CHAPTER

2

Fundamentals of Molecular Evolution*

OUTLINE

2.1 BIOINFORMATICS, MOLECULAR EVOLUTION, AND PHYLOGENETICS

Probably, the shortest classical definition of evolution is *descent with modification* from the ancestor. Evolutionary changes lead to changes in the inherited characters in a population[a]. The ultimate outcome of evolution is the formation of new species (**speciation**), but evolution can generate diversity at all possible levels of biological organization including at the level of macromolecules, such as DNA and proteins.

Molecular evolution is a relatively recent discipline that has developed since DNA and protein sequence information became available. Simply stated, molecular

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

[a]A population is composed of members of a species occupying a geographic area. A community is composed of members of different populations occupying the same geographic area.

evolution is evolution at the level of nucleic acids and proteins. At the molecular level, the primary cause of evolution is the accumulation of changes in genomic sequence (hence proteins as well[b]). Therefore, evolution results in alteration of the genetic composition (**gene pool**) of a population over time. Changes in gene pool are associated with changes in gene frequency in a population[c].

The work of Emile Zuckerkandl and Linus Pauling between 1960 and 1965, particularly their seminal publication in 1965,[1] is credited with ushering in a change in evolutionary thinking from the level of species to the level of macromolecular sequence. Such a paradigm shift in evolutionary thinking from population to macromolecular sequence essentially paved the way for the birth of a new field, molecular evolution. The classical definition of evolution as *descent with modification* refers to the event of speciation—that is, the formation of new species from an ancestral species. The same definition and concepts also apply to molecular evolution except for the fact that the targets of molecular evolution are nucleic acid and protein sequences. The causes of molecular evolution, such as mutation, recombination, gene conversion, duplication and divergence of genes, de novo origin of new genes, and structural and functional evolution of genomes, as well as changes in gene frequency in a population, are also at the heart of evolution at the level of species and beyond.

The availability of the complete genome sequence of many species provides a wealth of data and information for molecular evolutionary studies and comparative genomics. *Evolutionary biology provides the scientific context and bioinformatic analysis utilizes the analytical tools for comparative genomics*. In the context of evolutionary biology, the goal of various applications of bioinformatics, such as sequence alignment, sequence identity/similarity search, motif analysis, sequence homology analysis, chromosomal synteny analysis, and making phylogenetic trees, is to trace the signature and determine the rate of molecular evolution, as well as study the relatedness of taxa. Following the spirit of the now-famous statement by Dobzhansky that "nothing in biology makes sense except in the light of evolution, "Higgs and Attwood (2005) have stated, "nothing in bioinformatics makes sense except in the light of evolution".[2] This is a very astute way of summarizing the relationship between bioinformatics and molecular evolution.

It has become a standard practice in studies involving DNA or protein sequence to obtain a phylogenetic tree and assess sequence divergence. Freely available software on the web has made it almost effortless to input the data and quickly get an output. Because of such widespread use of DNA and protein sequence analysis and phylogenetic inference, it is important to understand the principles of molecular evolution. The following narrative summarizes some fundamental concepts of molecular evolution that help in understanding the evolutionary foundations of bioinformatics.

2.2 BIOLOGICAL EVOLUTION AND BASIC PREMISES OF DARWINISM

Biological evolution is most simply defined as *descent with modification*; the modification may be small scale (e.g. changes in gene/protein sequence) or large scale (e.g. speciation). After life had originated on Earth about 3.6 billion (3600 million) years ago, it evolved from simple to progressively complex forms, all from one primordial ancestral form, called the **last universal common ancestor** (**LUCA**). The evolutionary history of the descendants of LUCA constitutes the **tree of life**.

Evolution of life is a continuous process involving splitting of lineages, divergence of the descendants, and adaptive radiation into different environments (ecological niches) creating phenotypic diversity, and ultimately leading to reproductive isolation and the formation of new species (**speciation**). It is important to note in this context that even though "species" is an accepted taxonomic category, the concept of species and speciation is a hotly debated issue even 150 years after the publication of Darwin's *On the Origin of Species*. We will follow the most widely used definition of species, provided by the **biological species concept**.

Two pioneering architects of the biological species concept were Theodosius Dobzhansky and Ernst Mayr. According to Mayr's classical definition of species, "species are groups of actually or potentially interbreeding natural populations that are reproductively isolated from other such groups"[d].[3] In other

[b]Changes in genomic sequence include changes in the sequence of protein-coding genes, non protein-coding genes, and regulatory sequences, as well as intergenic regions. Such changes may result in altered gene expression and trigger genome evolution.

[c]A small-scale change within a population below the species level, such as a change in allele frequencies, is called microevolution. Microevolution can be observed over a short period of time, such as across a few generations (e.g. development of resistance). In contrast, large-scale changes and evolution at or above the species level and over a long period of time are called macroevolution.

[d]This definition of species was originally proposed in Mayr's now-classic book *Systematics and the Origin of Species* (1942, Columbia University Press, New York). However, Mayr's definition of species owed its origin to the concept of species proposed by Dobzhansky in his famous book *Genetics and the Origin of Species* (1937, Columbia University Press, New York). Dobzhansky conceptualized species as "that stage in the evolutionary process at which the once actually or potentially interbreeding array of forms becomes segregated in two or more separate arrays which are physiologically incapable of interbreeding."

words, a species is a reproductive community that represents a unique gene pool. Genetic exchange between members of two different gene pools is usually not successful in producing fertile offspring that could perpetuate the existence of the species. When populations within a species become isolated by geography, mate selection, or other means that interfere with mating, they may start to diverge and over time may evolve into new species.

Darwin's theory of evolution by natural selection states that (1) variations exist among the organisms of a population, (2) the resources (food and space) are limited, (3) the scarcity of resources would lead to competition among individuals, and (4) individuals with favorable variations are more likely to survive in the competition whereas those that do not have the favorable variations simply die out. Those that survive will reproduce, increase in number, and occupy a specific environment. This process, which removes some organisms from the population but favors (selects) others, is called **natural selection** and it is a **passive process** acting like a sieve. Natural selection could be **purifying** (**negative**) **selection** that removes deleterious variations, and **positive** (**Darwinian**) **selection** that fixes the beneficial variations in the population and promotes the emergence of new phenotypes. When the organisms with favorable variations reproduce, the variations spread in the population and help the population to better adapt to the environment. Over many generations, the population adapted to a specific environment evolves into a new species that becomes reproductively isolated from other such groups. The coupling of Darwinism with modern genetics transformed classical Darwinism into **neo-Darwinism** (also known as **modern synthesis** or the **synthetic theory of evolution**).

The Darwinian evolutionary process predicts that the pace of evolution is gradual because an evolving population accumulates small variations over a long period of time. Hence, the divergence of lineages is slow, steady, and stepwise. For example, for a species A to evolve into species B, it should go through many stages, such as A_1, A_2, A_3 ... A_n until it evolves into B. This gradual pace of evolution through incremental changes is known as **phyletic gradualism**. However, the fossil records for most species are incomplete and they do not show the existence of small incremental changes on the way to the new species[e]. To account for the lack of fossil records showing phyletic gradualism, paleontologists Stephen J. Gould and Niles Elredge[4] put forth a competing hypothesis, which claims that species are generally stable, changing little over long periods of time. This condition of little or no change is called **stasis**. The stasis is punctuated by rapid bursts of evolutionary changes that result in the formation of new species. As a result, this process leaves few fossils behind, which can explain the absence of many intermediate forms in the fossil record. Gould and Elredge termed this phenomenon **punctuated equilibrium**. In reality, both phyletic gradualism and punctuated equilibrium could have played a role in evolution.

A basic assumption of the Darwinian theory is that new mutations, both advantageous and deleterious, constantly arise in the population independent of need, and *evolution is caused by natural selection acting through beneficial mutations by fixing them in the population*. Darwinian evolution does not consider neutral mutations that do not confer any selective advantage or disadvantage to be of any importance in the evolutionary process. This long-held view of Darwinian evolution was challenged by the neutral theory of molecular evolution. The neutral theory is discussed later in this chapter.

2.2.1 First Experimental Demonstration of Evolutionary Principles in the Test Tube

Sol Spiegelman and colleagues[5] first demonstrated that **Darwinian evolutionary principles**—that is, **variation**, **selection**, and **amplification**—could lead to the evolution of biological macromolecules in the test tube in an extracellular environment. Spiegelman and coworkers explored the evolutionary consequences for a self-duplicating nucleic acid molecule put under selection pressure for faster growth. Bacteriophage Qβ is an RNA phage with an RNA genome (~3500 nucletotides (nt)) that codes for four proteins: viral coat protein, attachment protein, maturation protein, and β1 replicase, also called Qβ-replicase, which is an RNA-dependent RNA polymerase. When Qβ-replicase is incubated with Qβ-RNA template in the presence of ribonucleotides, it synthesizes new Qβ-RNA molecules.

The goal of the experiment was to determine how molecules evolve if the selection pressure is allowed to only select for molecules that can multiply increasingly faster. The experimental procedure involved serial transfer of the reaction mix in which the incubation time was progressively reduced over time. The first

[e]Among living species, the fossil record of the modern-day horse from *Hyracotherium* (previously known as *Eohippus*) to *Equus*, spanning a period of about 55 million years, is one of the better-preserved fossil records that show macroevolutionary changes. Most fossil records are not as well preserved.

reaction was allowed to proceed for 20 minutes, after which an aliquot was used to start the second reaction, and so on for the first 13 reactions. After the first 13 reactions, the incubation periods were reduced to 15 min (transfers 14–29), 10 min (transfers 30–38), 7 min (transfers 39–52), and 5 min (transfers 53–74). The progressive reduction in the incubation intervals between transfers maintained the selection pressure for the evolution of the most rapidly multiplying RNA template molecules. As the experiment progressed, the rate of RNA synthesis increased and the product became smaller. By the 74th transfer, the size of the replicating molecule had become ~17% of its original size by deleting most of the original genome, and replicated 15 times faster than the complete viral RNA. This short RNA template variant was found to have experienced a significant change in base composition as well. The fact that this RNA template variant replicated 15 times faster than the complete viral RNA suggested that in addition to becoming smaller, the variant increased the efficiency with which it interacted with the replicase. Therefore, the RNA molecules adapted to the new conditions by throwing away anything not needed for fast replication[f].

It should be emphasized in this context that Spiegelman's experiment was a demonstration of **directed evolution** because selection pressure was applied to achieve a predetermined evolutionary outcome. The goal of Spiegelman's experiment as stated by Mills et al. was, "What will happen to the RNA molecules if the only demand made on them is the Biblical injunction, multiply, with the biological proviso that they do so as rapidly as possible?" In contrast, natural evolutionary processes are not directed. Genetic variations are random and spontaneous; hence they arise in the population independent of need. The advantages or disadvantages of such variations become apparent only when selection pressure arises. Thus, the natural evolutionary process works as a **blind watchmaker**, as Richard Dawkins calls it to underscore the lack of purpose and direction in the process. However, in recent years, the concept of directed (adaptive) mutation and directed evolution in bacteria, originally proposed in 1988 by John Cairns and coworkers,[6] has garnered some support. This idea is still not mainstream in evolutionary biology and is beyond the scope of this book.

Since the experiment of Spiegelman, many more extracellular Darwinian experiments have been conducted to direct the evolution of desired traits in biological macromolecules, and many laboratories have reported some remarkable findings.

2.3 MOLECULAR BASIS OF HERITABLE GENETIC VARIATIONS—THE RAW MATERIALS FOR EVOLUTION

Genetic variations in a population evolve irrespective of need. Most genetic variations are deleterious or at best neutral, but some may be beneficial in a specific environment. It is the selection pressure that reveals the utility of a beneficial genetic variation. Four important sources of molecular genetic variations are mutation, recombination, gene flow, and creation of new genes.

2.3.1 Molecular Basis of Mutation

Mutation is the change of genomic sequence. Mutation can be a **point mutation** (alteration of just one nucleotide), a **frameshift mutation** (alteration of the open reading frame (ORF) of the gene), or a **chromosomal mutation**—that is, large-scale alterations of the chromosomal DNA (**insertion**, **deletion**, **inversion**, **duplication**, **translocation**) (Figure 2.1A). Chromosomal mutations can result in gene duplication and divergence, exon shuffling, retrotransposition, gene fission/fusion, and gene deletion; each of these events creates genetic diversity.

Based on the effect on the polypeptide product, a point mutation can be missense, nonsense, or silent. A **missense point mutation** changes an amino acid in the polypeptide; a **nonsense point mutation** creates a stop codon, thereby prematurely truncating the ORF and ending translation of the polypeptide; a **silent point mutation** does not change the amino acid sequence of the polypeptide (Figure 2.1B). Splice donor or acceptor site mutations as well as splicing signal site mutations can result in the exonization of a previous intron sequence or intronization of a previous exon sequence; these types of mutations frequently have pathological consequences. There are a number of reports in the literature describing such mutations.

Based on the type of base altered, a point mutation can be classified as a **transition** or a **transversion** mutation. A pyrimidine replaced by another pyrimidine (C→T or T→C) or a purine replaced by another purine (A→G or G→A) is a transition mutation. A common mechanism of transition mutations is the formation of tautomeric forms (amino→imino tautomer as occurs in A and C; and keto→enol tautomer as occurs in G and T), and mispairing of bases (Figure 2.1C). If the mispairing survives the DNA repair machinery (e.g. if the mispairing occurs during replication), then by the following replication cycle the

[f]The small, rapidly duplicating RNA template variant was later termed the **Spiegelman monster**.

