# MOLECULAR EVOLUTION AND PHYLOGENETICS

# 1

# Molecular Basis of Evolution

## 1.1. Evolutionary Tree of Life

From the time of Charles Darwin, it has been a dream for many biologists to reconstruct the evolutionary history of all organisms on Earth and express it in the form of a phylogenetic tree (Haeckel 1866). The ideal approach to this problem is to use the fossil record, but since the fossil record is fragmentary and incomplete, most investigators have used the methods of comparative morphology and comparative physiology. Using this approach, classical evolutionists have been able to infer the major aspects of the evolutionary history of organisms. However, the evolutionary change of morphological and physiological characters is so complex that this approach does not produce a clear-cut picture of evolutionary history, and the details of the phylogenetic trees reconstructed have almost always been controversial.

Recent advances in molecular biology have changed this situation drastically. Since the blueprint of all organisms is written in deoxyribonucleic acid (DNA) (ribonucleic acid [RNA] in some viruses), one can study the evolutionary relationships of organisms by comparing their DNA. This approach has several advantages over the classical approach in which morphological and physiological characters are used. First, DNA consists of the four types of nucleotides, adenine (A), thymine (T), cytosine (C), and guanine (G), and it can be used for comparing any groups of organisms, including bacteria, plants, and animals. In the classical approach, this is virtually impossible. Second, since the evolutionary change of DNA follows a more or less regular pattern, it is possible to use a mathematical model to formulate the change and compare DNAs from distantly related organisms. The evolutionary change of morphological characters is extremely complicated even for a short evolutionary time. Therefore, it is not clear whether various assumptions required for morphological phylogenetics are really satisfied or not. Third, the genomes of all organisms consist of long sequences of nucleotides and contain a much larger amount of phylogenetic information than morphological characters. For these reasons, molecular phylogenetics is expected to clarify many branching patterns of the tree of life that have been hard to resolve by the classical approach.

Systematics or taxonomy is one of the most controversial areas of biology. The definition of species, genera, families, and others is often subjective, and it is not uncommon that two experts working on the same group of organisms (e.g., *Drosophila*) vehemently disagree about the assignment of organisms to subspecies, species, genera, and so forth. Phylogenetics is less controversial than systematics, because it is primarily concerned with the evolutionary relationships of organisms, and the assignment of a group of organisms to a given taxonomic rank is of secondary importance. Nevertheless, the two areas of biology are closely related to each other, because the classification of organisms is conducted to reflect their evolutionary histories (Darwin 1859; Mayr 1968). In this sense, phylogenetics plays an important role in developing a scientific basis of systematics, though it may not solve all the problems of the latter discipline. Recent advances in molecular phylogenetics have already provided new insights into various aspects of classification of organisms, as will be mentioned later.

## 1.2. Mechanism of Evolution

The primary cause of evolution is the mutational change of genes. A mutant gene or DNA sequence caused by nucleotide substitution, insertions/deletions, recombination, gene conversion, and so forth may spread through the population by genetic drift and/or natural selection (see, e.g., Nei 1987; Hartl and Clark 1997) and eventually be fixed in a species. If this mutant gene produces a new morphological or physiological character, this character will be inherited by all the descendant species unless the gene mutates again. Therefore, if we establish a valid phylogenetic tree for a group of species, we are able to identify the lineage of species in which any specific character appeared by mutation.

This information is useful in understanding the mechanism of evolution of any specific character of interest. Comparison of the environmental conditions of this lineage of species with those of species lacking the character may suggest whether the character evolved by a particular process of natural selection or by genetic drift. If we can identify the genes involved and study their evolutionary change, we will know what kind of mutational change has generated the particular morphological or physiological character.

This type of study is already being conducted with respect to the enzymes (lysozymes and ribonucleases) that are associated with the evolution of the two-gut digestion system of ruminants and the langur monkey (Stewart et al. 1987; Jermann et al. 1995). The foregut of these animals harbors bacteria that can ferment grasses and tree leaves. These bacteria are then digested by lysozymes in the hind gut, and the RNA released is decomposed by ribonucleases. Fermentation mix, including digested bacteria, provides nutrients to host animals (Barnard 1969). It is now possible to infer the amino acid sequences of the proteins of ancestral organisms by statistical methods (e.g., Fitch 1971; Maddison and Maddison 1992; Yang et al. 1995b) and then reconstruct the ancestral proteins by site-directed mutagenesis. Therefore, we can study the catalytic activity of the ancient proteins (e.g., Jermann et al. 1995). In this way, it is possible to study the evolutionary change of gene function.

