cytoplasm).

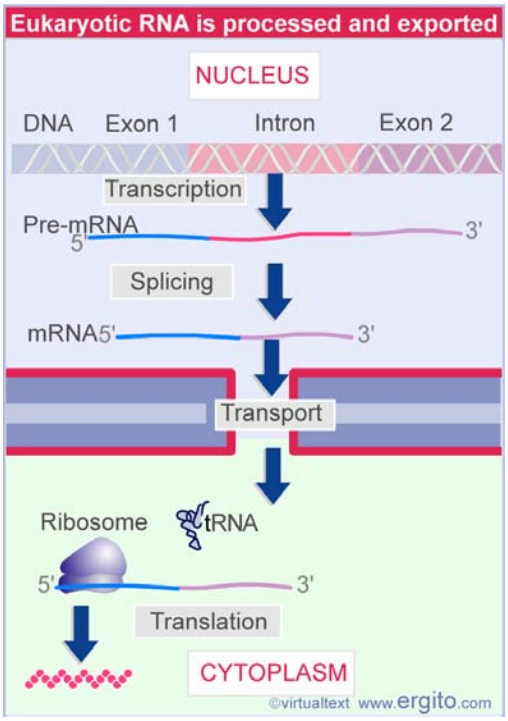

Figure 1.39 Gene expression is a multistage process.

The most important stage in processing is **RNA splicing**. Many genes in eukaryotes (and a majority in higher eukaryotes) contain internal regions that do not code for protein. The process of splicing removes these regions from the pre-mRNA to generate an RNA that has a continuous open reading frame (see **Figure 2.1**). Other processing events that occur at this stage involve the modification of the 5′ and 3′ ends of the pre-mRNA (see **Figure 5.16**).

Translation is accomplished by a complex apparatus that includes both protein and RNA components. The actual "machine" that undertakes the process is the *ribosome*, a large complex that includes some large RNAs (*ribosomal RNA*s, abbreviated to *rRNA*s) and many small proteins. The process of recognizing which amino acid corresponds to a particular nucleotide triplet requires an intermediate *transfer RNA* (abbreviated to *tRNA*); there is at least one tRNA species for every amino acid. Many ancillary proteins are involved. We describe translation in *Molecular Biology 2.5 Messenger RNA*, but note for now that the ribosomes are the large structures in **Figure 1.38** that move along the mRNA.

The important point to note at this stage is that the process of gene expression involves RNA not only as the essential substrate, but also in providing components of the apparatus. The rRNA and tRNA components are coded by genes and are generated by the process of transcription (just like mRNA, except that there is no subsequent stage of translation).

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.1.1.24

GENES ARE DNA

1.1.25 Proteins are *trans*-acting but sites on DNA are *cis*-acting

Key Terms

cis configuration describes two sites on the same molecule of DNA.

trans configuration of two sites refers to their presence on two different molecules of DNA (chromosomes).

A ***cis*-acting** site affects the activity only of sequences on its own molecule of DNA (or RNA); this property usually implies that the site does not code for protein.

Key Concepts

- All gene products (RNA or proteins) are *trans*-acting. They can act on any copy of a gene in the cell.
- *cis*-acting mutations identify sequences of DNA that are targets for recognition by *trans*-acting products. They are not expressed as RNA or protein and affect only the contiguous stretch of DNA.

A crucial step in the definition of the gene was the realization that all its parts must be present on one contiguous stretch of DNA. In genetic terminology, sites that are located on the same DNA are said to be in ***cis***. Sites that are located on two different molecules of DNA are described as being in ***trans***. So two mutations may be in *cis* (on the same DNA) or in *trans* (on different DNAs). The complementation test uses this concept to determine whether two mutations are in the same gene (see **Figure 1.27** in *Molecular Biology 1.1.16 Mutations in the same gene cannot complement*). We may now extend the concept of the difference between *cis* and *trans* effects from defining the coding region of a gene to describing the interaction between regulatory elements and a gene.

Suppose that the ability of a gene to be expressed is controlled by a protein that binds to the DNA close to the coding region. In the example depicted in **Figure 1.40**, messenger RNA can be synthesized only when the protein is bound to the DNA. Now suppose that a mutation occurs in the DNA sequence to which this protein binds, so that the protein can no longer recognize the DNA. As a result, the DNA can no longer be expressed.