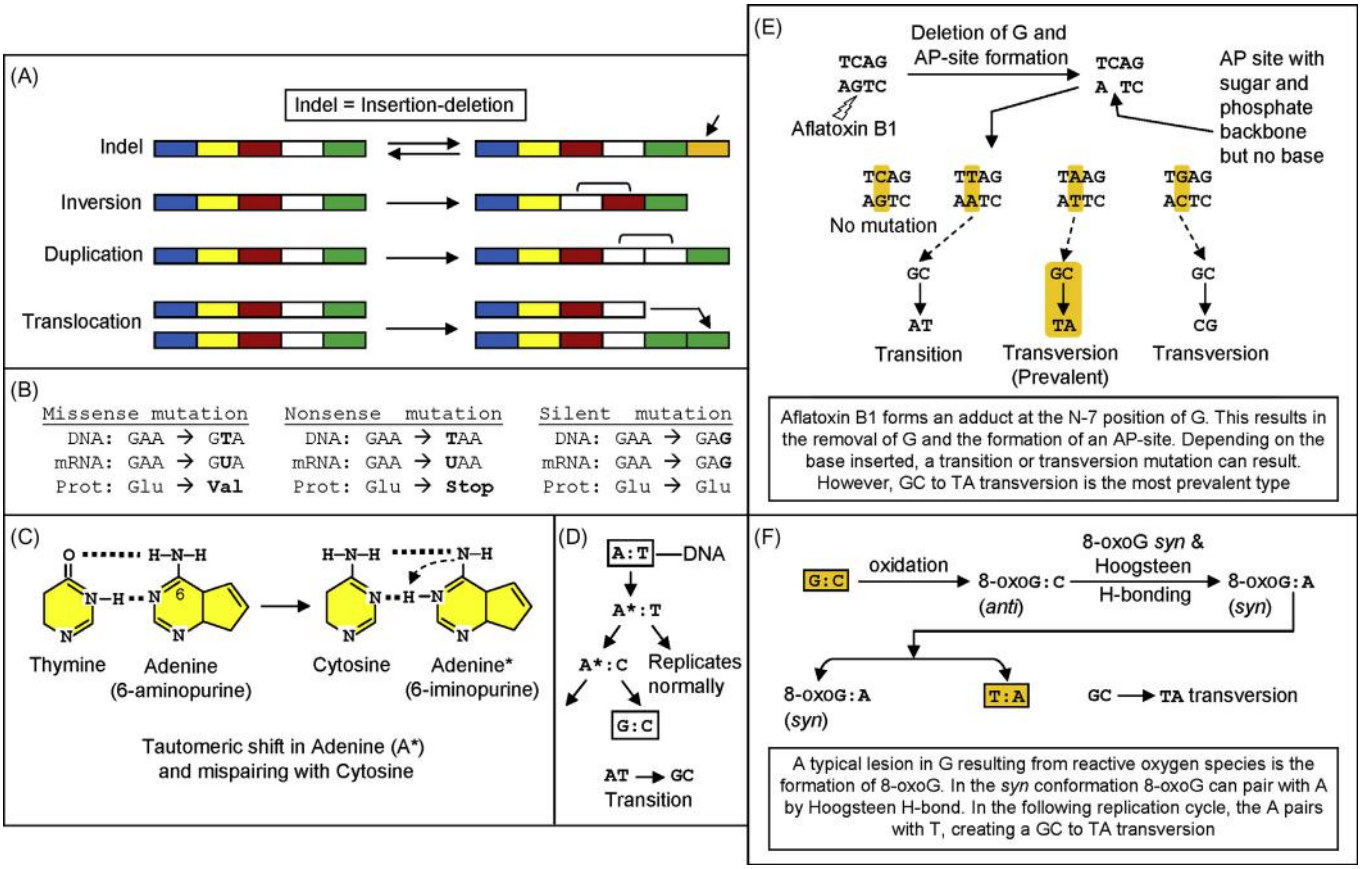

FIGURE 2.1 **Molecular basis of mutation.** (A) Various types of mutations affecting long DNA fragments, i.e. a chromosome. (B) Various effects of a one-base-pair mutation in DNA (only sense strand is shown). A missense mutation alters the amino acid sequence of a protein; a nonsense mutation disrupts the ORF and prematurely stops translation, whereas a silent mutation does not change the amino acid sequence of the protein. (C) Mechanism of transition mutation due to tautomeric shift in adenine resulting in 6-iminopurine from 6-aminopurine. (D) Wrong base pairing by imino tautomer of adenine results in AT-to-GC transition mutation in two replication cycles. (E) The mechanism of aflatoxin-B1-mediated transversion mutation (see text for details). (F) The mechanism of 8-oxoG-mediated transversion mutation (see text for details).

affected position of DNA has the base pair replaced by transition mutation (Figure 2.1D). Another mechanism of transition mutation in genomes is the spontaneous oxidative deamination of methylated C to form T, resulting in CG→TA transition over time. In contrast to transition mutation, a purine replaced by a pyrimidine or a pyrimidine replaced by a purine is a transversion mutation. Chemicals such as aflatoxin B1 can cause transversion mutation through adduct formation. Aflatoxin B1 forms an adduct at the N-7 position of guanine. This ultimately results in the removal of G and the formation of an AP-site (apurinic site). Depending on the base inserted for repair, a transition or transversion mutation can result. However, GC→TA transversion is the most prevalent type (Figure 2.1E).[7] Oxidation of guanine can also lead to transversion. A typical lesion in guanine resulting from oxidative stress is the formation of 8-oxoG. The 8-oxoG lesion in DNA is normally repaired by the dedicated enzyme 8-oxoG DNA glycosylase, which removes the oxoG with the concomitant cleavage of the DNA backbone. If the removal fails to take place, 8-oxoG tends to form the *syn* conformer, which then pairs with A by Hoogsteen H-bond during replication. In the following replication cycle, the A pairs with T, creating a GC→TA transversion (Figure 2.1F).[8] As mentioned above, *transition mutations are far more prevalent than transversion mutations*. In earlier literature, a point mutation was called a **single nucleotide polymorphism** (**SNP**) if it occurred in at least 1% of the population, but currently, any point mutation is regarded as an SNP. In the human genome, >65% of all SNPs are C→T transition mutations. SNPs and **copy number variations** (**CNVs**, also called **copy number polymorphisms or CNPs**) together constitute a significant source of inter-individual variation in a population.

In addition to the classical mutations described above, expansion or contraction of repeat sequences

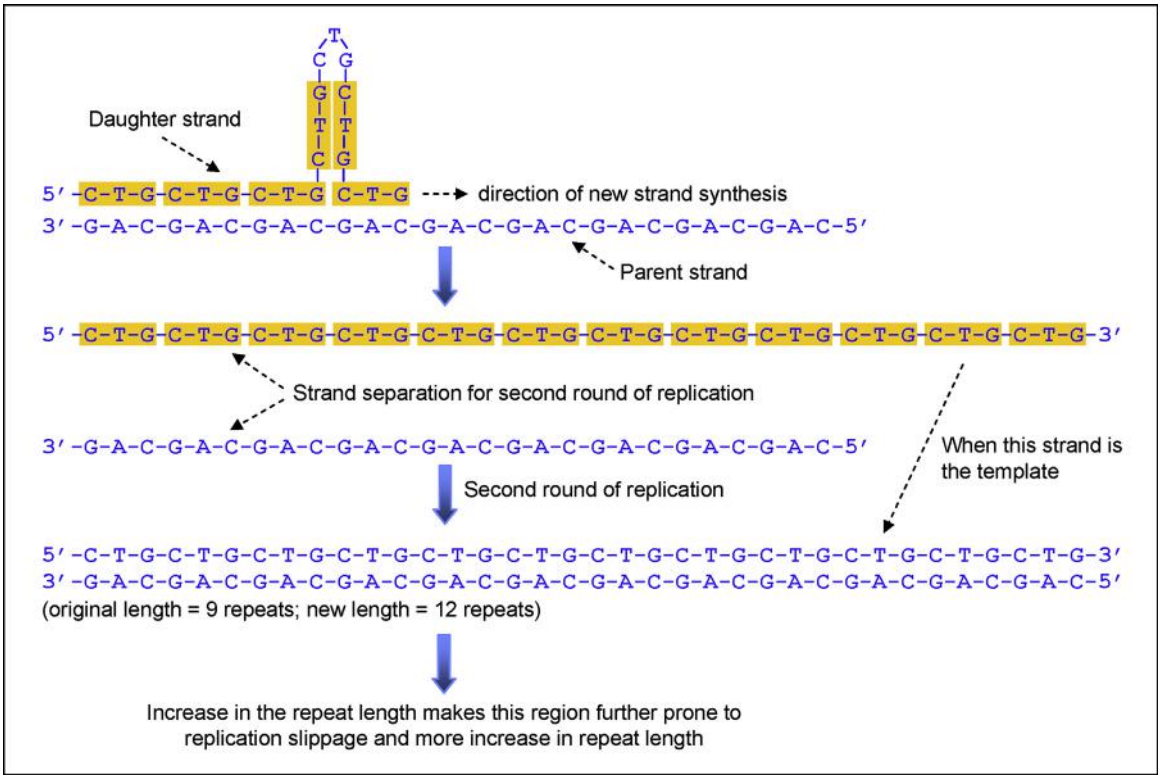

FIGURE 2.2 **Mechanism of expansion of triplet repeats through replication slippage.** The −C−T−G− triplet repeats in the gene are highlighted except the one forming loop. The increase in the number of repeats through replication slippage is a random process; it may be as few as one triplet or it may be multiple triplets. The figure shows an increase of three −C−T−G− triplet repeats in the gene in two rounds of replication. The strand of DNA containing the −C−T−G− triplets (highlighted) is the sense strand; therefore, the mRNA will have the same repeats as −C−U−G−.

constitutes another class of mutations. Repeat sequences in DNA can be expanded during replication. Two mechanisms can result in the expansion of repeat sequences: **replication slippage** (also called **slipped strand mispairing**) and **unequal crossing over**. In replication slippage, a long stretch of repeat sequences in the DNA folds back and pairs on itself, forming an internal hairpin or stem−loop structure, during replication. As a result, there is a net increase in the repeat sequences following replication in the daughter strand while the repeat length in the parent strand remains the same. The increased length of one strand propagates through subsequent rounds of replication (Figure 2.2). Misalignment of DNA involving blocks of the same repeat sequences may also occur during crossing over (unequal crossing over). As a result, in one chromosome the repeat length increases (insertion) while in the other chromosome it decreases (deletion), as shown in Figure 2.3.

The presence of uninterrupted trinucleotide repeats (triplet repeats) makes the sequence unstable and prone to further expansion through replication slippage. Increased numbers of triplet repeats are associated with a number of heritable genetic disorders in humans, such as Huntington's disease (CAG repeats), myotonic dystrophy (CTG repeats), fragile-X syndrome (CGG repeats). A higher number of uninterrupted triplet repeats is usually correlated with an earlier onset and a greater severity of the disease. *In contrast, interruption of the triplet repeats may reduce the predisposition of the carrier to the disease*. For example, fragile-X syndrome in humans is associated with the expansion of the CGG triplet repeats in the *FMR1* (fragile-X mental retardation 1) gene. However, if these CGG repeats are interspersed with AGG triplet repeats, the predisposition towards developing the disease is significantly reduced.[9] Populations that have a disproportionately large number of uninterrupted CGG-repeat-containing alleles, such as the Tunisian Jews, have a much higher incidence of fragile-X syndrome.[10]

Most mammals possess a small number of the CGG repeats in the *FMR1* gene (mean = 8 ± 0.8), but primates have a greater number of repeats (mean = 20 ± 2.3). Interestingly, nonhuman primates do not have fragile sites in the *FMR1* gene because they have many more interruptions in the CGG sequences.[11]

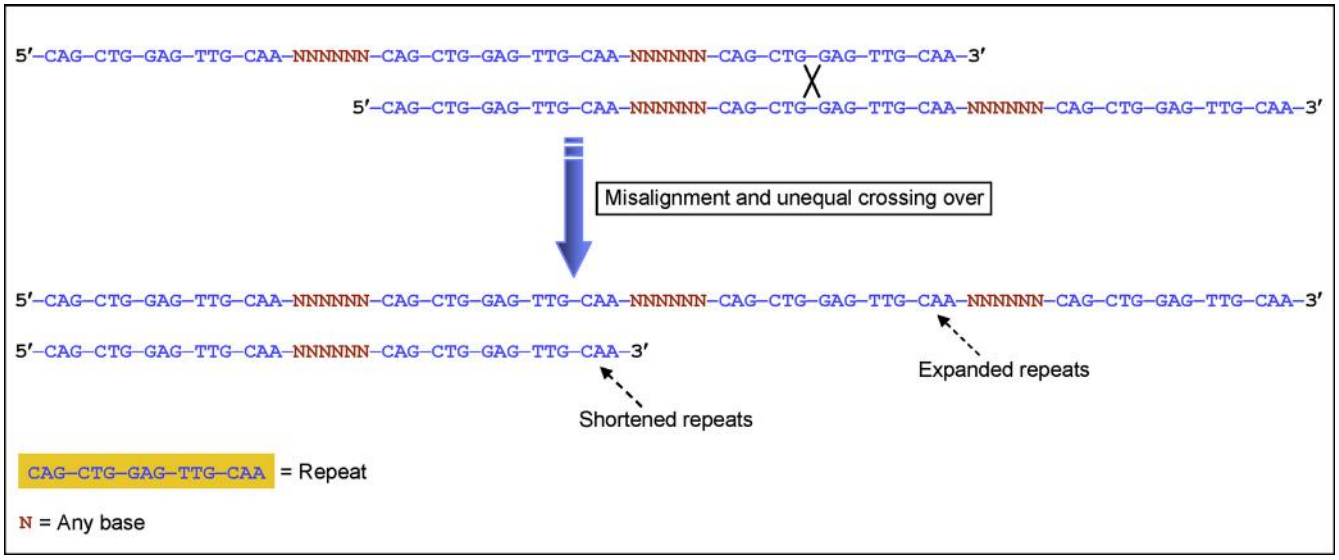

FIGURE 2.3 **Unequal crossing over altering the repeat length.** The block of repeat sequence used here as an example is –CAG–CTG–GAG–TTG–CAA–. The presence of blocks of the same repeat sequence makes the chromosomal misalignment and unequal crossing over possible.

2.3.2 Recombination and Generation of Genetic Diversity

In sexually reproducing organisms, meiotic recombination during gamete formation provides a means of creating genetic variation. In genetic recombination, a DNA segment moves from one DNA molecule to another DNA molecule. Recombination can take place between two homologous sequences or two nonhomologous sequences. Recombination between two homologous sequences is called **homologous recombination** and it occurs during meiosis between two homologous DNA molecules (homologous chromosomes) by crossing over. *The frequency of homologous recombination is low*. Recombination between two nonhomologous sequences can be mediated by **site-specific recombination**. Site-specific recombination occurs when two nonhomologous DNA molecules have only a small region of sequence identity; recombination occurs using this small region. *Recombination apparently depends on short stretches (could be as short as ~30 bp) of complete identity rather than long stretches of general similarity*.[12] Site-specific recombination helps in the integration of phage DNA into a bacterial chromosome; it can also help integrate transposable elements into the host DNA. Therefore, site-specific recombination provides a mechanism for introducing genetic diversity in the recipient genome.

Recombination between homologous chromosomes begins with double-strand breaks (DSBs). Because the non-sister chromatids of homologous chromosomes may not be identical in terms of their DNA sequence, mismatch repair synthesis during recombination may result in **gene conversion**. The mismatch repair enzyme corrects the sequence mismatch by partial resection of the broken DNA molecule followed by resynthesis of one of the strands using the corresponding DNA strand of the non-sister chromatid as the template. This results in a unidirectional transfer of the donor sequence to the acceptor sequence. It is easy to contemplate that if an allele is removed during resection, that allele is created during resynthesis based on the sequence of the allele of the donor strand. This phenomenon leads to gene conversion. Therefore, gene conversion involves nonreciprocal exchange of genetic material in which one sequence remains unchanged and the other sequence is altered.