Determination of the relative importance of mutation, natural selection, genetic drift, recombination, and so forth is an important subject in population genetics. For this purpose, population geneticists are now sequencing different alleles at a locus to clarify the evolutionary histories of the alleles. Here the problem is not the phylogenetic tree of different species but of different alleles within species. One of the interesting results obtained from this type of study is that some allelic lineages at the major histocompatibility complex (MHC) loci in mammals have persisted in the population for millions of years (e.g., Figueroa et al. 1988; Lawlor et al. 1988; McConnell et al. 1988; Hughes and Yeager 1998). This observation is consistent with the view that the antigen-recognition site of MHC molecules is subject to overdominant selection (Hughes and Nei 1988). Phylogenetic analyses of polymorphic alleles also have shown that intragenic recombination can occur with an appreciable frequency within species (e.g., Robertson et al. 1995; Fitch 1997).

Phylogenetic analysis of polymorphic alleles may also give important information about the extent of gene flow between two populations. Three decades ago, Prakash et al. (1969) examined the electrophoretic alleles at some enzyme loci in the North American and South American (Bogota, Colombia) populations of *Drosophila pseudoobscura* and found that many alleles are shared by the two populations. From this observation, they concluded that the Bogota population was formed only recently, possibly around 1950, from migrants from North America. However, this conclusion was questioned by Coyne and Felton (1977) in their detailed analysis of electrophoretic alleles. Later Schaeffer and Miller (1991) studied this problem by constructing a phylogenetic tree for the DNA sequences of polymorphic alleles of the alcohol dehydrogenase gene from the North and South American populations and concluded that the Bogota population was probably formed more than 100,000 years ago.

As is clear from the above examples, molecular phylogenetics has become an important tool for studying the mechanism of evolution.

## 1.3. Structure and Function of Genes

Although the molecular biology of the gene is beyond the scope of this book, let us explain the basic structure and function of genes that are important for understanding this book. In terms of function, genes can be classified into two groups: **protein-coding genes** and **RNA-coding genes.** Protein-coding genes are transcribed into **messenger RNAs** (**mRNA**), which are in turn translated into the amino acid sequences of proteins. RNA-coding genes are those that produce **transfer RNAs** (**tRNA**), **ribosomal RNAs** (**rRNA**), **small nuclear RNAs** (**snRNA**), and so on. These nonmessenger RNAs are the final products of RNA-coding genes. Ribosomal RNAs are components of ribosomes that are the core of the machinery of protein synthesis, whereas tRNAs are essential in transferring the genetic information of mRNAs into amino acid sequences of pro-

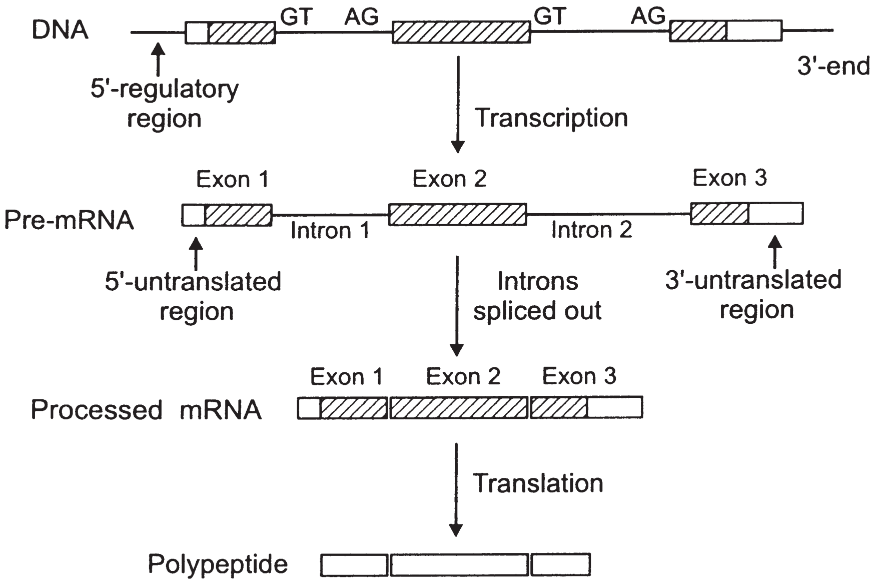

FIGURE 1.1. Basic structure of a eukaryotic protein-coding gene in relation to transcription and translation.

teins. snRNAs are confined to the nucleus, and several of them are involved in intron splicing or other RNA processing reactions.