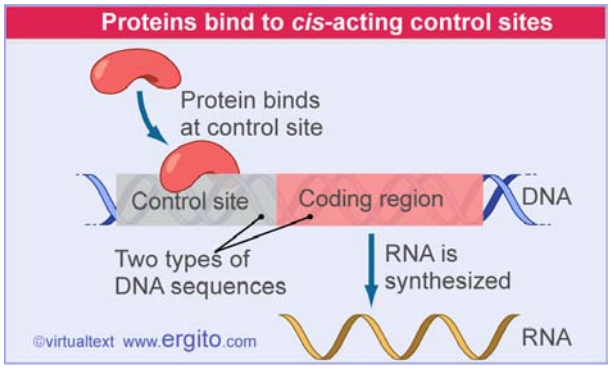

Figure 1.40 Control sites in DNA provide binding sites for proteins; coding regions are expressed via the synthesis of RNA.

So a gene can be inactivated either by a mutation in a control site or by a mutation in a coding region. The mutations cannot be distinguished genetically, because both have the property of acting only on the DNA sequence of the single allele in which they occur. They have identical properties in the complementation test, and a mutation in a control region is therefore defined as comprising part of the gene in the same way as a mutation in the coding region.

Figure 1.41 shows that a deficiency in the control site *affects only the coding region to which it is connected; it does not affect the ability of the other allele to be expressed.* A mutation that acts solely by affecting the properties of the contiguous sequence of DNA is called ***cis*-acting**.

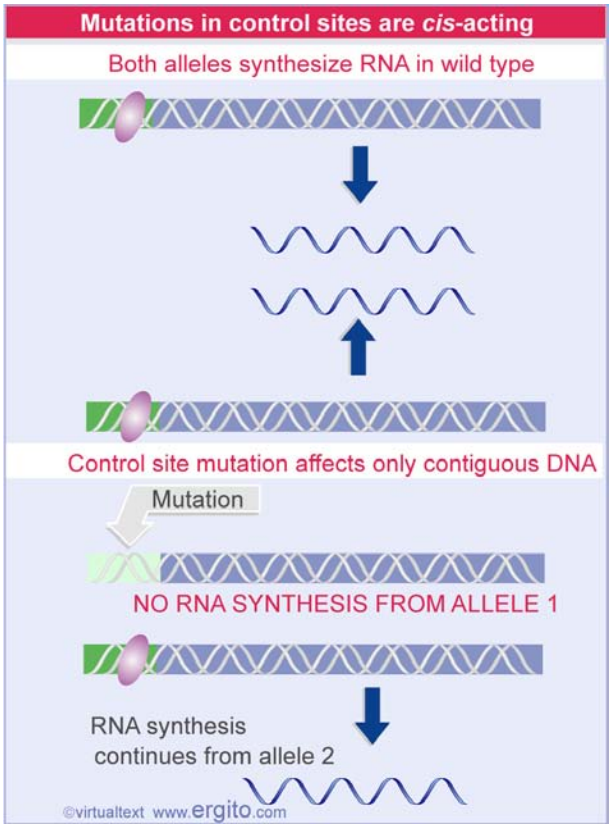

Figure 1.41 A *cis*-acting site controls the adjacent DNA but does not influence the other allele.

We may contrast the behavior of the *cis*-acting mutation shown in **Figure 1.41** with the result of a mutation in the gene coding for the regulator protein. **Figure 1.42** shows that the absence of regulator protein would prevent *both* alleles from being expressed. A mutation of this sort is said to be *trans*-acting.

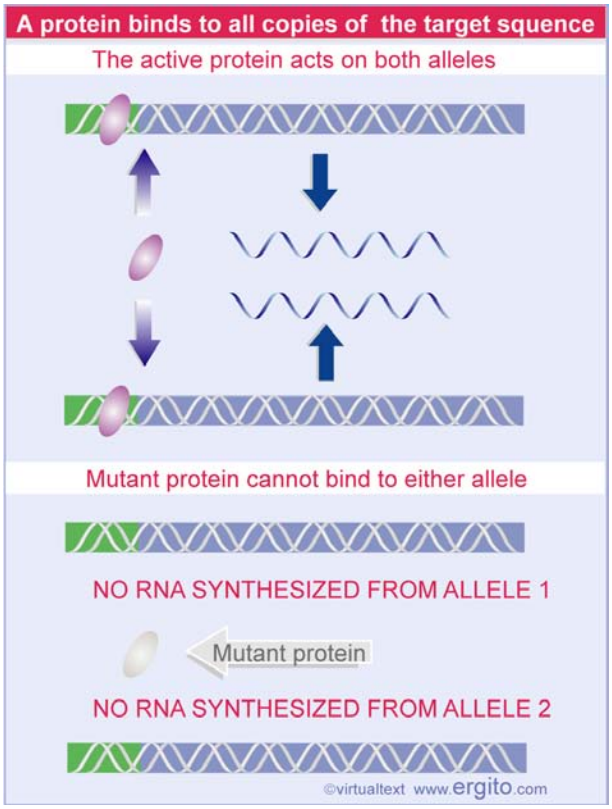

Figure 1.42 A *trans*-acting mutation in a protein affects both alleles of a gene that it controls.