Homologous recombination can also take place between two stretches of DNA that are not allelic. This is called **non-allelic homologous recombination** (**NAHR**). NAHR is driven by sequence identity, and it results in deletion in one chromosome and duplication in the other chromosome. Duplicated segments are predisposed to further NAHR. NAHR may lead to loss or increased copy number of specific genes, resulting in copy number variations (CNVs) of specific genes within the deleted or duplicated region. Such CNVs have major implications in health and disease as well as genome evolution. *In general, repeats provide hotspots of major structural alterations in the genome, ranging from microduplication and microdeletion to major segmental duplication and deletion, as well as repeat expansion and contraction*.

2.3.3 Gene Flow and Introduction of Genetic Diversity

Gene flow is also called **gene migration**. Gene flow is the transfer of genetic material from one population to another. Gene flow can take place between two populations of the same species through migration, and is mediated by reproduction and **vertical gene transfer** from parent to offspring. Alternatively, gene flow can take place between two different species through **horizontal gene transfer** (HGT, also known as **lateral gene transfer**), such as gene transfer from bacteria or viruses to a higher organism, or gene transfer from an endosymbiont to the host. HGT is discussed in detail later in this chapter. Gene flow within a population can increase the genetic variation of the population, whereas gene flow between genetically distant populations can reduce the genetic difference between the populations. Because gene flow can be facilitated by physical proximity of the populations, gene flow can be restricted by physical barriers separating the populations. Incompatible reproductive behaviors between the individuals of the populations also prevent gene flow.

2.3.4 Origin of New Genes, Creation of Genetic Diversity and Genome Evolution

Generation of new genes is an important mechanism for creating genetic novelties; hence, it is an important driving force of evolution in all organisms. New genes can be created by two major processes, (1) processes that use coding sequences (pre-existing genes) as the raw materials, and (2) processes that use noncoding sequences as the raw material.

2.3.4.1 Origin of New Genes from Coding Sequences (Pre-existing Genes)

These processes are better understood and include gene duplication, exon shuffling, gene fusion and fission, and lateral gene transfer.

2.3.4.1.A GENE DUPLICATION AND THE 2R HYPOTHESIS

Gene duplication creates paralogs. Susumu Ohno's seminal book *Evolution by Gene Duplication* (1970)[13] popularized the concept that gene duplication plays an important role in evolution. By comparing the genome size of different groups of non-vertebrate chordates and vertebrates, Ohno argued that the complexity of vertebrate genomes during evolution was achieved by whole-genome duplications in the lineage leading to vertebrates. Analysis of **orthologous genes** (**orthologs**[g]) showed that compared to urochordates (e.g. sea squirts), the genomes of jawless vertebrates, such as lamprey and hagfish, contain at least two orthologs and the genomes of mammals contain three or more orthologs. Ohno proposed that the ancestors of reptiles, birds, and mammals had experienced at least one tetraploid evolution either at the stage of fish or at the stage of amphibians. Since the turn of the millennium, the modern version of Ohno's hypothesis, known as the **two rounds (2R) hypothesis**, has resurfaced and gained popularity. There are disagreements regarding the stages of evolution when genome duplications took place. The most popular version of the 2R hypothesis proposes that one round of genome duplication took place at the root of the vertebrate lineage—that is, after the emergence of urochordates—followed by another around the time Agnatha (jawless vertebrates, e.g. lamprey and hagfish) and Gnathostomata (jawed vertebrates) split—that is, before the radiation of jawed vertebrates.[14–16] There are, however, debates about the 2R hypothesis, but that is beyond the scope of this section.

Ohno considered whole-genome duplication to be more important as an evolutionary mechanism than individual gene duplication, but gene duplication is now known to be a major mechanism for the creation of novel genetic material and an important driver of genome evolution. Genome sequencing shows that gene duplication is prevalent in all three domains of life (Bacteria, Archaea, Eukarya). In multicellular eukaryotes, including humans, ~40–60% genes have been produced through duplication, depending on the species. Several publications have reported on the rate of gene duplication in various eukaryotic species, but the results vary significantly. For example, based on observations from the genomic databases for several eukaryotic species, Lynch and Conery estimated that in eukaryotes the average rate of gene duplication is approximately 0.01 per gene per million years (i.e. the probability of duplication of a eukaryotic gene is at least 1% per million years[h,i]).[17,18] However, Cotton and Page estimated a gene duplication rate that is one order of magnitude lower than the estimate of Lynch and Conery.[19] Many duplicated genes are inactivated

[g]Orthologous genes or orthologs are homologs in different species—that is, they evolved from a common ancestral gene through speciation. Orthologs often retain the same or similar function(s).

[h]The duplication event per gene per million years was estimated to be 0.0023 for *Drosophila melanogaster*, 0.0083 for *Saccharomyces cerevisiae*, and 0.0208 for *Caenorhabditis elegans*, the average being ~0.01. So, it was the highest for *C. elegans*.

[i]The duplication event per gene per million years was estimated to be 0.009 for humans. In this publication, the rates calculated were slightly lower for *Drosophila*, yeast, and *C. elegans*, but the average was still ~0.01.

by accumulating degenerative mutations and become pseudogenes. Gene duplication can result from unequal crossing over, retrotransposon insertion, segmental duplication, and chromosomal (whole-genome) duplication.

If the rate of gene duplication is assumed to be somewhere in between the two estimates cited above, then it becomes close to the rate of fixed nucleotide substitutions, particularly in protein-coding genes. Using data from human and rodents, and assuming 80 million years as the time of divergence between the two lineages, the average fixed nucleotide substitution rate in protein-coding genes was calculated to be 0.74 per nonsynonymous site and 3.51 per synonymous site per billion (10^9) years.[20] However, such average estimates could still vary significantly in different species.

Unequal crossing over usually generates tandem duplication, which could involve the entire gene or part of a gene. Figure 2.3 shows duplication of a section of the gene through unequal crossing over. Duplication of the entire gene involves duplication of the introns as well as the regulatory sequences. The insertion of **processed** (**retrotransposed**) **pseudogenes** can also introduce genetic variability to the genome, particularly if the retrotransposed pseudogenes recruit new promoters and become functional. Some expressed pseudogenes regulate the mRNA expression of the normal gene. For example, *Makorin1-p1* in mice is a transcribed pseudogene, which regulates the expression of the normal gene *Makorin1*.[21] Pseudogenes are of two main types: (I) **duplicated** (**nonprocessed**) and (II) **retrotransposed** (**processed**). Duplicated pseudogenes arise from genomic DNA duplication or unequal crossing over. They retain the original exon−intron organization of the functional gene (hence nonprocessed), but their protein-coding potential is lost because of the loss of transcription regulatory elements, such as promoters or enhancers, or mutations disrupting the ORF, such as frameshifts or premature stop codons. In contrast, processed pseudogenes result from retrotransposition—that is, they arise from reverse transcription of mRNA into complementary DNA (cDNA) followed by the integration of the cDNA into the genome. As a result, processed pseudogenes lack introns and promoter, and they typically contain the poly(A) tail. Because they are retrotransposed, they are flanked by direct repeats. Processed pseudogenes are usually nonfunctional unless they are integrated under the influence of an active promoter, or recruit new promoters over time to become functional. Another type of pseudogene is known as the **unitary pseudogene**. A unitary pseudogene is a regular gene that has lost the protein-coding potential because of spontaneous mutation in the coding region; so it is neither duplicated nor retrotransposed. Because most pseudogenes are nonfunctional, they are not under selection pressure and are free to accumulate further mutations and increasingly diverge from the parent sequence from which they were derived. Pseudogenes have been identified in all known genomes, but their numbers greatly vary. For example, the estimated number of pseudogenes is 10,000−20,000 in humans, but only 110 in *Drosophila*.[22]

Human genome sequencing has revealed the widespread occurrence of **segmental duplications**, which often involve blocks of 1−200-kb (or longer) sequences that have been copied from one region of the genome and integrated into another region. Hence, segmental duplications create paralogous loci. The duplicated regions represent low-copy repeats and have >90% identity. Such strong sequence identity suggests that they are relatively recent in origin. The finished sequence of the human genome reported about 5.3% of the genome as segmental duplications.

Chromosomal (**whole-genome**) **duplication** is thought to arise by the breakdown of the normal mitotic or meiotic process. If chromosomes duplicate but do not separate (chromosomal non-disjunction) and are maintained in the same cell, a diploid gamete is produced. Fertilization of a diploid gamete by a normal haploid gamete would produce a triploid organism. The same mechanism can produce tetraploidy and even higher ploidy. In addition to the above mechanism of polyploidy, termed **autopolyploidy**, genome duplication and polyploidy can also be produced by hybridization of two related species that produce viable offspring. Such polyploidy is called **allopolyploidy**, and allopolyploids produce a diverse set of gametes. During evolution, whole-genome duplication resulting in polyploidy occurred frequently in plants but infrequently in animals.

The **evolutionary fate of duplicated genes** involves either acquiring new function or becoming nonfunctional. In most cases, the duplicated genes are free to acquire degenerative mutations and become pseudogenes (**pseudogenization**) because there are no functional constraints and the genes are not under selection pressure. Thus, pseudogenization is a neutral process. In order for the gene to escape pseudogenization and functional death, selection pressure must force the duplicated gene to drift towards fixation through **neofunctionalization**. Gene duplication followed by neofunctionalization of the duplicated gene provides an important mechanism for the genome to diverge both structurally and functionally. Neofunctionalization involves acquiring new function by the duplicated gene at the expense of the ancestral function—that is, the duplicated gene acquires a function that was not present in the ancestral gene. For example, the type III antifreeze protein (*AFPIII*) gene in the Antarctic zoarcid fish evolved from a sialic acid synthase (*SAS*) gene after duplication,

divergence, and neofunctionalization. The SAS is an old cytoplasmic enzyme present in microbes through vertebrates, whereas AFPIIIs are secreted plasma proteins that bind to invading ice crystals and arrest ice growth to prevent fish from freezing. The *SAS* gene possesses both sialic acid synthase and rudimentary ice-binding activities. Following duplication, the N-terminal SAS domain was deleted and replaced by a nascent signal peptide needed for the extracellular export of the mature protein. Further optimization of the C-terminal domain's ice-binding ability through amino acid changes led to the evolution of AFPIII as a neofunctionalized secreted protein capable of non-colligative freezing-point depression.[23] Another example is the retinoic acid receptor (*RAR*) gene. Mammals have three *RAR* paralogs—*RAR*α, β, and γ—created by genome duplications at the time of origin of vertebrates. Using pharmacological ligands selective for specific paralogs, it was demonstrated that RARβ kept the ancestral RAR role, whereas RARα and RARγ diverged both in ligand-binding capacity and in expression patterns. Therefore, neofunctionalization occurred at both the expression and the functional levels to shape RAR roles during development in vertebrates.[24] Many other examples of neofunctionalization have been reported in the literature.

Neofunctionalization does not always have to arise following gene duplication. A beneficial mutation of the wild-type gene may create a mutant allele with new function. If the beneficial mutant allele is maintained by balancing selection, the carrier (heterozygote) will have increased fitness. If the beneficial mutant allele becomes the source of the duplicated gene, then the duplicated gene will be quickly fixed in the population by positive selection.[25]

Another functional outcome of gene duplication and divergence is **subfunctionalization**. Like pseudogenization, subfunctionalization is also a neutral process. Subfunctionalization occurs when the duplicated copies (paralogs) partition the attributes of the ancestral gene, such as function and/or expression. Following a duplication event, both paralogs experience a period of relaxed selection and accelerated evolution. This is because natural selection does not distinguish which paralog should be under selection and which paralog should be free from selective constraint. Thus, both genes might accumulate mutations that impair ancestral gene function. Under this condition, each paralog may retain one part of the function (subfunction) of the ancestral gene. Alternatively, each individual paralog may lose its ability to substitute for the ancestral gene function, but together the two paralogs may still be able to complement each other in producing ancestral gene function. Subfunctionalization has been proposed as an alternative mechanism driving duplicate gene retention in organisms with small effective population sizes.[26] A model to explain the high retention of duplicated genes through subfunctionalization was provided early on by the **duplication−degeneration−complementation (DDC) model**.[27] According to the DDC model, originally proposed in the context of *cis*-regulatory elements, subfunctionalization is driven entirely by degenerative mutations. Degenerative changes occur in regulatory sequences of both duplicated copies such that the expression pattern of the original gene can only be achieved when the two duplicated genes can complement each other. Therefore, degenerative mutations in the regulatory elements may increase the chance of duplicate gene retention. An implication of the DDC model is that the paralogs can not accumulate same inactivating mutations that would interfere with their ability of complementation. A number of examples of subfunctionalization have been reported in the literature. A common example is the normal human hemoglobin, which is composed of two α-chains and two β-chains (α2β2) encoded by α-globin and β-globin genes, respectively. The α- and β-globin genes are products of gene duplication and subsequent subfunctionalization because they complement each other in producing normal functional hemoglobin.[28] An example of subfunctionalization in terms of differential expression of paralogs is that of the *pax6a* and *pax6b* genes in zebrafish; these paralogs arose following a whole-genome duplication event about 350 million years ago. The expression patterns of *pax6a* and *pax6b* have diverged from each other since the duplication event. Whereas *pax6a* is widely expressed in the brain compared to *pax6b*, only *pax6b* is expressed in the developing pancreas. Such differential expression of *pax6b* in brain and pancreas is due to the loss of a brain-specific downstream regulatory element but gain of an upstream pancreas enhancer element.[29] An example of subfunctionalization has also been reported in Archaea. When Tocchini-Valentini and coworkers searched the genome of *Sulfolobus solfataricus* (Archaea; Crenarchaeota) for **homologs**[j] of *Methanocaldococcus jannaschii* (Archaea; Euryarchaeota) tRNA endonuclease, they found two paralogs of the tRNA endonuclease gene of *M. jannaschii* in the genome of the *S. solfataricus*. Characterization of these two paralogous gene products revealed that both are required for tRNA endonuclease activity, each complementing the other for complete

[j]Homologous genes, or homologs, are related to each other by descent from a common ancestral gene. Homologs may or may not have the same or similar function. Therefore, the orthologs and paralogs described above are two different types of homologous genes.