The basic structure of a protein-coding gene in eukaryotes is presented in Figure 1.1. It is a long linear arrangement of four nucleotides, A, T, C, and G, and consists of a transcribed part of DNA and the 5′ and 3′ nontranscribed flanking regions. The flanking regions are necessary for controlling transcription and processing of **pre-messenger RNAs** (**pre-mRNA**). A pre-mRNA consists of coding regions and noncoding regions. Coding regions contain information for encoding amino acids in the polypeptide produced by the gene, whereas noncoding regions contain some information necessary for regulation of polypeptide production. Some segments of noncoding regions are spliced out in the process of production of a **mature mRNA.** These segments are called **introns,** and the remaining regions are called **exons** (Figure 1.1). The number of exons in a gene varies from gene to gene. Prokaryotic genes have no introns, whereas some eukaryotic genes (e.g., muscular dystrophy gene) have as many as 78 introns (Roberts et al. 1992). The functional role of introns is not well understood. Usually, an intron begins with the dinucleotide GT and ends in AG. These dinucleotides provide context for correct intron splicing.

The genetic information carried by the nucleotide sequence of a gene is first transferred to mRNA by a simple process of one-to-one transcription of the nucleotides. The genetic information transferred to mRNA determines the amino acid sequences of the protein produced. Nucleotides of mRNA are read sequentially, three at a time. Each such triplet or **codon** is translated into a particular amino acid in the growing polypeptide chain according to the genetic code.

Table 1.1 Standard or “universal” genetic code.

| Codon | | Codon | | Codon | | Codon | |
|---|---|---|---|---|---|---|---|
| UUU | Phe | UCU | Ser | UAU | Tyr | UGU | Cys |
| UUC | Phe | UCC | Ser | UAC | Tyr | UGC | Cys |
| UUA | Leu | UCA | Ser | UAA | Ter | UGA | Ter |
| UUG | Leu | UCG | Ser | UAG | Ter | UGG | Trp |
| CUU | Leu | CCU | Pro | CAU | His | CGU | Arg |
| CUC | Leu | CCC | Pro | CAC | His | CGC | Arg |
| CUA | Leu | CCA | Pro | CAA | Gln | CGA | Arg |
| CUG | Leu | CCG | Pro | CAG | Gln | CGG | Arg |
| AUU | Ile | ACU | Thr | AAU | Asn | AGU | Ser |
| AUC | Ile | ACC | Thr | AAC | Asn | AGC | Ser |
| AUA | Ile | ACA | Thr | AAA | Lys | AGA | Arg |
| AUG | Met | ACG | Thr | AAG | Lys | AGG | Arg |
| GUU | Val | GCU | Ala | GAU | Asp | GGU | Gly |
| GUC | Val | GCC | Ala | GAC | Asp | GGC | Gly |
| GUA | Val | GCA | Ala | GAA | Glu | GGA | Gly |
| GUG | Val | GCG | Ala | GAG | Glu | GGG | Gly |

The **genetic code** for nuclear genes seems to be universal for both prokaryotes and eukaryotes with a few exceptions. The same genetic code (“**universal**” or **standard genetic code**) is used for chloroplast genes, but mitochondrial genes use slightly different genetic codes. The standard genetic code is presented in Table 1.1. In this table, amino acids are represented by three-letter codes (see Table 1.2). There are $4^3 = 64$ possible codons for the four different nucleotides, uracil (U), cytosine (C), adenine (A), and guanine (G). (U corresponds to T in nucleotide sequences.) Three of the codons (UAA, UAG, UGA) are, however, **termination** or **stop codons** and do not code for any amino acid. Each of the remaining 61 codons (**sense codons**) codes for a particular amino acid, but since there are only 20 amino acids (Table 1.2) used for making proteins, there are many codons that code for the same amino acid. Codons coding for the same amino acid are called **synonymous codons.** In the genetic code table, codon AUG codes for methionine, but this codon is also used as the **initiation codon.** The methionine encoded by the initiation codon is in a modified form and is later removed from the polypeptide. Recent studies have shown that CUG and UUG are also used as the initiation codons in some nuclear genes (Elzanowski and Ostell 1996). These initiation codons should be excluded in a study of DNA sequence evolution, because they remain unchanged in most cases. Termination codons should also be eliminated.