Reversing the argument, if a mutation is *trans*-acting, we know that its effects must be exerted through some diffusible product (typically a protein) that acts on multiple targets within a cell. But if a mutation is *cis*-acting, it must function via affecting directly the properties of the contiguous DNA, which means that it is *not expressed in the form of RNA or protein*.

Last updated on January 15, 2004

GENES ARE DNA

1.1.26 Genetic information can be provided by DNA or RNA

Key Terms

The **central dogma** describes the basic nature of genetic information: sequences of nucleic acid can be perpetuated and interconverted by replication, transcription, and reverse transcription, but translation from nucleic acid to protein is unidirectional, because nucleic acid sequences cannot be retrieved from protein sequences.

A **retrovirus** is an RNA virus with the ability to convert its sequence into DNA by reverse transcription.

Reverse transcription is synthesis of DNA on a template of RNA. It is accomplished by the enzyme reverse transcriptase.

Key Concepts

- Cellular genes are DNA, but viruses and viroids may have genes of RNA.
- DNA is converted into RNA by transcription, and RNA may be converted into DNA by reverse transcription.
- The translation of RNA into protein is unidirectional.

The **central dogma** defines the paradigm of molecular biology. Genes are perpetuated as sequences of nucleic acid, but function by being expressed in the form of proteins. Replication is responsible for the inheritance of genetic information. Transcription and translation are responsible for its conversion from one form to another.

Figure 1.43 illustrates the roles of replication, transcription, and translation, viewed from the perspective of the central dogma:

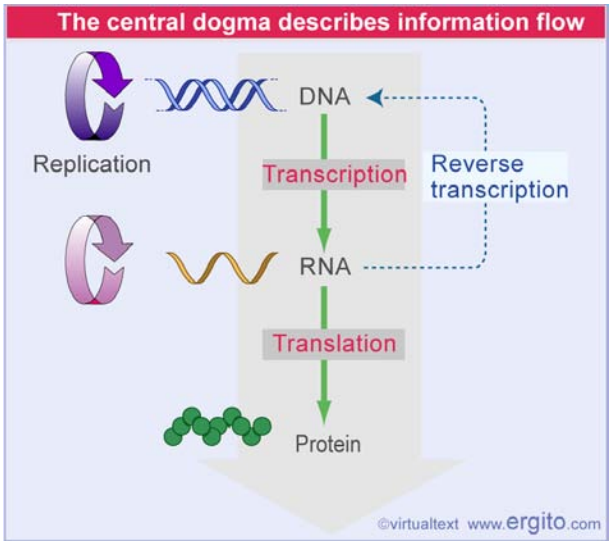

Figure 1.43 The central dogma states that information in nucleic acid can be perpetuated or transferred, but the transfer of information into protein is irreversible.

- *The perpetuation of nucleic acid may involve either DNA or RNA as the genetic material*. Cells use only DNA. Some viruses use RNA, and replication of viral RNA occurs in the infected cell.
- *The expression of cellular genetic information usually is unidirectional.* Transcription of DNA generates RNA molecules that can be used further *only* to generate protein sequences; generally they cannot be retrieved for use as genetic information. Translation of RNA into protein is always irreversible.

These mechanisms are equally effective for the cellular genetic information of prokaryotes or eukaryotes, and for the information carried by viruses. The genomes of all living organisms consist of duplex DNA. Viruses have genomes that consist of DNA or RNA; and there are examples of each type that are double-stranded (ds) or single-stranded (ss). Details of the mechanism used to replicate the nucleic acid vary among the viral systems, but the principle of replication via synthesis of complementary strands remains the same, as illustrated in **Figure 1.44**.