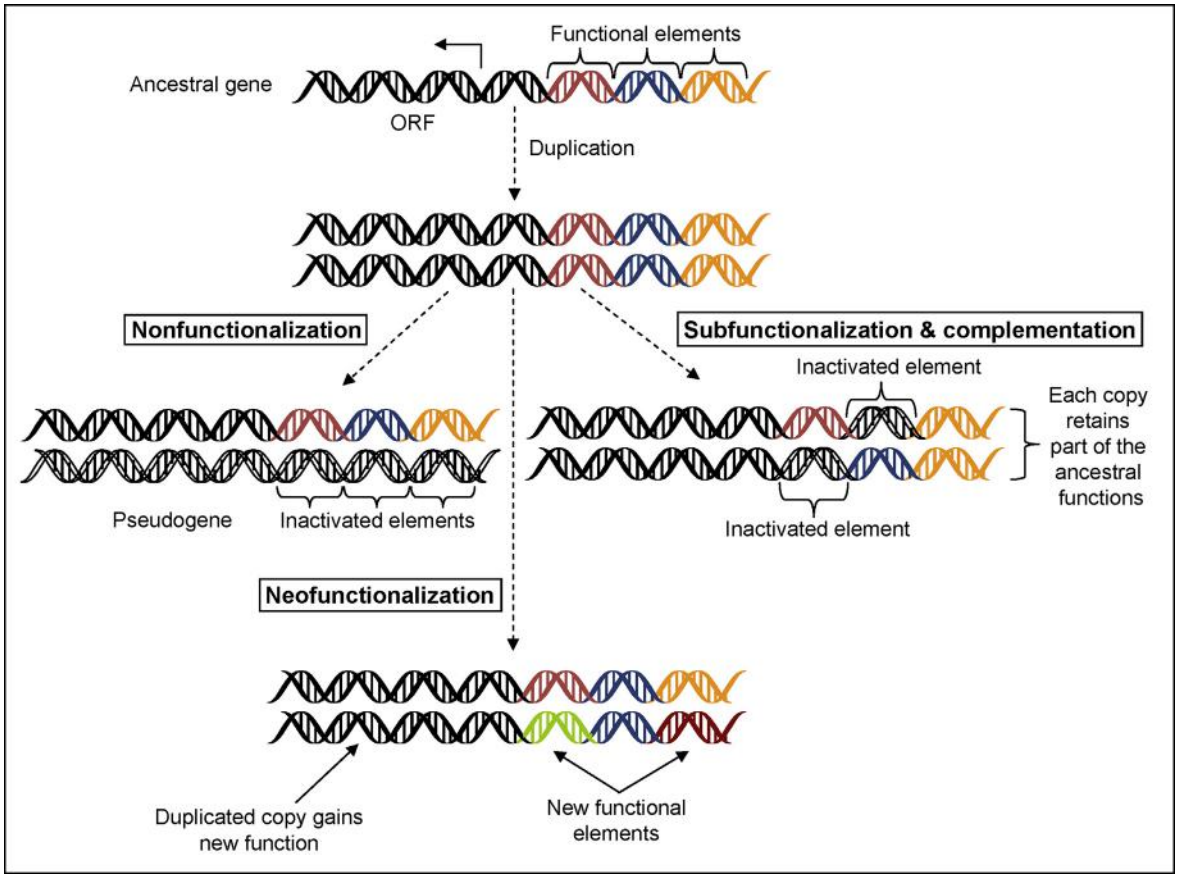

FIGURE 2.4 **Three possible fates of duplicated genes: pseudogenization (nonfunctionalization), neofunctionalization, and subfunctionalization using *cis*-regulatory modules as targets of divergence.** Duplicated genes are not under selection pressure; hence, there are no functional constraints and a duplicated gene is free to acquire degenerative mutations and become a pseudogene. Sometimes, the acquisition of new function by the duplicated gene (neofunctionalization) provides an important mechanism for the genome to diverge both structurally and functionally. The newly acquired function is not present in the ancestral gene. Subfunctionalization occurs when the duplicated copies (paralogs) partition the attributes of the ancestral gene, such as function and/or expression. The figure shows that degenerative changes occurred in regulatory sequences of both paralogs such that the expression pattern of the original gene can only be achieved when the two duplicated genes complement each other (see text for examples).

activity. Detailed analysis of the amino acid sequences of the two proteins demonstrated that these two sequences had evolved by duplication of the ancestral sequence followed by divergence and subfunctionalization of the sequences.[30] Figure 2.4 shows the three fates of duplicated genes discussed here (pseudogenization, neofunctionalization, subfunctionalization) using *cis*-regulatory modules as targets of divergence.

2.3.4.1.B EXON SHUFFLING

The natural process of creating new combinations of exons by intronic recombination is called exon shuffling.[31] Following the discovery of introns, Walter Gilbert suggested that the presence of introns allowed exon shuffling, which resulted in genomes being more complex and diversified. Exon shuffling is largely responsible for protein-domain shuffling.[32] The diversity of protein-domain combinations increased with the evolution of organismal complexity. However, most protein domains are ancestral; only few new domains have been invented in the vertebrate lineage. For example, about 7% of the protein families in human genome seem to be specific to vertebrates. The majority of the proteins necessary for the maintenance of basic cellular functions evolved early. Hence, the evolution of proteome complexity was driven by the reshuffling of pre-existing components into a richer collection of domain architectures.[33] Therefore, protein-domain shuffling, which refers to the duplication of a domain or the insertion of a domain from one gene into another, has been a major factor in the evolution of human phenotypic complexity. Kaessmann et al.[34] systematically analyzed intron phase distributions in the coding sequence of human protein domains to identify signatures of exon shuffling resulting in domain shuffling. Introns of symmetrical phase combinations

(i.e. 0−0, 1−1, and 2−2[k]) were found to be predominant at the boundaries of domains, whereas non-boundary introns showed no excess symmetry, suggesting that exon shuffling primarily involved rearrangement of structural and functional domains. Domains flanked by phase 1 introns (i.e. 1−1 symmetrical domains) were found to have dramatically expanded in the human genome due to domain shuffling. The observation of predominance and extracellular location of 1−1 symmetrical domains among metazoan protein-specific domains suggested an association with the evolution of multicellularity. In contrast, 0−0 symmetrical domains were found mostly overrepresented among ancient protein domains that are shared between the eukaryotic and prokaryotic kingdoms. Franca et al.[35] investigated the intron phase distribution in 10 genomes to generate a catalog of putative exon shuffling events in several eukaryotic species, including non-metazoans (choanoflagellate *Monosiga brevicollis*), early branching metazoans (the sea anemone *Nematostella vectensis*), the smallest chordate (urochordate *Ciona intestinalis*), and representative species from all vertebrate lineages except reptiles (zebrafish, *Xenopus*, chicken, mouse, and human). They confirmed previous observations that exon shuffling mediated by phase 1 introns (1−1 exon shuffling) is the predominant kind in multicellular animals, whereas exon shuffling mediated by phase 0 introns (0−0 exon shuffling) is the predominant type in non-metazoan species. They also concluded that such a pattern was achieved since the early steps of animal evolution.

Intronic recombination generating exon shuffling was most likely facilitated by two important events at a later stage during the evolution of eukaryotes: the emergence of spliceosomal introns, and the insertion of repetitive sequences within spliceosomal introns.[36] Although the presence of repetitive sequences in introns could facilitate intron recombination, insertion of repetitive sequences in self-splicing introns would not have been tolerated because self-splicing introns encode an essential function. In contrast, insertion of repetitive sequences would have been tolerated in spliceosomal introns because of the lack of such functional constraints. Hence, recombination involving self-splicing introns early in life's evolution could not have played an important role in exon shuffling, and consequently in the evolution of ancient proteins. Exon shuffling most likely increased in parallel with the evolution and expansion of spliceosomal introns and the concomitant appearance of less compact genomes.

Patthy analyzed the evolutionary distribution of some proteins that could be identified as modular proteins (containing specific functional modules) and seemingly evolved by intronic recombination. His analysis revealed that modular multidomain proteins produced by exon shuffling are restricted in their evolutionary distribution[l]. The majority of these proteins are functionally linked to the evolution of multicellularity of animals, such as constituents of the extracellular matrix, proteases involved in tissue remodeling, various proteins of body fluids, and proteins associated with cell−cell and cell−matrix interactions. Some examples include selectins, interleukin-2 receptor, cartilage link protein, follistatin, C-type lectin, and tolloid. The results suggest that exon shuffling acquired major significance at the time of metazoan radiation.

2.3.4.1.C GENE FUSION AND FISSION

During evolution, many complex proteins were apparently produced by gene fusion and less complex proteins by gene fission. Gene fusion results in the creation of a composite protein. In contrast, gene fission results in the creation of two or more smaller, split proteins. For example, the basic biochemistry of fatty acid synthesis is very similar from *E. coli* to mammals. However, the six enzymes and the acyl carrier protein involved in fatty acid synthesis exist as independent polypeptides in *E. coli*, whereas in mammals these exist as one composite polypeptide containing all the activities because of the fusion of genes encoding them.

Snel and coworkers[37] analyzed all ORFs of 17 completely sequenced bacterial genomes using the Smith−Waterman sequence comparison algorithm; the analysis showed evidence for numerous cases of gene fusion and fission. In general, they observed that

[k]As mentioned in Chapter 1, introns can be divided into three types based on phases: phase 0, phase 1, and phase 2. A phase 0 intron does not disrupt a codon, a phase 1 intron disrupts a codon between the first and the second bases, and a phase 2 intron disrupts a codon between the second and third bases. An exon flanked by two introns of the same phase (e.g. 0−0, 1−1, 2−2) is called a symmetrical exon, whereas an exon flanked by two introns of different phases (e.g. 0−1, 1−2, 2−0, etc.) is called an asymmetrical exon. Legitimate alternative splicing involves the removal of a symmetrical exon. In contrast, alternative splicing involving an asymmetrical exon results in a change of the ORF downstream of the 3′-splice site (Figure 1.5), but this is very rare.

[l]In the analysis, protein modules were considered to be generated through exon shuffling if: (1) the modules were homologous (i.e. modules derived from a common ancestor) but present in otherwise nonhomologous proteins, and (2) the transposition of the module was mediated by exon shuffling through intronic recombination. Evidence of exon shuffling through intronic recombination was considered if the module was flanked by introns of same phase. Thus, the introns of these modular proteins were shown to have a marked intron-phase bias.

fusion occurred more often than fission. Using the same approach (sequence-based comparison) Enright and Ouzounis[38] identified 7224 components and 2365 composite unique proteins across the 24 species considered in the study. These 24 genomes included those of bacteria and eukaryotes, including *Drosophila melanogaster* and *Caenorhabditis elegans*. They found a number of functional associations. For example, MXR1 (peptide methionine sulfoxide reductase, involved in antioxidative processes) and YCL033C (function unknown) were predicted to be functionally associated by virtue of gene fusion in three species—*Helicobacter pylori*, *Haemophilus influenzae*, and *Treponema pallidum*—and this observation was supported by experimental results. Likewise, Yanai et al.[39] identified groups of closely related proteins that have undergone fusion or fission. For example, the genes for glycolytic enzymes triosephosphate isomerase (TPIA), phosphoglycerate kinase (PGK), and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) in the parasitic bacterium *Mycoplasma genitalium*, are linked by fusion events in other species, such as TPIA + PGK in *Thermotoga maritima* and TPIA + GAPDH in *Phytophthora infestans*.

Using domain architecture comparison, Kummerfeld et al.[40] performed a comprehensive analysis of divergent sequences in distantly related organisms to identify evidence of gene fusion and fission during evolution. The authors considered proteins at the level of domain architecture because structural domains reveal more about distant evolutionary relationships than simple sequence alignment. The domain information was collected from the Structural Classification of Proteins (SCOP) database, which provides an evolutionary definition of domains based on three-dimensional structure. The authors studied proteins across 131 genomes (17 Archaea, 98 Bacteria, and 16 Eukarya), and investigated 7116 domain architectures to identify protein domains that evolved by fusion or fission. In order to do that, the authors looked for domain architectures that were present as a single protein (i.e. the composite form) in at least one genome, and as a set of shorter proteins (i.e. the split forms) in other genomes, which would suggest that the composite protein was split by fission or the split proteins were fused at some stage during evolution. The authors identified 2869 groups of multidomain proteins as a single protein in certain organisms and as two or more smaller proteins with equivalent domain architectures in other organisms. They also found that fusion events were approximately four times more common than fission events, which is consistent with the observation by Snel et al. The authors discussed the possible contribution of horizontal gene transfer in the evolution of composite proteins, which is more prevalent in Bacteria and Archaea.

2.3.4.1.D HORIZONTAL GENE TRANSFER

Horizontal gene transfer, also known as **lateral gene transfer**, refers to nonsexual transmission of genetic material between unrelated genomes; hence, horizontal gene transfer involves gene transfer across species boundaries. The phenomenon of horizontal gene transfer throws a wrench in the concepts of last common ancestor, syntenic relationship between genomes, phylogeny and the evolution of discrete species units, taxonomic nomenclature, etc.[m] The majority of examples of horizontal gene transfer are known in prokaryotes. In bacteria, three principal mechanisms can mediate horizontal gene transfer: **transformation** (uptake of free DNA), **conjugation** (plasmid-mediated transfer), and **transduction** (phage-mediated transfer). In plants, **introgression** can mediate horizontal gene transfer; this means gene flow from one gene pool to another gene pool—that is, from one species to another species by repeated backcrossing between an interspecific hybrid and one of its parent species. Therefore, introgression depends on the extent of reproductive isolation between the two species. Introgression has also been reported between duck species, between butterfly species involved in mimicry, and between human and Neanderthal.[41]

Horizontal gene transfer in animals is not common, but there are some reports. For example, Acuña et al.[42] identified the gene *HhMAN1* from the coffee berry borer beetle, *Hypothenemus hampei*, which shows clear evidence of horizontal gene transfer from bacteria. *HhMAN1* encodes the enzyme mannanase, which hydrolyzes galactomannan. Phylogenetic analyses of the mannanase from both prokaryotes and eukaryotes revealed that mannanases from plants, fungi, and animals formed a distinct eukaryotic clade, but *HhMAN1* was most closely related to prokaryotic mannanases, grouping with the *Bacillus* clade. *HhMAN1* was not detected in the closely related species *H. obscurus*, which does not colonize coffee beans. The authors hypothesized that the acquisition of the *HhMAN1* gene from bacteria was likely an adaptation in response to need in a specific ecological niche.

[m]During evolution, different lineages split from a common ancestor (the last common ancestor of those lineages) and evolve to ultimately form reproductively isolated groups (species). However, lineages descending from a common ancestor still maintain many ancestral genes in groups and in the same order but scattered in different chromosomes (syntenic relationship between genomes). This scenario of evolution does not consider the possibility of exchange of genetic material between groups belonging to different lineages. The phenomenon of horizontal gene transfer is an exception to this paradigm.

There are also some examples of horizontal gene transfer from fungi to arthropods, such as aphids (insects) and mites (arachnids). Phylogenetic analysis revealed the evidence of horizontal transfer of genes encoding carotenoid desaturase and carotenoid cyclase–carotenoid synthase from fungi to pea aphid,[43] and to spider mite.[44] Notably, the fused carotenoid cyclase–carotenoid synthase gene is characteristic of fungi but not of plants or bacteria. The authors discussed the possible mechanism of such gene transfer. Gene transfer into a single arthropod ancestor of both spider mites and aphids is not likely because it would require subsequent loss of these genes in most other living arthropod taxa. The most likely scenario is the transfer of these genes through symbiosis, which probably occurred independently in both aphids and spider mites. It has been suggested that the frequent association of mites with viruses makes them ideal horizontal gene transfer vectors, including incorporation of mobile genes into their own genomes.