Table 1.3 shows the genetic code for vertebrate mitochondrial genes. There are a few differences between this genetic code and the standard nuclear genetic code. In the mitochondrial genetic code, codon UGA is not a termination codon but codes for tryptophan. By contrast, codons AGA and AGG are termination codons instead of an arginine codon. AUA, which codes for isoleucine in the nuclear code, is used for encod-

Table 1.2 One- and three-letter amino acid codes.

| | Code | | |
|---|---|---|---|
| Name | 1-Letter | 3-Letter | Property of the Side Chain at pH 7 |
| Alanine | A | Ala | Nonpolar (hydrophobic) |
| Cysteine | C | Cys | Polar |
| Aspartic acid | D | Asp | Polar (hydrophilic, acidic) |
| Glutamic acid | E | Glu | Polar (hydrophilic, acidic) |
| Phenylalanine | F | Phe | Nonpolar (hydrophobic) |
| Glycine | G | Gly | Nonpolar |
| Histidine | H | His | Polar (hydrophilic, basic) |
| Isoleucine | I | Ile | Nonpolar (hydrophobic) |
| Lysine | K | Lys | Polar (hydrophilic, basic) |
| Leucine | L | Leu | Nonpolar (hydrophobic) |
| Methionine | M | Met | Nonpolar (hydrophobic) |
| Asparagine | N | Asn | Polar (hydrophilic, neutral) |
| Proline | P | Pro | Nonpolar |
| Glutamine | Q | Gln | Polar (hydrophilic, neutral) |
| Arginine | R | Arg | Polar (hydrophilic, basic) |
| Serine | S | Ser | Polar |
| Threonine | T | Thr | Polar |
| Valine | V | Val | Nonpolar (hydrophobic) |
| Tryptophan | W | Trp | Nonpolar |
| Tyrosine | Y | Tyr | Polar |

ing methionine. The mitochondrial genetic code for vertebrates does not necessarily apply to nonvertebrate organisms. In fact, ascidian, echinoderms, *Drosophila,* yeast, plants, and protozoans are known to have slightly different genetic codes, as shown in Table 1.4. The genetic codes of nuclear genes of ciliated protozoans such as *Tetrahymena* and *Par-*

Table 1.3 Vertebrate mitochondrial genetic code. Differences from the standard genetic code are shown in boldface.

| Codon | | Codon | | Codon | | Codon | |
|---|---|---|---|---|---|---|---|
| UUU | Phe | UCU | Ser | UAU | Tyr | UGU | Cys |
| UUC | Phe | UCC | Ser | UAC | Tyr | UGC | Cys |
| UUA | Leu | UCA | Ser | UAA | Ter | UGA | **Trp** |
| UUG | Leu | UCG | Ser | UAG | Ter | UGG | Trp |
| CUU | Leu | CCU | Pro | CAU | His | CGU | Arg |
| CUC | Leu | CCC | Pro | CAC | His | CGC | Arg |
| CUA | Leu | CCA | Pro | CAA | Gln | CGA | Arg |
| CUG | Leu | CCG | Pro | CAG | Gln | CGG | Arg |
| AUU | Ile | ACU | Thr | AAU | Asn | AGU | Ser |
| AUC | Ile | ACC | Thr | AAC | Asn | AGC | Ser |
| AUA | **Met** | ACA | Thr | AAA | Lys | AGA | **Ter** |
| AUG | Met | ACG | Thr | AAG | Lys | AGG | **Ter** |
| GUU | Val | GCU | Ala | GAU | Asp | GGU | Gly |
| GUC | Val | GCC | Ala | GAC | Asp | GGC | Gly |
| GUA | Val | GCA | Ala | GAA | Glu | GGA | Gly |
| GUG | Val | GCG | Ala | GAG | Glu | GGG | Gly |

Table 1.4 Some other genetic codes that differ from the standard code.

| | Codons | | | | | | |
|---|---|---|---|---|---|---|---|
| Organelle/Organisms | UGA | AUA | AAA | AGR | CUN | CGG | UAR |
| Standard genetic code | Ter | Ile | Lys | Arg | Leu | Arg | Ter |
| Mitochondrial code | | | | | | | |
| Vertebrate | Trp | Met | • | Ter | • | • | • |
| Ascidian | Trp | Met | • | Gly | • | • | • |
| Echinoderm | Trp | • | Asn | Ser | • | • | • |
| *Drosophila* | Trp | Met | • | Ser | • | • | • |
| Yeast | Trp | Met | • | • | Thr | • | • |
| Protozoan | Trp | • | • | • | • | • | • |
| Mold | Trp | • | • | • | • | • | • |
| Coelenterate | Trp | • | • | • | • | • | • |
| Nuclear code | | | | | | | |
| *Tetrahymena* | • | • | • | • | • | • | Gln |
| *Mycoplasma* | Trp | • | • | • | • | • | • |
| Euplotid | Cys | • | • | • | • | • | • |

*Note:* • Indicates identity with the standard code. R = A or G and N = T, C, A, or G.