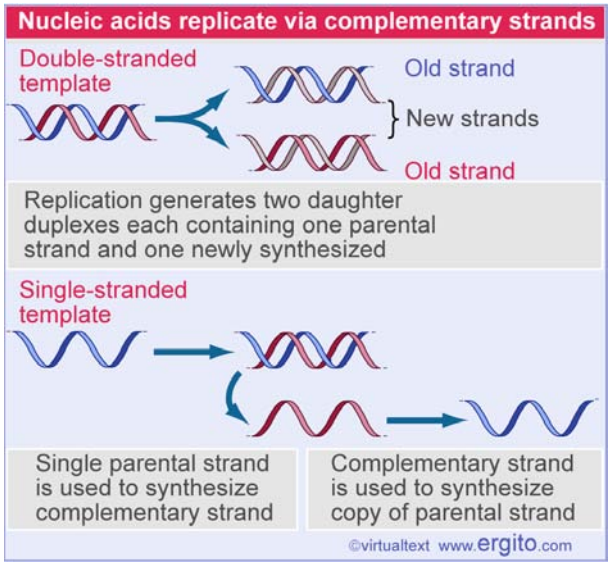

Figure 1.44 Double-stranded and single-stranded nucleic acids both replicate by synthesis of complementary strands governed by the rules of base pairing.

Cellular genomes reproduce DNA by the mechanism of semi-conservative replication. Double-stranded virus genomes, whether DNA or RNA, also replicate by using the individual strands of the duplex as templates to synthesize partner strands.

Viruses with single-stranded genomes use the single strand as template to synthesize a complementary strand; and this complementary strand in turn is used to synthesize its complement, which is, of course, identical with the original starting strand. Replication may involve the formation of stable double-stranded intermediates or use double-stranded nucleic acid only as a transient stage.

The restriction to unidirectional transfer from DNA to RNA is not absolute. It is overcome by the **retroviruses**, whose genomes consist of single-stranded RNA molecules. During the infective cycle, the RNA is converted by the process of **reverse transcription** into a single-stranded DNA, which in turn is converted into a double-stranded DNA. This duplex DNA becomes part of the genome of the cell, and is inherited like any other gene. *So reverse transcription allows a sequence of RNA to be retrieved and used as genetic information.*

The existence of RNA replication and reverse transcription establishes the general principle that *information in the form of either type of nucleic acid sequence can be converted into the other type.* In the usual course of events, however, the cell relies on the processes of DNA replication, transcription, and translation. But on rare occasions (possibly mediated by an RNA virus), information from a cellular RNA is converted into DNA and inserted into the genome. Although reverse transcription plays no role in the regular operations of the cell, it becomes a mechanism of potential importance when we consider the evolution of the genome.

The same principles are followed to perpetuate genetic information from the massive genomes of plants or amphibians to the tiny genomes of mycoplasma and the yet smaller genetic information of DNA or RNA viruses. **Figure 1.45** summarizes some

examples that illustrate the range of genome types and sizes.

Genomes have nucleic acids

Genome	Gene Number	Base Pairs
Organisms		
Plants	<50,000	$<10^{11}$
Mammals	30,000	~3 x 10^9
Worms	14,000	~10^8
Flies	12,000	1.6 x 10^8
Fungi	6,000	1.3 x 10^7
Bacteria	2-4,000	$<10^7$
Mycoplasma	500	$<10^6$
dsDNA Viruses		
Vaccinia	<300	187,000
Papova (SV40)	~6	5,226
Phage T4	~200	165,000
ssDNA Viruses		
Parvovirus	5	5,000
Phage fX174	11	5,387
dsRNA Viruses		
Reovirus	22	23,000
ssRNA Viruses		
Coronavirus	7	20,000
Influenza	12	13,500
TMV	4	6,400
Phage MS2	4	3,569
STNV	1	1,300
Viroids		
PSTV RNA	0	359

Figure 1.45 The amount of nucleic acid in the genome varies over an enormous range.

Throughout the range of organisms, with genomes varying in total content over a 100,000 fold range, a common principle prevails. *The DNA codes for all the proteins that the cell(s) of the organism must synthesize; and the proteins in turn (directly or indirectly) provide the functions needed for survival.* A similar principle describes the function of the genetic information of viruses, whether DNA or RNA. *The nucleic acid codes for the protein(s) needed to package the genome and also for any functions additional to those provided by the host cell that are needed to reproduce the virus during its infective cycle.* (The smallest virus, the satellite tobacco necrosis virus [STNV], cannot replicate independently, but requires the simultaneous presence of a "helper" virus [tobacco necrosis virus, TNV], which is itself a normally infectious virus.)