2.3.4.2 *Origin (de Novo) of New Genes from Noncoding Sequences*

The processes of how a new gene is created de novo from noncoding sequence are not well understood. For a noncoding DNA to give birth to a protein-coding gene, two features are needed: the DNA must be transcription-competent, and the DNA must acquire an open reading frame. It is being increasingly appreciated that a rare but consistent feature of eukaryotic genomes is the evolution of new genes de novo. Every genome contains genes that lack homologs in other taxonomic lineages. These new genes are called **orphan genes**. Orphan genes may arise by duplication and rearrangement followed by rapid divergence, but their de novo origin from noncoding DNA appears to be a very important mechanism.[45] If orphan genes are born through a duplication–divergence mechanism, they have to diverge beyond recognition as paralogs. In contrast, the de novo origin of orphan genes from noncoding DNA requires the emergence of sequence features forming functional signals, such as transcription initiation signal, polyadenylation signal, splice signal, etc., and finally the sequence would have to come under regulatory control in order for the gene to be expressed. Further accumulation of additional regulatory elements can expand the tissue expression pattern of a newly evolved orphan gene. *One characteristic of genes originated de novo is that these genes are usually simple (mostly single exon) so that their evolution de novo would be possible.*

In recent years, following the sequencing of many genomes, there have been multiple reports of identification of genes born de novo from noncoding DNA. Begun and coworkers,[46,47] reported de novo origin of orphan genes from noncoding DNA in *Drosophila*. By comparing the genome sequences of various species of *Drosophila*, Levine et al. described five novel genes in *D. melanogaster* that were derived from noncoding DNA. These genes have no homologs in any other species. Begun et al. subsequently used testis-derived expressed sequence tags (ESTs) from *D. yakuba* to identify genes that have likely arisen either in *D. yakuba* or in the *D. yakuba*/*D. erecta* ancestor. They identified eleven such genes. The genes described in these two publications are mostly X-linked, expressed in the testis, and have male germ-line functions. Zhou et al.[48] identified nine genes that originated de novo, and estimated that about 12% of the new genes that originated in the *Drosophila* lineage had arisen de novo. In recent years, efforts have turned to the human genome in order to find genes that most likely originated de novo. By building blocks of conserved synteny between human and chimpanzee genome and using 1:1 orthologs identified as BLASTP hits (hits in the protein database using Basic Local Alignment Search Tool (BLAST)) with no other similarly strong hits, Knowles and McLysagh reported three human protein-coding genes—*CLLU1*, *C22orf45*, and *DNAH10OS*—that seemingly had de novo origin in the human genome. Each of these three genes is a single-exon gene; however, they do contain introns in the untranslated regions. In order to minimize the chance that the genes could be annotation artifact, the authors only considered human genes that are classified as "known" by Ensembl and that have expressed sequence tag (EST) support for transcription.[49] Another de novo protein-coding gene, *C20orf203*, which is associated with brain function in humans, was reported in 2010.[50]

More recently, the identification of the most extensive set of human genes born de novo from noncoding DNA was reported by Wu et al.[51] Using a similar approach as that of Knowles and McLysaght, they reported 60 new protein-coding genes that apparently originated de novo in the human lineage since its divergence from the chimpanzee. Their data are supported by both transcriptional and proteomic evidence. Using RNA sequencing, the highest expressions of these genes were found to be in the cerebral cortex and testes, suggesting that these genes may contribute to phenotypic traits that are unique to humans, including the development of cognitive ability. Interestingly, the earlier finding of Knowles and McLysagh on the three human genes identified as having a de novo origin (*CLLU1*, *C22orf45*, and *DNAH10OS*) was not supported by the findings of Wu et al. The discrepancy was due to changes in gene annotation in the different versions of the databases used by these two groups (version 46 used by Knowles and McLysaght versus version 56 used by Wu et al.). This discrepancy also underscores the fundamental challenge of identifying

genes of de novo origin accurately based on annotated genome. A major challenge remains to demonstrate the functionality of these genes.

Exonization of previous intron sequences through mutation and abolition of splice sites is another mechanism of increasing the proportion of coding sequences derived from noncoding sequences in the genome. Examples include exonization of intronic *Alu* sequences,[52,53] and of intronic sequences in the collagen IV gene.[54] However, exonization of introns may also be associated with pathological outcomes.[55,56]

2.4 FACTORS THAT AFFECT GENE FREQUENCY IN A POPULATION

The mechanism of molecular evolution also involves the accumulation of genetic diversity, which leads to changes in gene frequency and genetic structure of the population. Changes in allele frequency initially result in microevolution, which introduces genetic variations in a population through processes such as mutation, migration, selection, genetic drift, population bottlenecks, and even relaxation of purifying selection.

A simple model for calculating gene frequency in a diploid population is provided by the **Hardy–Weinberg equilibrium** principle (see Box 2.1). It states that *the gene frequency in a diploid population remains constant through generations provided five conditions are met: no mutation, no migration, no selection, no genetic drift, and panmixis (random mating)*. For example, two alleles A_1 and A_2 can produce three possible genotypes: A_1A_1, A_1A_2, and A_2A_2. According to the Hardy–Weinberg principle, if the frequency of A_1 is p, and the frequency of A_2 is q ($q = 1 - p$, because $p + q = 1$, i.e. 100%), then the frequencies of A_1A_1, A_1A_2, and A_2A_2 are p^2, $2pq$, and q^2, respectively, and $p^2 + 2pq + q^2$ will also be 1 (i.e. 100%). A population in which the genotypic ratios are maintained is said to be in Hardy–Weinberg equilibrium.

BOX 2.1

Hardy–Weinberg Equilibrium at a Single Locus with Two Alleles

		Sperm	
		A_1 (p)	A_2 (q)
Egg	A_1 (p)	A_1A_1 (p^2)	A_1A_2 (pq)
	A_2 (q)	A_1A_2 (pq)	A_2A_2 (q^2)

Hence, the frequencies are: $A_1A_1 = p^2$, $A_1A_2 = 2pq$, $A_2A_2 = q^2$.

The sum of the frequencies of alleles as well as the genotypes is always 1.

Hence, for the alleles, $p + q = 1$ (=100%), and for the genotype, $(p + q)^2 = 1$, or $p^2 + 2pq + q^2 = 1$ (=100%).

Example: If the frequency of $A_1 = 0.7$ and the frequency of $A_2 = 0.3$ (=1 − 0.7), then the frequencies of the genotypes in the population are as follows:

$A_1A_1 = (0.7)^2 = 0.49 = 49\%$;

$A_1A_2 = 2(0.7)(0.3) = 0.42 = 42\%$;

$A_2A_2 = (0.3)^2 = 0.09 = 9\%$.

Hardy–Weinberg Equilibrium at a Single Locus with Three or More Alleles (Multiple Alleles)

If the locus under study has three or more alleles (multiple alleles), the derivation of frequencies is similar to that used for two alleles. If the alleles are A_1, A_2, and A_3, and the frequencies are, p, q, and r respectively, then:

The gene frequency p (A_1) + q (A_2) + r (A_3) = 1.
The genotype frequency $(p + q + r)^2 = 1$, or
p^2 (A_1A_1) + q^2 (A_2A_2) + r^2 (A_3A_3) + $2pq$ (A_1A_2) + $2pr$ (A_1A_3) + $2qr$ (A_2A_3) = 1.

Hardy–Weinberg Equilibrium at Two or More Loci

Let's assume, at one locus, the alleles are A_1 and A_2 and their frequencies are p and q, respectively.

At a separate, independently assorting locus, the alleles are B_1 and B_2, and their frequencies are r and s, respectively. Hence, $p + q = 1$, and $r + s = 1$.

The four types of allelic combinations in the gametes are: A_1B_1, A_1B_2, A_2B_1, and A_2B_2; their frequencies will be pr, ps, qr, and qs, respectively, and $pr + ps + qr + qs = 1$.

If all the alleles are at equilibrium, then the genotype frequencies will be $(pr + ps + qr + qs)^2$. The genotype frequencies of offspring can also be easily calculated using the Punnett square; for example, a cross $A_1A_2B_1B_2 \times A_1A_2B_1B_2$ will yield p^2r^2 $A_1A_1B_1B_1$; $2p^2rs$ $A_1A_1B_1B_2$; $2pqr^2$ $A_1A_2B_1B_1$; ... q^2s^2 $A_2A_2B_2B_2$.

The Hardy–Weinberg equilibrium principle is a very simplistic representation of the maintenance of gene frequencies in a population, and it does not take into account most of the complexities associated with actual populations. The conditions that need to be met for a population to remain in Hardy–Weinberg equilibrium also underscore the conditions that can introduce genetic variations in a population and cause microevolution, as discussed below.

2.4.1 Mutation

Genetic variation in a population is derived from a wide assortment of different alleles. Mutation or change in the genetic material is one of the primary sources of generation of genetic diversity in the population. As discussed above, a mutation can be a point mutation, a change in the open reading frame of a gene, or a chromosomal mutation. Chromosomal mutations are large-scale changes in chromosomal structure and organization, exemplified by insertion–deletion (indel), inversion, duplication, and translocation (Figure 2.1A).

The spontaneous point mutation rate (see Box 2.2) varies depending on the gene and the species. The mutation rate can be expressed differently. Studies utilizing breeding of control mice and monitoring mutations in five coat-color loci demonstrated an average mutation rate of $\sim 12 \times 10^{-6}$ per locus per gamete for forward mutations from the wild type, and $\sim 2 \times 10^{-6}$ per locus per gamete for reverse mutations from recessive alleles.[57,58] Mouse mutation data summarized from different radiation experiments showed a forward mutation rate of 6.6×10^{-6} per locus per generation.[59] The average forward mutation rate of the hypoxanthine phosphoribosyltransferase (*HPRT*) gene of the human promyelocytic leukemia cell line HL-60 was reported to be $\sim 2\text{–}6 \times 10^{-7}$/cell/generation.[60] When the mutation rate is calculated based on the evolution of pseudogenes, it turns out to be one or two orders of magnitude higher. This is expected because pseudogenes are mostly free from selective constraints. For example, the mutation rate based on the evolution of pseudogenes in humans was estimated to be $\sim 2 \times 10^{-8}$ per base per generation.[61] However, a different estimate, based on determining the substitution rate in pseudogenes, calculated the average mutation rate in mammalian nuclear DNA to be $3\text{–}5 \times 10^{-9}$ nucleotide substitutions per nucleotide site per year.[62]

Therefore, changes in allele frequency due to mutations alone are very small. Nevertheless, for a large population, the cumulative effect of mutation over many generations can be significant. Recently, it was demonstrated that natural genetic variations in the human genome are caused by small insertions and deletions.[63] The authors reported almost 2 million small insertions and deletions (indels) ranging from 1 to 10,000 bp in length in the genomes of 79 diverse humans. These variants include 819,363 small indels that map to human genes. Small indels were frequently found in the coding exons of these genes, and several lines of evidence indicate that such variations are a major determinant of human biological diversity.

BOX 2.2

ESTIMATION OF MUTATION RATE

The mutation rate in haploid organisms can be directly measured because the mutation will be expressed and the mutant phenotype can be observed.

Determination of the mutation rate in diploid organisms is more challenging because a recessive mutation can be masked by the dominant allele. Hence, the expression of the mutant phenotype and the actual occurrence of the mutation can be separated by many generations. Some major contributions on the estimation of mutation rate in mammals were made by a number of different groups from the 1950s to the 1970s. The contributions of Gunther Schlager and Margaret Dickie (cited above) of the Jackson Laboratory, Bar Harbor, Maine, are worth mentioning simply because of the volume of the work they did. They analyzed in excess of 7 million mice over many years for five coat-color loci (*nonagouti*, *brown*, *albino*, *dilute*, *leaden*) for estimating the average mutation rate.

For direct estimation, as done by Schlager and Dickie, the mutation rate in a single generation is used. In this scenario, the parental genotypes are known. If the offspring shows a mutant phenotype, it is backcrossed with the parents, and also crossed with a mouse homozygous for that mutation, and with a mouse that does not carry the mutation, in order to confirm the mutation. The mutation rate is calculated as follows:

$$\mu = x/2N,$$

where μ = mutation rate, x = number of mutant offspring, and N = total number of offspring examined. The factor 2 is used because each offspring develops from fertilization involving two haploid gametes. Each haploid gamete contains one allele that can potentially be the mutant allele. Therefore, the mutation rate calculated this way is expressed as "per locus per gamete." When using cell

BOX 2.2 (*cont'd*)

culture, the mutation rate can also be expressed "per cell division."

Example: If eight offspring are born with a mutant phenotype out of 1 million (10^6) progeny, and if three of those offspring had affected parents, then five offspring were born with the new mutation. Therefore, the mutation rate will be $5/(2 \times 10^6) = 2.5 \times 10^{-6}$ per locus per gamete.

Because an accurate estimation of mutation rate involves using animals with known genotype, many forward crosses and backcrosses with parents, and careful analysis of a large number of progeny, it may be difficult to determine the true mutation rate if parental genotype information is not available. In this situation, the **mutation frequency** (instead of mutation rate) can be calculated using the same formula. The mutation frequency does not tell when the mutation first appeared in the population; however, mutation frequency can provide an approximation of the true mutation rate.

2.4.2 Migration (Gene Flow)

Migration is the movement of organisms from one location to another. It involves movement from one subpopulation to another subpopulation, or dispersal of groups of individuals from one central population into different geographic locations. The various subpopulations of a species that has broad geographic distribution do not have the same genetic makeup; therefore, the relative frequency of various alleles may differ significantly. In such cases, migration of individuals from one subpopulation to another can add significant genetic variation to the receiving subpopulation. If the individuals from the two subpopulations then mate (panmixis), the relative frequencies of various alleles and genotypes eventually change and come to equilibrium again. In contrast, if groups of individuals move out of one central population into different geographic locations, then over time those subpopulations accumulate genetic variations independently and consequentially genetically diverge form one another.

The gene frequencies in the resulting population can be calculated by taking into account the fraction of the migrant subpopulation, the fraction of the native subpopulation, and the gene frequencies in those subpopulations, as exemplified in Box 2.3.

2.4.3 Natural Selection

Natural variations exist among the individuals in any population. Many of these differences do not affect survival or reproductive fitness (e.g. the eye color variations in humans), but some differences may improve the chances of survival of a particular group of individuals. Natural selection results in the fixation of these advantageous variations in the population, leading to greater adaptability to and reproductive success in the environment. Thus, natural selection drives the evolutionary engine.

Natural selection can be of two types, based on its effect on the fate of genetic variations: purifying (negative) selection and positive (Darwinian) selection. Purifying selection removes deleterious variations, whereas positive selection fixes beneficial variations in the population and promotes the emergence of new phenotypes. As a result, natural selection acts on populations to determine the allele frequency and distribution of quantitative traits[n] over generations. The principal types of selection determining the distribution of traits across a population are directional, stabilizing, disruptive, and balancing selection.