*amecium* are also slightly different from the standard genetic code. Here, UAA and UAG do not appear to be termination codons even in nuclear genes but code for glutamine. Furthermore, in the prokaryotic organism *Mycoplasma capricolum,* the usual termination codon UGA is used for encoding tryptophan (see Osawa 1995).

In plant mitochondrial genes, codon CGG is not directly translated into tryptophan, but the nucleotide C in this codon is converted to U after the mRNA is formed, and this converted codon UGG encodes tryptophan using the standard genetic code. This process is called **RNA editing** (Covello and Gray 1993). In the comparison of amino acid sequences from different plant species, however, one can treat CGG as though it were a tryptophan codon. Actually, RNA editing occurs in some mitochondrial genes of the other eukaryotic kingdoms as well, and one should be cautious in translating DNA sequences into amino acid sequences in these genes.

## 1.4. Mutational Changes of DNA Sequences

Since all morphological and physiological characters of organisms are ultimately controlled by the genetic information carried by DNA, any mutational changes in these characters are due to some change in DNA molecules. There are four basic types of changes in DNA. They are **substitution** of a nucleotide for another nucleotide (Figure 1.2A), **deletion** of nucleotides (Figure 1.2B), **insertion** of nucleotides (Figure 1.2C), and **inversion** of nucleotides (Figure 1.2D). Insertion, deletion, and inversion may occur with one or more nucleotides as a unit. If insertions or deletions occur in a protein coding gene, they may shift the reading frame of the nucleotide sequence. These insertions and deletions are called **frameshift mutations.**

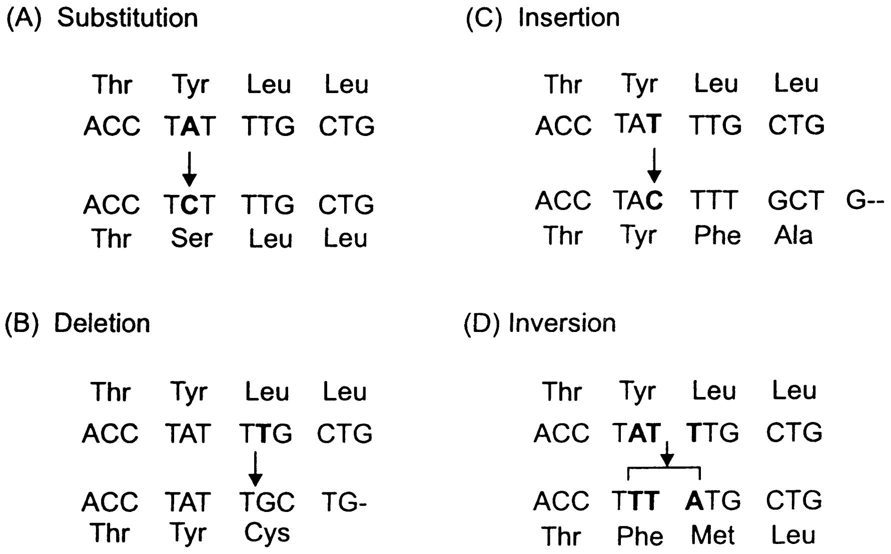

FIGURE 1.2. Four basic types of mutation at the nucleotide level. Nucleotide sequences are presented in units of codons or nucleotide triplets in order to show how the amino acids encoded are affected by the nucleotide changes. The nucleotides affected by the mutational changes are shown in boldface.

Nucleotide substitutions can be divided into two classes: **transitions** and **transversions.** A transition is the substitution of a **purine** (adenine or guanine) for another purine or the substitution of a **pyrimidine** (thymine or cytosine) for another pyrimidine (Figure 1.3). Other types of nucleotide substitutions are called transversions. In most DNA segments, transitional nucleotide substitutions are known to occur more frequently than transversions (e.g., Fitch 1967; Gojobori et al. 1982; Kocher and Wilson 1991). In the case of protein-coding genes, nucleotide substitutions that result in synonymous codons are called **synonymous** or **silent**

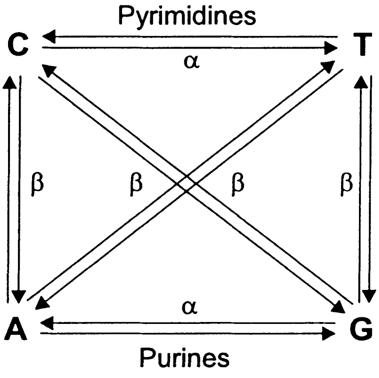

FIGURE 1.3. Transitional (A↔G and T↔C) and transversional (others) nucleotide substitutions. α and β are the rates of transitional and transversional substitutions, respectively.