GENES ARE DNA

1.1.27 Some hereditary agents are extremely small

Key Terms

A **viroid** is a small infectious nucleic acid that does not have a protein coat.

Virion is the physical virus particle (irrespective of its ability to infect cells and reproduce).

A **subviral pathogen** is an infectious agent that is smaller than a virus, such as a virusoid.

Scrapie is a infective agent made of protein.

A **prion** is a proteinaceous infectious agent, which behaves as an inheritable trait, although it contains no nucleic acid. Examples are PrP^{Sc}, the agent of scrapie in sheep and bovine spongiform encephalopathy, and Psi, which confers an inherited state in yeast.

PrP is the protein that is the active component of the prion that causes scrapie and related diseases. The form involved in the disease is called PrP^{Sc}.

Key Concepts

- Some very small hereditary agents do not code for protein but consist of RNA or of protein that has hereditary properties.

Viroids are infectious agents that cause diseases in higher plants (for review see 2525). They are very small circular molecules of RNA. Unlike viruses, where the infectious agent consists of a **virion**, a genome encapsulated in a protein coat, *the viroid RNA is itself the infectious agent.* The viroid consists solely of the RNA, which is extensively but imperfectly base paired, forming a characteristic rod like the example shown in **Figure 1.46**. Mutations that interfere with the structure of the rod reduce infectivity.

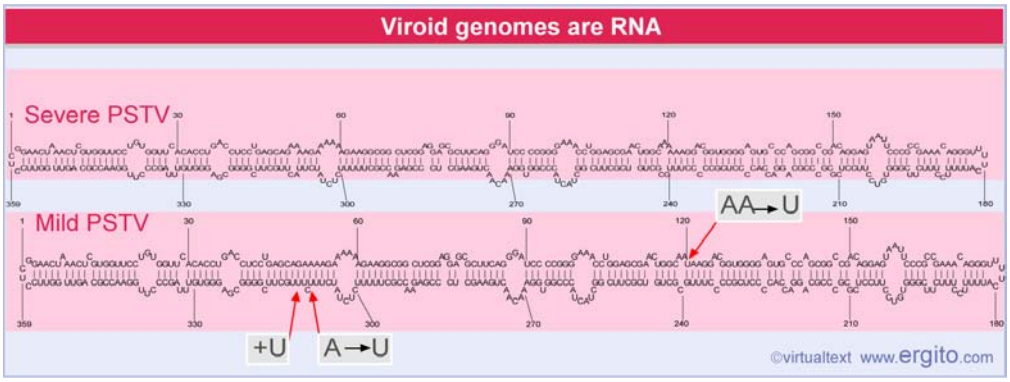

Figure 1.46 PSTV RNA is a circular molecule that forms an extensive double-stranded structure, interrupted by many interior loops. The severe and mild forms differ at three sites.

A viroid RNA consists of a single molecular species that is replicated autonomously

in infected cells. Its sequence is faithfully perpetuated in its descendants. Viroids fall into several groups. A given viroid is identified with a group by its similarity of sequence with other members of the group. For example, four viroids related to PSTV (potato spindle tuber viroid) have 70-83% similarity of sequence with it. Different isolates of a particular viroid strain vary from one another, and the change may affect the phenotype of infected cells. For example, the *mild* and *severe* strains of PSTV differ by three nucleotide substitutions.

Viroids resemble viruses in having heritable nucleic acid genomes. They fulfill the criteria for genetic information. Yet viroids differ from viruses in both structure and function. They are sometimes called **subviral pathogens**. Viroid RNA does not appear to be translated into protein. So it cannot itself code for the functions needed for its survival. This situation poses two questions. How does viroid RNA replicate? And how does it affect the phenotype of the infected plant cell?

Replication must be carried out by enzymes of the host cell, subverted from their normal function. The heritability of the viroid sequence indicates that viroid RNA provides the template.

Viroids are presumably pathogenic because they interfere with normal cellular processes. They might do this in a relatively random way, for example, by sequestering an essential enzyme for their own replication or by interfering with the production of necessary cellular RNAs. Alternatively, they might behave as abnormal regulatory molecules, with particular effects upon the expression of individual genes (for review see 12).

An even more unusual agent is **scrapie**, the cause of a degenerative neurological disease of sheep and goats. The disease is related to the human diseases of kuru and Creutzfeldt-Jakob syndrome, which affect brain function.