Directional selection favors the advantageous allele so that its proportion (and the associated phenotype) increases in the population. As a result, both the allele frequency and the phenotype are skewed in one direction and away from the average phenotype (Figure 2.5A). A popular example is the phenomenon of **industrial melanism** in the peppered moth (*Biston betularia*). This species has both light- and dark-colored phenotypes. Before the industrial revolution in England, the light-colored phenotype was predominant. During the industrial revolution, the trees on which the peppered moths

[n]A **quantitative trait** is a phenotype that is influenced by multiple genes as well as by the environment. Each gene involved in influencing a quantitative trait segregates according to Mendel's law. Because of polygenic influence, quantitative traits vary over a continuous range; hence, they are also known as **continuous traits**. As the name implies, quantitative traits can be measured. Some examples of quantitative trait phenotype in humans are skin color, height, blood pressure, and IQ. The (statistical) analysis that helps find the association between the phenotype and the molecular data in order to explain the genetic basis of complex traits is known as **quantitative trait locus** (**QTL**) analysis.

BOX 2.3

EFFECT OF MIGRATION ON GENE AND GENOTYPE FREQUENCIES

If a migrant subpopulation M migrates into a native subpopulation N, forming the resulting population R, the fraction of the migrant population in the resulting population is M/R, and that of the native population is N/R; hence, M/R + N/R = 1 (i.e. 100%).

If:

The frequency of $A_1 = p_M$ and that of $A_2 = q_M$ in subpopulation M
The frequency of $A_1 = p_N$ and that of $A_2 = q_N$ in subpopulation N
The frequency of $A_1 = p_R$ and that of $A_2 = q_R$ in the resulting population R

then:

$p_R = [(M/R \times p_M) + (N/R \times p_N)]$
$q_R = [(M/R \times q_M) + (N/R \times q_N)]$.

Example: If 300 individuals from a subpopulation (M) migrate into a native subpopulation (N) of 700 individuals, the resulting population (R) will contain 1000 individuals.

So, M/R = (300/1000) = 0.3 (i.e. 30% of the resulting population is migrant population); N/R = (700/1000) = 0.7 (i.e. 70% of the resulting population is native population).

Originally, if:

The frequency of A_1 in subpopulation M (p_M) = 0.45, and that of A_2 (q_M) = 0.55
The frequency of A_1 in subpopulation N (p_N) = 0.75, and that of A_2 (q_N) = 0.25

then:

The frequency of A_1 in the resulting population R (p_R) = $[(M/R \times p_M) + (N/R \times p_N)] = [(0.3 \times 0.45) + (0.7 \times 0.75)] = 0.66$
The frequency of A_2 in the resulting population R (q_R) = $[(M/R \times q_M) + (N/R \times q_N)] = [(0.3 \times 0.55) + (0.7 \times 0.25)] = 0.34$

Therefore, the frequencies of A_1 and A_2 in the resulting population are different from those of both the migrant and native populations.

With the change in gene frequencies, the genotype frequencies of A_1A_1, A_1A_2, and A_2A_2 in the resulting population R would change as well, and can be calculated following the Hardy–Weinberg equilibrium principle.

rested were blackened by soot. The darker background gave the dark-colored moths an advantage in hiding from predatory birds and at the same time made the light-colored moth more visible and prone to predation. As a result, over time the dark-colored moths proliferated and became the predominant phenotype while the light-colored moth population was significantly reduced. Through regulation and legislation, the environment started clearing up. As a result, the balance between light-colored and dark-colored varieties was reversed and the light-colored variety proliferated again.

Stabilizing selection is known to be the most prevalent type of natural selection; it favors the intermediate (average) phenotype of the trait, and in doing so it removes the extreme phenotypes of the trait from the population (Figure 2.5B). Thus, stabilizing selection reduces genetic variability in the population. *It is generally accepted that stabilizing selection maintains the DNA and protein sequences over evolutionary time*. However, Kimura[64] demonstrated that under stabilizing selection, extensive neutral evolution can occur through random genetic drift. In other words, many cryptic neutral genetic changes may occur in natural populations while maintaining the phenotype unchanged. A common example of stabilizing selection is the mortality and birth weight in human babies. It is well known that both very large and very small human babies suffer high mortality rates; hence, the intermediate weight is the most favored phenotype for survival.

Disruptive selection (**diversifying selection**) favors the two extreme phenotypes of the trait and minimizes the average phenotype. Thus, disruptive selection creates a bimodal distribution of a trait in the population; consequently, it is the opposite of stabilizing selection in the outcome (Figure 2.5C). Disruptive selection is an important driving force behind sympatric speciation[o]. An example of disruptive selection is provided by the mimicry and survival of the African butterfly *Pseudacraea eurytus*. In this species, the coloration

[o]**Sympatric speciation** is the process by which new species evolve from an ancestral species through the evolution of reproductive barriers while inhabiting the same geographic region. This is in contrast to **allopatric speciation**, in which geographical isolation separates two populations of a species resulting in reproductive isolation and speciation.

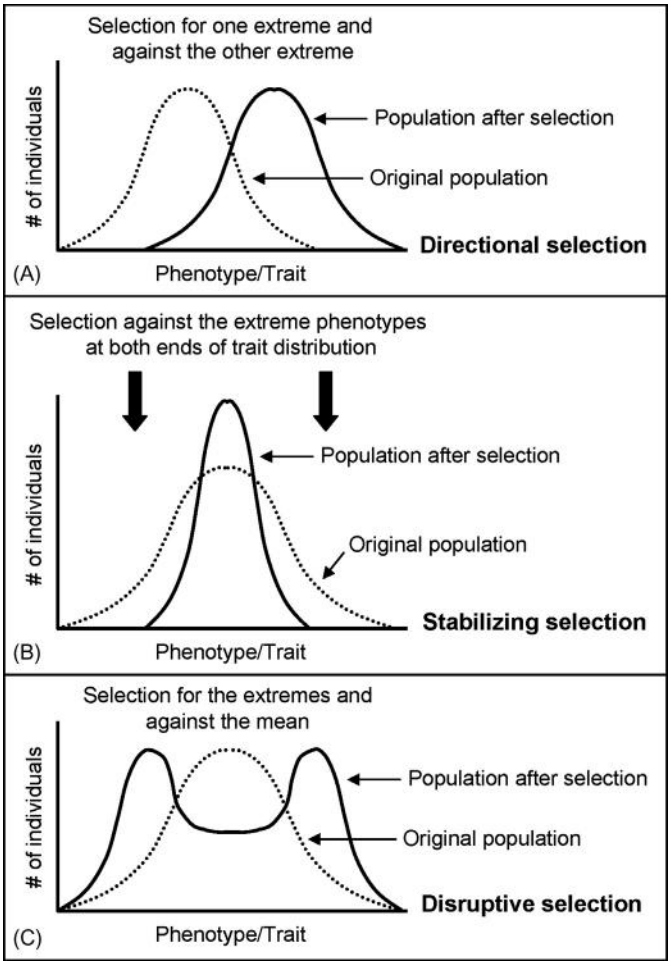

FIGURE 2.5 **Three types of natural selection.** (A) Directional selection; (B) stabilizing selection; (C) disruptive selection. See text for details.

ranges from reddish yellow to blue, with some intermediate colors. The extreme colors mimic other butterflies that are not normally preyed upon by the local predatory birds. In contrast, butterflies with intermediate coloration are devoured by the predators in greater numbers. Therefore, butterflies with extreme coloration survive in greater proportion compared to those with intermediate coloration. Another example of disruptive selection is the selection of the two extreme trophic phenotypes in the spadefoot toad (*Spea multiplicata*). Using a mark-recapture experiment in a natural pond, Martin and Pfennig[65] showed that the spadefoot toad can have different trophic phenotypes depending on the resource availability. However, disruptive selection favors the two extreme phenotypes, the small-headed "omnivore phenotype," which feeds mostly on detritus, and a large-headed "carnivorous" phenotype, which feeds on and whose phenotype is induced by the fairy shrimp. By foraging more effectively on the two alternative resource types, these extreme phenotypes avoid competition for food resources and are favored by disruptive selection, whereas the intermediate phenotypes are reduced in number.

Balancing selection (**balanced polymorphism**) maintains polymorphism in the population with respect to an allele of a trait. Therefore, balancing selection maintains genetic diversity in the population. A classic example of balancing selection is the **heterozygote advantage** in areas in Africa with high incidence of malaria. Sickle cell anemia reduces life expectancy and is caused if an individual is homozygous for a variant of hemoglobin (HbS/HbS). A red blood cell (RBC) containing HbS becomes sickle-shaped and is extremely sensitive to oxygen deprivation. However, the malarial parasite *Plasmodium* cannot survive in such sickle-shaped RBCs. Thus, heterozygous individuals, containing one normal copy and one variant copy of the hemoglobin gene (HbA/HbS), are at a survival advantage in areas with high incidence of malaria. In contrast, individuals homozygous for normal hemoglobin (HbA/HbA) are at an increased risk of death by malaria. Thus, selection maintains the apparently deleterious HbS allelic variant in the population, and balances between strong selection against both HbA/HbA and HbS/HbS genotypes by providing a selective advantage to the HbA/HbS genotype.

Based on the scale of changes, selection can lead to microevolution and macroevolution. **Microevolution** means small changes in the genome and is also associated with changes in gene frequency in a population. Over time, the accumulated small changes collectively can be significant enough to create certain new traits so that the group possessing those traits could be assigned an infra-species category, such as a **subspecies** or **variety** under the original species. In contrast, **macroevolution** means evolutionary changes leading up to the formation of species or higher taxa. The mechanisms for both micro- and macroevolutionary processes are generally the same.

2.4.4 Genetic Drift

Genetic drift (**also called random genetic drift**) means a change in the gene pool strictly by chance fixation of alleles. The effects of genetic drift can be acute in small populations and for infrequently occurring alleles, which can suddenly increase in frequency in the population or be totally wiped out. The alleles thus fixed by chance (genetic sampling error) may be neutral—that is, they may not confer any survival or reproductive advantage. Therefore, for small populations, genetic drift can result in a significant change in gene frequency in a short period of time.

Genetic drift can be caused by a number of chance phenomena, such as differential number of offspring left by different members of a population so that certain genes increase or decrease in number over generations independent of selection, sudden immigration or emigration of individuals in a population changing gene frequency in the resulting population, or population bottleneck. Of these, population bottleneck can cause a radical change in allele frequencies in a very short time. A population bottleneck occurs when a population suddenly shrinks in size owing to random events, such as sudden death of individuals due to environmental catastrophe, habitat destruction, predation, or hunting. When the small number of surviving individuals gives rise to a new population, there is a radical change in the gene frequency in the resulting population, in which certain genes (including rare alleles) of the original population may radically increase in proportion while others may radically decrease or be wiped out completely, independently of selection. Additionally, the resulting population contains a small fraction of the genetic diversity of the original population. The **founder effect** is a severe case of population bottleneck and happens when a few individuals migrate out of a population to establish a new subpopulation. Random genetic drift accompanies such founder effect, to severely reduce the genetic variation that exists in the original population. In the new population, the founder effect can rapidly increase the frequency of an allele whose frequency was very low in the original population. If the allele is a disease-related allele, the founder effect can lead to the prevalence of the disease in the new population. An increase in a specific disease in a human population due to the founder effect is seen in the Old Order Amish of eastern Pennsylvania,[66] and in the Afrikaner population of South Africa.[67]

The current Amish population has descended from a small number of German immigrants who settled in the United States during the eighteenth century. The incidence of Ellis–van Creveld syndrome (a form of dwarfism with polydactyly, abnormalities of the nails and teeth, and heart problems) is many times more prevalent in this Amish population than in the American population in general. The origin of this disease can be traced back to one couple, Samuel King and his wife, who came to the area in 1744. The mutated gene that causes the syndrome was passed along from the Kings and their offspring. The Amish population practices endogamy (individuals tend to mate within their own subgroup). Additionally, in this community the **gene flow is centrifugal**—that is, members may leave the community but outsiders do not join the community—therefore, there has been no introduction of exogenous genes into the Amish gene pool. As a result, the frequency of the disease gene has rapidly increased over generations.

Another example of founder effect comes from the Afrikaner population of South Africa, which is mainly descended from one group of European (mainly Dutch, but also German and French) immigrants that landed there in 1652. The present-day Afrikaner population has a very high prevalence of Huntington's disease; over 200 affected individuals in more than 50 supposedly unrelated families have been found to be ancestrally related through a common progenitor in the seventeenth century. Thus, the root of the disease can be traced back over 14 generations to a common progenitor who supposedly carried the gene for Huntington's disease. Huntington's disease is an autosomal dominant disease caused by triplet (CAG) repeat expansion in the gene (and the mRNA), containing 40 to >100 CAG triplets. The onset and severity of the disease is directly correlated with the number of repeats.

2.4.5 Nonrandom Mating

Changes in gene frequency by genetic drift are influenced in a large part by the breeding structure of the population—that is, whether the population practices random mating or nonrandom mating. **Inbreeding** is the most common form of nonrandom mating. Inbreeding occurs when genetically related individuals preferentially mate with each other (e.g. mating between relatives). The most extreme form of inbreeding is **self-fertilization**. Inbreeding produces a larger excess of homozygotes in the population than would be expected from random mating. Consequently, inbreeding also increases the frequency of homozygotes of rare alleles, including rare recessives, which will be subject to selection. If a rare allele is deleterious, its frequency can rise through homozygosity because of significant inbreeding in a normally outbreeding population. This phenomenon is called **inbreeding depression**.

Inbreeding is measured by the **inbreeding coefficient** (F), which is a measure of the probability that two alleles are identical by descent. This means the degree to which two alleles are more likely to be homozygous than heterozygous simply because the parents are genetically related. The value of F can theoretically range from 0 (0%; hence no inbreeding, completely random mating) to 1 (100%; hence complete inbreeding, all alleles are identical by descent).

If the frequency of allele A is p and the frequency of allele a is q, and the value of F is known, then the frequencies of genotypes AA, Aa and aa are determined as follows:

$$AA = p^2 + Fpq; \quad Aa = 2pq - 2Fpq; \quad Aa = q^2 + Fpq. \quad (2.1)$$

2.5 THE NEUTRAL THEORY OF EVOLUTION

The Darwinian theory of evolution by natural selection is based on the assumption that new mutations that constantly arise in the population are mostly adverse but some are beneficial. Natural selection filters out the adverse mutations, while fixing beneficial mutations in the population. In other words, evolution is caused by natural selection acting through beneficial mutations fixed in the population. Thus, it is an underlying assumption by Darwinian evolutionists that neutral mutations that do not confer any selective advantage or disadvantage are very rare, if they exist at all. A corollary to this assumption is that genetic drift, which causes chance fixation of neutral alleles, could not have played any role in evolution.