**substitutions,** whereas those that result in nonsynonymous codons are called **nonsynonymous** or **amino acid replacement substitutions.** In addition, there are mutations that result in stop codons, and they are called **nonsense mutations.**

Because of the properties of the genetic codes, most synonymous substitutions occur at the third nucleotide position of codons, but some occur at the first position. All nucleotide substitutions at the second position are either nonsynonymous or nonsense mutations. If we assume that all codons are equally frequent in the genome and the probability of substitution is the same for all pairs of nucleotides, the proportions of synonymous, nonsynonymous, and nonsense mutations are about 25, 71, and 4%, respectively (Nei 1975; Li 1997). In practice, of course, the assumption of equal frequencies of codons and random nucleotide substitution is not realistic, but these percentages give a rough idea of the relative frequencies of different mutations at the nucleotide level.

Recent data indicate that insertions or deletions occur quite often, particularly in noncoding regions of DNA. The number of nucleotides involved in an insertion or deletion event varies from a few nucleotides to a large block of DNA. Short insertions or deletions are apparently caused by errors in DNA replication. Long insertions or deletions seem to be mainly due to unequal crossover or DNA transposition. DNA transposition, that is, the movement of a DNA segment from one chromosomal position to another, may occur by the aid of **transposons** or **transposable elements.** Transposable elements are known to cause various sorts of mutational changes (e.g., Kidwell and Lisch 1997). Another possible mechanism of gene insertion is **horizontal gene transfer** between species that is apparently mediated by transposable elements.

The possible role of unequal crossover in increasing the number of genes in the genome has been known for many years (Bridges 1936; Stephens 1951). However, only after the initiation of molecular study of DNA has it been realized that it plays an important role in evolution in increasing or decreasing DNA content (Ohno 1967, 1970; Nei 1969). Particularly in multigene families such as immunoglobulin and ribosomal RNA genes, unequal crossover seems to have played an important role in generating multiple copies of genes. A genetic event related to unequal crossover is **gene conversion.** Gene conversion is the alteration of a segment of DNA that makes the segment identical with another segment of DNA. This event is believed to occur by the repair of mismatched bases in heteroduplex DNA (Radding 1982) and is capable of homogenizing the member genes of a multigene family, but it does not change the number of gene copies.

## 1.5. Codon Usage

If nucleotide substitution occurs at random at each nucleotide site, every nucleotide site is expected to have one of the four nucleotides, A, T, C, and G, with equal probability. Therefore, if there is no selection and no mutational bias, one would expect that the codons encoding the same amino acid are on average in equal frequencies in protein coding regions

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Phe UUU | 15 (0.51) | Ser UCU | 32 (1.86) | Tyr UAU | 18 (0.64) | Cys UGU | 5 (1.00) |
| **UUC** | 44 (1.49) | UCC | 38 (2.21) | **UAC** | 38 (1.36) | UGC | 5 (1.00) |
| Leu UUA | 2 (0.07) | UCA | 2 (0.12) | Ter UAA | | Ter UGA | |
| UUG | 8 (0.27) | UCG | 5 (0.29) | Ter UAG | | Trp UGG | 8 (1.00) |
| Leu CUU | 11 (0.36) | Pro CCU | 9 (0.48) | His CAU | 5 (0.36) | Arg **CGU** | 89 (3.93) |
| CUC | 18 (0.60) | CCC | 0 (0.00) | CAC | 23 (1.64) | **CGC** | 46 (2.03) |
| CUA | 1 (0.03) | CCA | 11 (0.59) | Gln CAA | 15 (0.34) | CGA | 1 (0.04) |
| **CUG** | 141 (4.67) | **CCG** | 55 (2.93) | **CAG** | 73 (1.66) | CGG | 0 (0.00) |
| Ile AUU | 29 (0.69) | Thr **ACU** | 19 (0.78) | Asn AAU | 4 (0.11) | Ser AGU | 3 (0.17) |
| **AUC** | 98 (2.31) | **ACC** | 63 (2.57) | **AAC** | 66 (1.89) | AGC | 23 (1.34) |
| AUA | 0 (0.00) | ACA | 3 (0.12) | Lys **AAA** | 77 (1.35) | Arg AGA | 0 (0.00) |
| Met AUG | 60 (1.00) | ACG | 13 (0.53) | AAG | 37 (0.65) | AGG | 0 (0.00) |
| Val **GUU** | 55 (1.53) | Ala GCU | 30 (0.94) | Asp GAU | 60 (0.83) | Gly **GGU** | 78 (2.40) |
| GUC | 21 (0.58) | GCC | 19 (0.59) | GAC | 85 (1.17) | GGC | 47 (1.45) |
| **GUA** | 34 (0.94) | **GCA** | 30 (0.94) | Glu **GAA** | 147 (1.52) | GGA | 0 (0.00) |
| **GUG** | 34 (0.94) | **GCG** | 49 (1.53) | GAG | 46 (0.48) | GGG | 5 (0.15) |