The infectious agent of scrapie does not contain nucleic acid. This extraordinary agent is called a **prion** (proteinaceous infectious agent) (for review see 2523). It is a 28 kD hydrophobic glycoprotein, **PrP**. PrP is coded by a cellular gene (conserved among the mammals) that is expressed in normal brain. The protein exists in two forms. The product found in normal brain is called PrP^{c}. It is entirely degraded by proteases. The protein found in infected brains is called PrP^{sc}. It is extremely resistant to degradation by proteases. PrP^{c} is converted to PrP^{sc} by a modification or conformational change that confers protease-resistance, and which has yet to be fully defined (383).

As the infectious agent of scrapie, PrP^{sc} must in some way modify the synthesis of its normal cellular counterpart so that it becomes infectious instead of harmless (see *Molecular Biology 5.23.24 Prions cause diseases in mammals*). Mice that lack a PrP gene cannot be infected to develop scrapie, which demonstrates that PrP is essential for development of the disease (386).

Reviews

12. Diener, T. O. (1986). *Viroid processing: a model involving the central conserved region and hairpin.* Proc. Natl. Acad. Sci. USA 83, 58-62.

2523. Prusiner, S. B. (1998). *Prions.* Proc. Natl. Acad. Sci. USA 95, 13363-13383.

2525. Diener, T. O. (1999). *Viroids and the nature of viroid diseases.* Arch. Virol. Suppl. 15, 203-220.

References

383. McKinley, M. P., Bolton, D. C., and Prusiner, S. B. (1983). *A protease-resistant protein is a structural component of the scrapie prion*. Cell 35, 57-62.

386. Bueler, H. et al. (1993). *Mice devoid of PrP are resistant to scrapie*. Cell 73, 1339-1347.

GENES ARE DNA

1.1.28 Summary

Two classic experiments proved that DNA is the genetic material. DNA isolated from one strain of *Pneumococcus* bacteria can confer properties of that strain upon another strain. And DNA is the only component that is inherited by progeny phages from the parental phages. DNA can be used to transfect new properties into eukaryotic cells.

DNA is a double helix consisting of antiparallel strands in which the nucleotide units are linked by 5′–3′ phosphodiester bonds. The backbone provides the exterior; purine and pyrimidine bases are stacked in the interior in pairs in which A is complementary to T while G is complementary to C. The strands separate and use complementary base pairing to assemble daughter strands in semiconservative replication. Complementary base pairing is also used to transcribe an RNA representing one strand of a DNA duplex.

A stretch of DNA may code for protein. The genetic code describes the relationship between the sequence of DNA and the sequence of the protein. Only one of the two strands of DNA codes for protein. A codon consists of three nucleotides that represent a single amino acid. A coding sequence of DNA consists of a series of codons, read from a fixed starting point. Usually only one of the three possible reading frames can be translated into protein.

A chromosome consists of an uninterrupted length of duplex DNA that contains many genes. Each gene (or cistron) is transcribed into an RNA product, which in turn is translated into a polypeptide sequence if the gene codes for protein. An RNA or protein product of a gene is said to be *trans*-acting. A gene is defined as a unit on a single stretch of DNA by the complementation test. A site on DNA that regulates the activity of an adjacent gene is said to be *cis*-acting.

A gene may have multiple alleles. Recessive alleles are caused by a loss-of-function. A null allele has total loss-of-function. Dominant alleles are caused by gain-of-function.

A mutation consists of a change in the sequence of A·T and G·C base pairs in DNA. A mutation in a coding sequence may change the sequence of amino acids in the corresponding protein. A frameshift mutation alters the subsequent reading frame by inserting or deleting a base; this causes an entirely new series of amino acids to be coded after the site of mutation. A point mutation changes only the amino acid represented by the codon in which the mutation occurs. Point mutations may be reverted by back mutation of the original mutation. Insertions may revert by loss of the inserted material, but deletions cannot revert. Mutations may also be suppressed indirectly when a mutation in a different gene counters the original defect.

The natural incidence of mutations is increased by mutagens. Mutations may be concentrated at hotspots. A type of hotspot responsible for some point mutations is caused by deamination of the modified base 5-methylcytosine.

Forward mutations occur at a rate of ~10^{-6} per locus per generation; back mutations are rarer. Not all mutations have an effect on the phenotype.

Although all genetic information in cells is carried by DNA, viruses have genomes of double-stranded or single-stranded DNA or RNA. Viroids are subviral pathogens that consist solely of small circular molecules of RNA, with no protective packaging. The RNA does not code for protein and its mode of perpetuation and of pathogenesis is unknown. Scrapie consists of a proteinaceous infectious agent.