This long-held view of molecular evolution was challenged by the neutral theory of molecular evolution, proposed by Kimura.[68] In brief, the neutral theory postulates that evolutionary changes at the molecular level are not caused by natural selection alone acting only on advantageous mutations, but are mostly caused by random chance fixation of selectively neutral or near-neutral alleles (genetic drift). Therefore, genetic drift plays an important role in molecular evolution. To expand the concept, according to neutral theory, the majority of new mutations are either deleterious or neutral. Deleterious mutations adversely affect the fitness of the carrier whereas neutral mutations do not affect the fitness of the carrier (hence, selectively neutral). *Fitness in the context of evolution means the ability to reproduce, and contribute to the gene pool of the next generation*. Deleterious mutations that adversely affect fitness are removed from the population by purifying selection. In contrast, neutral mutations are subject to chance sampling and random fixation in every generation. In this process, some neutral mutations are fixed randomly by sheer chance while others are removed from the population. Once a neutral mutation is fixed by chance, its frequency increases by genetic drift, which leads to genetic polymorphism in the population. These genetic variations in the population provide the raw materials for molecular evolution. The allele carrying the new fixed mutation is called a **derived allele**, as opposed to the **ancestral allele** from which it is derived. As mentioned above, extensive neutral evolution can occur through random genetic drift while the phenotype is still maintained unchanged under stabilizing selection.[64]

It should be remembered that neutral theory does not deny the role of natural selection in evolution—that is, it does not deny the importance of positive selection in the origin of adaptations—it simply complements the Darwinian view by emphasizing the role of neutral mutations as additional raw materials for evolution and genetic drift as an additional mechanism of evolution. The neutral theory also predicts that purifying selection is ubiquitous, but positive selection is rare.[69]

2.5.1 Synonymous and Nonsynonymous Substitutions, Constraints on Changes in Gene and Protein Sequence, and Evolution

A nucleotide substitution that changes the corresponding amino acid in the protein is called a **nonsynonymous substitution** (denoted as K_A), whereas a nucleotide substitution that does not change the amino acid in the protein is called a **synonymous substitution** (denoted as K_S).

The neutral theory predicts that synonymous substitutions will be tolerated, but nonsynonymous substitutions will be removed by purifying selection. Consequently, nonsynonymous substitutions will be fewer than synonymous substitutions. Consistent with this prediction, it is known that synonymous substitutions typically exceed nonsynonymous substitutions in protein-coding genes, and functionally constrained regions of genes evolve at a slower rate than regions that are not functionally constrained. However, if a nonsynonymous substitution confers some selective advantage, then it will be rapidly fixed in the population by positive selection. The average rates of synonymous and nonsynonymous substitutions previously calculated were 4.7 substitutions/synonymous site versus 0.88 substitutions/nonsynonymous site per 10^9 (billion) years, respectively.[70] This estimate was subsequently revised to 3.51 substitutions/synonymous site versus 0.74 substitutions/nonsynonymous site per 10^9 (billion) years in rodents and humans, as stated earlier in this chapter.

2.5.2 Signatures of Positive Selection

A prediction of the neutral theory is that if the substitutions are all neutral, then for a given protein-coding gene the K_A/K_S ratio between two species should be very similar to the same ratio within species (null hypothesis), and it is the deviation from this prediction that provides support for positive selection (with some exceptions, such as relaxation of purifying selection and population bottleneck). McDonald and Kreitman[71] proposed a simple method to determine signatures of positive selection in protein sequence (see Box 2.4). The test relies on determining statistically significant deviation from the prediction of the neutral theory (the null hypothesis) that if the

BOX 2.4

THE MCDONALD–KREITMAN TEST

The McDonald–Kreitman method tests the neutral theory as the null hypothesis (H_0) against the (positive) selection hypothesis as the alternative hypothesis (H_1). In this test, two DNA sequences are aligned. Nucleotide substitutions in the coding region are classified in two ways: (1) **synonymous** versus **replacement**, and (2) **fixed difference** versus **polymorphic**.

1. **Synonymous** versus **replacement** substitutions: Synonymous substitutions result in a synonymous codon and no amino acid change in the protein, whereas replacement (or **nonsynonymous**) substitutions result in a nonsynonymous codon and amino acid change.
2. **Fixed difference** versus **polymorphic** substitutions: Polymorphic substitutions show variations within species, whereas fixed difference (also called **fixed divergence**) substitutions differ between species but not within species. Such dual classification allows the use of a 2×2 table. McDonald and Kreitman studied the sequence evolution of the *Adh* gene in *Drosophila melanogaster*, *Drosophila simulans*, and *Drosophila yakuba*. Tabulating the alignment data provided the following table:

	Fixed Difference (between species)	**Polymorphism (within species)**
Synonymous (K_S) (no amino acid change)	17	42
Replacement (K_A) (amino acid change)	7	2

$G = 7.43$; $P = 0.0006$.

McDonald and Kreitman used the *G*-test for statistical independence to determine if the cells in the 2×2 table were independent. In other words, whether the proportion of replacement versus synonymous changes was independent of whether the changes were fixed or polymorphic; similarly whether the proportion of fixed difference versus polymorphism was independent of whether the changes were synonymous or replacement.

The replacement/synonymous substitution ratio (K_A/K_S) of the fixed differences between species is 7/17 (= 0.41), whereas the same ratio of the polymorphic sites within species is 2/42 (= 0.048). *Thus, there is a more than eight-fold excess of replacement mutations between species compared to polymorphic mutations within species*. Similarly, the fixed difference/polymorphic substitution ratio among synonymous sites is 17/42 (= 0.40), whereas the same ratio among replacement sites is 7/2 (= 3.5). *Thus, there is a more than eight-fold excess of replacement substitutions compared to synonymous substitutions between species*. If all these substitutions were neutral, no such statistically significant differences would be expected. Therefore, the result of the *G*-test of independence indicates deviation from the assumptions of neutral evolution, thereby signifying a strong signature of positive selection.

substitutions are all neutral, then for a given protein-coding gene, the K_A/K_S ratio at divergent sites **between** species should be very similar to the same ratio at polymorphic sites **within** species. Deviation from the null hypothesis will constitute evidence of positive selection.

Signatures of positive selection, however, are not very widespread, except in some select groups of genes, such as genes important in host–pathogen interactions, as well as in sex-related genes. For example, strong signatures of positive selection, with K_A/K_S ratios ranging from 1.36 to 5.15, were observed when two proteins (16 and 18 kDa) in the acrosomal vesicle of abalone spermatozoa were compared. These values were among the highest for full-length sequences analyzed so far.[72]

2.5.3 Selective Sweep and the Hitchhiking Effect

If a new mutation offers increased fitness to the carrier, it is fixed in the population through positive selection, and its frequency rapidly increases. Such rapid fixation of an advantageous mutation is called **selective sweep**. As the frequency of the new mutation

BOX 2.5

NEUTRAL EVOLUTION–MUTATION RELATIONSHIP

1. The probability of fixation of a mutation (p) in a diploid population of size N is $1/2N$ (i.e. $p = 1/2N$).
2. The rate of substitution per unit time (k) in a diploid population of size N = the number of mutations fixed per unit time in a diploid population of size $N \times$ the probability of fixation of a mutation (p).
3. Because the number of mutations fixed per unit time is the mutations rate μ, and the number of any gene in a diploid population of size N is $2N$, the number of mutations fixed per unit time in a diploid population of size $N = 2N \times \mu$.
4. Hence, point (2) stated above can be expressed as $k = 2N \times \mu \times p$.
5. Because $p = 1/2N$, p can be substituted for $1/2N$ and point (2) can be rewritten as $k = 2N \times \mu \times 1/2N$; or $k = \mu$.
6. In other words, the rate of substitution per unit time—i.e. the rate of neutral evolution (k)—is equal to the mutation rate (μ) of neutral alleles, and is independent of the population size.

increases, the frequency of the genes/sequences around it that are very closely linked and not easily separated by recombination also increases. The net result is a loss of sequence variability around the newly fixed mutation in the population. The increase in frequency of the neighboring genes/sequences, simply because of their close proximity to the newly fixed mutation, is called the **hitchhiking effect**, or **genetic hitchhiking**. Selective sweep and the hitchhiking effect are the results of strong positive selection. The hitchhiking effect may also lead to an increase in the proportion of somewhat disadvantageous or deleterious mutations in the population.[73]

2.6 MOLECULAR CLOCK HYPOTHESIS IN MOLECULAR EVOLUTION

Kimura's neutral theory derived support from the **molecular clock hypothesis**. The molecular clock hypothesis states that the rate of molecular evolution of a gene (the rate of nucleotide substitution) or a protein (the rate of amino acid substitution) is approximately constant over evolutionary time. In other words, the number of replacements in the gene or protein is proportional to the time since their origin—that is, the number of replacements per unit time is similar. The hypothesis was based on the initial observation of amino acid substitutions in human and horse hemoglobin by Zuckerkandl and Pauling in 1962. This was followed by similar observations on cytochrome c from seven different eukaryotic species: horse, human, pig, rabbit, chicken, tuna, and baker's yeast.[74] The term "molecular clock hypothesis" was coined by Zuckerkandl and Pauling in 1965. The concept of the molecular clock fits well with Kimura's neutral theory because the rate of neutral evolution is equal to the mutation rate of neutral alleles, as shown in Box 2.5.

However, after more protein sequences were studied in the 1970s, it was realized that the rate of substitution could differ significantly in different proteins and different organisms. Nonetheless, the molecular clock represents a valuable tool in studies of evolution and molecular systematics, and it has been widely used in estimation of divergence times and reconstruction of phylogenetic trees.

2.7 MOLECULAR PHYLOGENETICS

Phylogeny refers to the evolutionary history of organisms or populations. **Phylogenetics** is the study of phylogenies—that is, the study of the evolutionary relationships among various organisms and populations. According to evolutionary theory, the similarity among organisms and groups of organisms is attributable to their descent from a common ancestor. This similarity extends even to the structure and function of molecules, such as DNA and proteins. Traditional phylogenetics considered morphological features. Modern phylogenetics uses information from DNA and protein sequences. The use of DNA and protein sequence information and their change over evolutionary time in order to infer the evolutionary relationship among a set of homologous genes or proteins is referred to as **molecular phylogenetics**. The goal of molecular phylogenetics is to estimate the evolutionary divergence of the DNA and protein sequences from a common ancestral sequence, and thus reconstruct the correct evolutionary relationships among these sequences in the form of a phylogenetic

tree. With the advent of molecular biology techniques, particularly DNA sequencing, molecular phylogenetic studies have become very common. Sometimes molecular phylogenetics is used to infer the evolutionary relationships among organisms. *In general, inference on evolutionary relationships based on protein sequences is preferred to that based on nucleic acid sequences.*

2.7.1 From Systematics and Biological Classification to Molecular Phylogenetics

> Systematics is the scientific study of the kinds and diversity of organisms and of any and all relationships among them ... Classification of organisms is an activity that belongs exclusively to systematics. ***G. G. Simpson***[75]

Biological classification is concerned with ordering (arranging) organisms or groups of organisms, both **living** (**extant**) and **fossil** (**extinct**), into hierarchical and multilevel categories based on their evolutionary relationships. Therefore, the conceptual foundation of the science of systematics and the activity of biological classification is the evolutionary (phylogenetic) relationship among taxa. The expression **phylogenetic systematics** (also known as **cladistics**, discussed in Section 2.7.2.2) underscores the link between systematics and phylogeny. Because classification of organisms takes into consideration their evolutionary relationships, the revision of older classification schemes with modern data, particularly ancestral and derived characters and homology (discussed later under cladistics), has affected only minor details.[76] With the availability of the vast amount of molecular data and analytical tools, molecular phylogenetics has become the norm for studying the evolutionary relationships. Nevertheless, for historical reasons it is appropriate to consider molecular phylogenetics against the backdrop of systematics and biological classification.

The first systematic way of classifying organisms was introduced by the Swedish botanist Carl Linnaeus. Linnaeus's classification scheme involved categorizing organisms based solely on morphological characters without any evolutionary context. He published his work as a book called ***Systema Naturae***. The 10th edition of *Systema Naturae*, published in 1758, is considered to be the beginning of biological classification and the **binomial nomenclature** system in biology. In binomial nomenclature, an organisms is given a name composed of two parts, usually using latinized expression; the first part identifies the genus to which the species belongs and the second part identifies the species within the genus. The original Linnaean classification scheme is called **Linnaean hierarchy**, and it had seven categories: kingdom, phylum, class, order, family, genus, and species. These categories are called **taxonomic categories**. Organisms that are the subjects of classification are called **taxa** (singular: **taxon**). Modern biological classification systems have many more taxonomic categories compared to the seven originally proposed by Linnaeus.

Linnaeus introduced his system of classification 100 years before the theory of evolution was proposed by Darwin; hence, it had no evolutionary context. Linnaeus's classification scheme was based on choosing "similar" characters, and such choice was more or less arbitrary. With a greater understanding of genetics—including population genetics, mechanism of evolution, and relationships among the living and extinct organisms at the biochemical and molecular levels—it became apparent that biological classification should reflect the relationships among organisms or groups of organisms by their descent from a common ancestor during evolution. *The meaning of "similarity" in modern biological classification is ancestral similarity (homology).*

2.7.2 Systems of Biological Classification

The three main systems of modern biological classification are **phenetics**, **cladistics**, and **evolutionary classification**. For all practical purposes, phenetics is no longer used as a phylogenetic method, whereas cladistics has become the most widely used method for molecular phylogenetic analysis.

2.7.2.1 Phenetics and Phenograms

Phenetics, also known as **numerical taxonomy**, was introduced in the 1950s.[77] Phenetics attempts to group species into higher taxa based on overall similarity, usually in morphology or other observable traits, and regardless of their phylogeny or evolutionary relationships. Many different characteristics are used to calculate a similarity coefficient, varying between 0 (no similarity) to 1 (highest similarity), between all pairs of organisms that are subjects of phenetic classification. Similarity coefficients are used to create a similarity matrix and develop a **phenogram**, which is a tree-like network expressing phenetic relationships. According to the proponents of phenetics, similarity is expected among the descendants of a common ancestor; therefore, grouping together the most similar taxa automatically produces phylogenetic classification. Although phenetics is not used anymore, its historical importance lies in introducing computer-based numerical algorithms, which are now essential in all modern phylogenetic analyses.

2.7.2.2 Cladistics, Clades, and Cladograms

The main proponent of cladistics was the German entomologist Willi Hennig in the mid-twentieth century.