FIGURE 1.4. Codon frequencies observed in the RNA polymerase genes (rpo B and D genes) of the bacterium *Escherichia coli.* The codons optimal for the translational system are shown in boldface. Relative synonymous codon usages (RSCU) given in the parentheses were computed by Equation (1.1). Data from Ikemura (1985).

of DNA. For example, amino acid valine (Val) is encoded by four codons, GUU, GUC, GUA, and GUG. So, if we examine a large number of Val codons in a gene, the relative frequencies of GUU, GUC, GUA, and GUG are all expected to be nearly equal to 25%.

In practice, the frequencies of different codons for the same amino acid are usually different, and some codons are used more often than others. Figure 1.4 shows the frequencies of use of each codon (number of times used) in the RNA polymerase of the bacterium *Escherichia coli (E. coli)* (Ikemura 1985). In the case of amino acid valine, the four codons are used nearly equally, though the use of GUU is more than two times higher than that of GUC. In arginine, however, codons CGU and CGC are used almost exclusively, and codons CGA, CGG, AGA, and AGG are almost never used. This type of **codon usage bias** is generally observed in both prokaryotic and eukaryotic genes.

What causes the codon usage bias? There are several factors. First, Ikemura (1981, 1985) showed that in *E. coli* and yeast, the frequency of codon usage in highly expressed genes is correlated with the relative abundance of the isoaccepting tRNAs in the cell. In other words, the tRNAs that correspond to frequently used codons are more abundant than those corresponding to rarely used codons. For example, in the case of arginine codons, CGU and CGC are often used, because the tRNAs corresponding to these two codons are more abundant than the tRNAs corresponding to the other codons. This suggests that the translationary machinery tends to use abundant tRNAs to produce proteins. Ikemura (1985) showed that the abundance of a particular tRNA is correlated with the number of copies of the gene that encodes the tRNA. Therefore, as far as highly expressed genes are concerned, the codon usage bias is essentially the same for all genes in the same organism. The codon usage bias

Table 1.5 Relative synonymous codon usage (*RSCU*) in bacteria (*E. coli*), yeast (*S. cerevisiae*), fruit fly (*D. melanogaster*), and human.

| Amino | | Bacteria | | Yeast | | Fruit fly | | Human | |
|---|---|---|---|---|---|---|---|---|---|
| Acid | Codon | High[a] | Low[b] | High | Low | High | Low | G+C[c] | A+T[d] |
| Leu | UUA | 0.06 | 1.24 | 0.49 | 1.49 | 0.03 | 0.62 | 0.05 | 0.99 |
| | UUG | 0.07 | 0.87 | **5.34** | **1.48** | 0.69 | 1.05 | 0.31 | 1.01 |
| | CUU | 0.13 | 0.72 | 0.02 | 0.73 | 0.25 | 0.80 | 0.20 | 1.26 |
| | CUC | 0.17 | 0.65 | 0.00 | 0.51 | 0.72 | 0.90 | 1.42 | 0.80 |
| | CUA | 0.04 | 0.31 | 0.15 | 0.95 | 0.06 | 0.60 | 0.15 | 0.57 |
| | CUG | **5.54** | **2.20** | 0.02 | 0.84 | **4.25** | **2.04** | **3.88** | **1.38** |
| Val | GUU | **2.41** | **1.09** | **2.07** | **1.13** | 0.56 | 0.74 | 0.09 | 1.32 |
| | GUC | 0.08 | 0.99 | 1.91 | 0.76 | 1.59 | 0.93 | 1.03 | 0.69 |
| | GUA | 1.12 | 1.63 | 0.00 | 1.18 | 0.06 | 0.53 | 0.11 | 0.80 |
| | GUG | 0.40 | 1.29 | 0.02 | 0.93 | **1.79** | **1.80** | **2.78** | **1.19** |
| Ile | AUU | 0.48 | 1.38 | 1.26 | 1.29 | 0.74 | 1.27 | 0.45 | 1.60 |
| | AUC | **2.51** | **1.12** | **1.74** | **0.66** | **2.26** | **0.95** | **2.43** | **0.76** |
| | AUA | 0.01 | 0.50 | 0.00 | 1.05 | 0.00 | 0.78 | 0.12 | 0.64 |
| Phe | UUU | 0.34 | 1.33 | 0.19 | 1.38 | 0.12 | 0.86 | 0.27 | 1.20 |
| | UUC | **1.66** | **0.67** | **1.81** | **0.62** | **1.88** | **1.14** | **1.73** | **0.80** |

*Source:* Modified from Sharp et al. (1988).
*Note:* Codons with the highest usage in the *High* genes and their corresponding usages in the *Low* genes are shown in boldface.