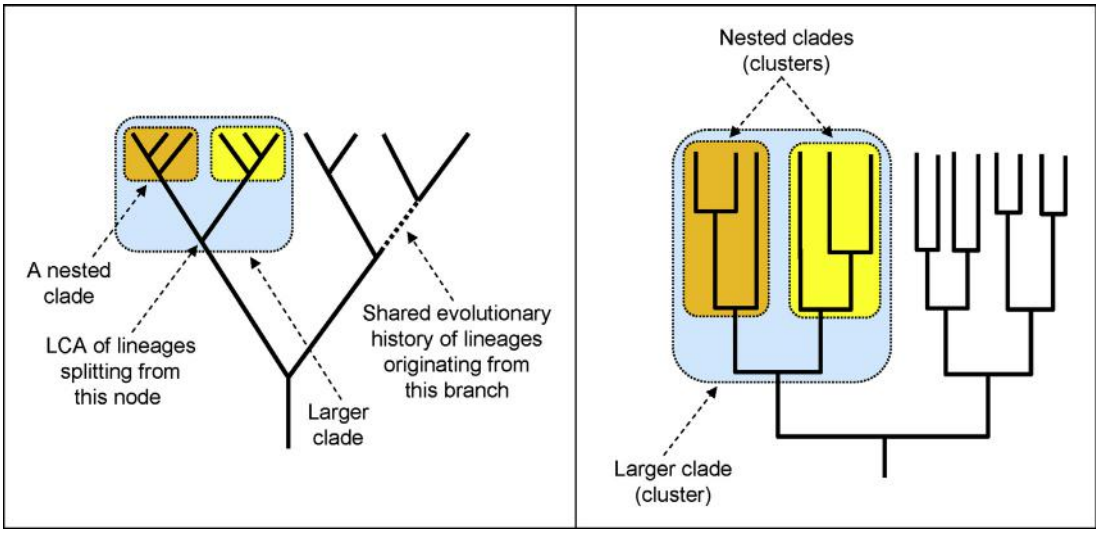

FIGURE 2.6 **Nested clades within a larger clade in a phylogenetic tree.** A typical cladogram on the left and a typical dendrogram on the right. In a phylogenetic tree, each branching point (node) represents the LCA of the lineages (including nodes) arising from this point. A branch preceding a node represents the shared evolutionary history of lineages that split from the node.

Cladistics is also known as **phylogenetic systematics** or **phylogenetic classification**. Cladistics classifies organisms based on shared derived characters. Therefore, taxa that share specific derived characters are grouped more closely together than those who do not. The groups are called **clades**; each clade consists of an ancestor and all of its descendants. The relationships between clades are shown in a branching hierarchical tree called a **cladogram**. Depending on the branching of the cladogram, it is possible to identify smaller clades within a larger clade; the smaller clades are called **nested clades.** Figure 2.6 shows nested clades within a larger clade in a phylogenetic tree. The phylogenetic tree has been represented as a typical cladogram on the left and as a typical **dendrogram** on the right. The dendrogram is sometimes loosely called a cladogram. In a phylogenetic tree (cladogram), each branching point (node) represents the **last common ancestor** (**LCA**) of the lineages (including nodes) arising from this point. The separation of taxa along the cladogram is driven by evolutionary innovation of new characters (evolutionary novelties or apomorphies, discussed below).

2.7.2.2.A SOME IMPORTANT TERMINOLOGY OF CLADISTICS

Terms used to describe various character states that are relevant in the discussion of cladistics include **apomorphy**, **synapomorphy**, **plesiomorphy**, **symplesiomorphy**, **autapomorphy**, and **homoplasy**. The terms are described below with examples.

A **primitive** or **ancestral character state** is called **plesiomorphy** (**plesiomorphic character**), and a shared plesiomorphy is called a **symplesiomorphy**. For example, hair is a unique mammalian character that evolved with the evolution of mammals. Mammalian evolution was followed by further evolution of various mammalian groups and subgroups based on evolutionary novelties. For example, primates form a more recently evolved mammalian group. Therefore, hair is a plesiomorphy (ancestral character) for primates. Because hair, as an ancestral mammalian character, is shared by all primates, it is also a symplesiomorphy (shared plesiomorphy) for primates in general.

In contrast to an ancestral character state, a **derived character state** (**evolutionary novelty**) is called **apomorphy** (**apomorphic character**), and a shared apomorphy is a **synapomorphy**. For example, hair is an apomorphy for mammals as a group because it distinguishes mammals from other vertebrate clades, such as reptiles. Because hair is shared by all mammals, it is also the synapomorphy (shared apomorphy) for mammals in general. Among mammals, different groups have their own apomorphies. For example, an opposable thumb is an apomorphy for primates because it is an evolutionary novelty for primates and is not found in non-primate mammals. Similarly, the feather is an apomorphy for birds. Therefore, an apomorphy for a larger clade can be a plesiomorphy for a smaller nested clade within that larger clade.

An apomorphy that is unique to a taxon is called **autapomorphy**. An example of a non-anatomical autapomorphy in modern humans is speech, which is unique to humans.

A character state that evolved because of **convergent evolution** but was not acquired through common evolutionary lineage is called **homoplasy**, and the

character is called a **homoplastic character**. Homoplastic characters evolve independently in multiple taxa in different evolutionary lineages in response to adaptation; these characters are not present in their common ancestor. For example, fins evolved independently in sharks (cartilaginous fish) and dolphins (mammals) to perform the same function, but they are structurally different and were not derived from their common ancestor. Hence, the fin is a homoplastic character for sharks and dolphins. In contrast to homoplasy, **homology** is a character state shared by a set of species and is present in their common ancestor. The term homology is pervasive in the evolutionary literature, including molecular evolution.

2.7.2.3 *Evolutionary Classification*

The third system of modern biological classification is referred to as **evolutionary classification**, also known as **Darwinian classification**, **evolutionary taxonomy**, and **evolutionary systematics**. It is actually the oldest of the three approaches and its strongest proponents include renowned evolutionary biologists such as Ernst Mayr, George Gaylord Simpson, and Julian Huxley. Mayr and Bock[78] emphasized that, contrary to the general belief, not all biological classifications are evolutionary classifications. They opined that evolutionary classification is more inclusive than ordering systems (e.g. phenetics and cladistics), which are based on just the pattern of branching points. Nevertheless, ordering systems producing dendrograms and cladograms are still useful phylogenetic classification schemes. Proponents of evolutionary classification maintain that classifications should reflect the two aspects of evolutionary change: (1) the splitting of the phyletic lineages—that is, the branching in the phylogenetic tree—and (2) the invasion of new environmental niches—that is, adaptation and evolutionary divergence. Therefore, the amount of evolutionary change after the branching points is an important consideration in evolutionary classification. In order to take account of this, evolutionary classification weighs the evolutionary innovations (apomorphic characters) that determine the branching point in the tree. Major evolutionary innovations that help a new phyletic lineage adapt to a new environment and drive adaptive evolution are given greater weight. Therefore, evolutionary classification tries to tell the evolutionary history of the taxonomic group.

Each of the three methods discussed above has its own strengths and shortcomings, and the proponents of each method claim that their method is the best. However, cladistics has become the method of choice for molecular phylogenetic analysis because of the molecular (sequence) data used to measure divergence from an ancestral taxon. This is probably why the use of cladistics has progressively increased with the increase in the number of entries in DNA and protein sequence databases, and has now become commonplace in molecular phylogenetic analysis.

2.7.3 Phylogenetic Tree

A **phylogenetic tree** or **evolutionary tree** is a diagrammatic representation of the evolutionary relationship among various taxa. The phylogenetic tree, including its reconstruction and reliability assessment, is discussed in more detail in Chapter 9. The terms **evolutionary tree**, **phylogenetic tree**, and **cladogram** are often used interchangeably to mean the same thing—that is, the evolutionary relationships among taxa. The term dendrogram is also used interchangeably with cladogram, although there are subtle differences, discussed in Chapter 9. Thus, it is important to be aware that usage of the vocabulary is not always consistent in the literature, although the context is the same, that is, representation of the evolutionary relationships of taxa.

References

1. Zuckerkandl E, Pauling L. *J Theor Biol* 1965;**8**:357–66.
2. Higgs PG, Attwood TK. *Bioinformatics and molecular evolution*. MA: Blackwell; 2005. pp. 1–11
3. Mayr E. *Populations, species, and evolution*. New York: Belknap Harvard; 1970. pp. 10–20
4. Gould SJ, Eldredge N. *Paleobiology* 1977;**3**:115–51.
5. Mills DR, et al. *Proc Natl Acad Sci USA* 1967;**58**:217–24.
6. Cairns J, et al. *Nature* 1988;**335**:142–5.
7. Trottier Y, et al. *Mol Carcinogen* 1992;**6**:140–7.
8. Bruner SD. *Nature* 2000;**403**:859–66.
9. Eichler EE, et al. *Hum Mol Genet* 1995;**4**:2199–208.
10. Falik-Zaccai TC, et al. *Am J Human Genet* 1997;**60**:103–12.
11. Eichler EE, et al. *Nat Genet* 1995;**11**:301–8.
12. Cabot EL, et al. *Genetics* 1993;**135**:477–87.
13. Ohno S. *Evolution by gene duplication*. New York: Springer-Verlag; 1970.
14. Hokamp K, et al. *J Struct Funct Genom* 2003;**3**:95–110.
15. Kasahara M. *Curr Opin Immunol* 2007;**19**:547–52.
16. Makałowski W. *Genome Res* 2001;**11**:667–70.
17. Lynch M, Conery JS. *Science* 2000;**290**:1151–5.
18. Lynch M, Conery JS. *J Struct Func Genom* 2003;**3**:35–44.
19. Cotton JA, Page RDM. *Proc R Soc Lond B* 2005;**272**:277–83.
20. Graur D, Li WH. *Fundamentals of molecular evolution*. 2nd ed. Sunderland, MA: Sinauer; 2000. pp. 99–164
21. Hirotsune S, et al. *Nature* 2003;**423**:91–6.
22. Thibaud-Nissen F, et al. *BMC Genomics* 2009;**10**:317.
23. Deng C, et al. *Proc Natl Acad Sci USA* 2010;**107**:21593–8.
24. Escriva H, et al. *PLoS Genet* 2006;**2**:e102.
25. Lynch M. *The origins of genome architecture*. Sunderland, MA: Sinauer; 2007. pp. 193–235
26. Rastogi S, Liberles DA. *BMC Evol Biol* 2005;**5**:28.
27. Force A, et al. *Genetics* 1999;**151**:1531–45.
28. Hahn MW. *J Heredity* 2009;**100**:605–17.
29. Kleinjan DA, et al. *PLoS Genet* 2008;**4**:e29.

30. Tocchini-Valentini GD, et al. *Proc Natl Acad Sci USA* 2005;**102**:8933–8.
31. Kolkman JA, Stemmer WP. *Nat Biotechnol* 2001;**19**:423–8.
32. Graur D, Li WH. *Fundamentals of molecular evolution*. 2nd ed. Sunderland, MA: Sinauer; 2000. pp. 249–322
33. International Human Genome Sequencing Consortium. *Nature* 2001;**409**:860–921.
34. Kaessmann H, et al. *Genome Res* 2002;**12**:1642–50.
35. Franca GS, et al. *Genetica* 2012;**140**:249–57.
36. Patthy L. *Gene* 1999;**238**:103–14.
37. Snel B, et al. *Trends Genet* 2000;**16**:9–11.
38. Enright A, Ouzounis C. *Genome Biol* 2001;**2**:34.1–7.
39. Yanai I, et al. *Proc Natl Acad Sci USA* 2001;**98**:7940–5.
40. Kummerfeld SK, Sarah A. *Trends Genet* 2005;**21**:25–30.
41. Syvanen M. *Annu Rev Genet* 2012;**46**:341–58.
42. Acuña R, et al. *Proc Natl Acad Sci USA* 2012;**109**:4197–202.
43. Moran NA, Jarvik T. *Science* 2010;**328**:624–7.
44. Altincicek B, et al. *Biol Lett* 2012;**8**:253–7.
45. Tautz D, Domazet-Lošo T. *Nat Rev Gen* 2011;**12**:692–702.
46. Levine MT, et al. *Proc Natl Acad Sci USA* 2006;**103**:9935–9.
47. Begun DJ, et al. *Genetics* 2007;**176**:1131–7.
48. Zhou Q, et al. *Genome Res* 2008;**18**:1446–55.
49. Knowles DG, McLysaght A. *Genome Res* 2009;**19**:1752–9.
50. Li C-Y, et al. *PLoS Comput Biol* 2010;**6**:e1000734.
51. Wu D-D, et al. *PLoS Genet* 2011;**7**:e1002379.
52. Sorek R, et al. *Mol Cell* 2004;**14**:221–31.
53. Schmitz J, Brosius J. *J Biochim* 2011;**93**:1928–34.
54. Butticè G, et al. *J Mol Evol* 1990;**30**:479–88.
55. Purevsuren J, et al. *Mol Genet Metab* 2008;**95**:46–51.
56. Raponi M, et al. *Hum Mut* 2006;**27**:294–5.
57. Schlager G, Dickie MM. *Genetics* 1967;**57**:319–30.
58. Schlager G, Dickie MM. *Mutat Res* 1971;**11**:89–96.
59. Russell LB, Russell WL. *Proc Natl Acad Sci USA* 1996;**93**:13072–7.
60. Monant Jr. RJ. *Cancer Res* 1989;**49**:81–7.
61. Drake JW, et al. *Genetics* 1998;**148**:1667–86.
62. Graur D, Li WH. *Fundamentals of molecular evolution*. 2nd ed. Sunderland, MA: Sinauer; 2000. pp. 5–38
63. Mills RE, et al. *Genome Res* 2011;**21**:830–9.
64. Kimura M. *Proc Natl Acad Sci USA* 1981;**78**:5773–7.
65. Martin RA, Pfennig DW. *Am Nat* 2009;**174**:268–81.
66. McKusick VA. *Nat Genet* 2000;**24**:203–4.
67. Hayden MR, et al. *S Afr Med J* 1980;**58**:197–200.
68. Kimura M. *Nature* 1968;**217**:624–6.
69. Hughes AL. *Heredity* 2007;**99**:364–73.
70. Li WS, et al. *Mol Biol Evol* 1985;**2**:150–74.
71. McDonald JH, Kreitman M. *Nature* 1991;**351**:652–4.
72. Vacquier VD, et al. *J Mol Evol* 1997;**44**(Suppl. 1):S15–22.
73. Chun S, Fay JC. *PLoS Genet* 2011;**7**:e1002240.
74. Margoliash E. *Proc Natl Acad Sci USA* 1963;**50**:672–9.
75. Simpson GG. *Principles of animal taxonomy*. New York: Columbia University Press; 1961. pp. 1–33
76. Mayr E, Ashlock PD. *Principles of systematic zoology*. 2nd ed. New York: McGraw-Hill. 1991, pp. 113–58.
77. Mayr E, Ashlock PD. *Principles of systematic zoology*. 2nd ed. New York: McGraw-Hill. 1991, pp. 195–205.
78. Mayr E, Bock WJ. *J Zool Syst Evol Res* 2002;**40**:169–94.