[a] *High* denotes genes with high levels of gene expression.
[b] *Low* denotes genes with low levels of gene expression.
[c] In humans, "G + C" refers to genes in GC-rich regions.
[d] In humans, "A + T" refers to genes in AT-rich regions.

in yeast is known to be quite different from that in *E. coli* (Table 1.5), but this bias can also be explained by the relative contents of tRNAs.

However, the above rule does not necessarily apply to moderately expressed genes (e.g., Thr and Trp synthetase genes in *E. coli*). In these genes, the codon usage tends to be more even for all codons that encode the same amino acid. The reason for this seems to be that the translation does not occur quickly, so that rare isoaccepting tRNAs can be used.

The above observations suggest that nucleotides mutate more or less at random, but the codons that do not correspond to abundant tRNAs are eliminated by **purifying selection** in highly expressed genes, because they are inefficient in protein synthesis. In moderately expressed genes, the selection pressure is apparently so low that many different codons are used. This pattern of natural selection is observed in many single-cell organisms or even in the fruit fly *Drosophila melanogaster* (Table 1.5). However, this rule does not seem to apply to human genes (Sharp et al. 1988).

Although the relative abundance of isoaccepting tRNAs is an important factor, there is another factor that affects the codon usage; it is the **biased mutation pressure.** In bacteria, the relative frequency of nucleotides G and C (GC content) in the genome is known to vary from about 25 to 75% (Osawa 1995), and this variation is believed to be largely due to the difference between the forward and backward mutation rates of the

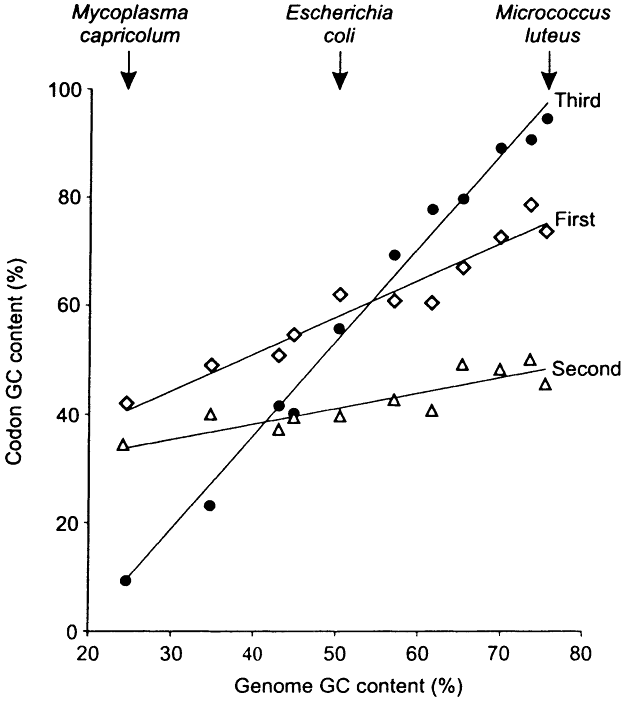

Figure 1.5. Relationships of the total genome GC content and the GC contents of first, second, and third nucleotide positions of genes from 11 different bacterial species, of which three species names are given. Modified from Muto and Osawa (1987).

GC and the AT pairs in the nucleotide sequence (Sueoka 1962). In some bacteria (e.g., *Mycoplasma capricolum*), the mutation pressure from GC to AT is so high that the nucleotides at silent third codon positions are almost always A or T (Muto and Osawa 1987). In some other bacteria (e.g., *Micrococcus luteus*), mutation pressure occurs in the opposite direction (AT → GC), so that the most often used nucleotide at third codon positions is either G or C (Figure 1.5).

Of course, for a protein to maintain its function, even the GC content at third positions is expected to be different from the equilibrium frequency determined by the mutation pressure alone, because some nucleotide substitutions at third positions result in amino acid changes and thus would be subject to purifying selection. Nucleotide substitutions at second positions are all nonsynonymous, so they are primarily controlled by functional constraints rather than by mutation pressure, whereas a small proportion of substitutions at first positions are synonymous, so that the effect of mutation pressure is expected to be intermediate between the effects at third and second codon positions.

Figure 1.5 shows the relationships between the GC content at first, second, and third codon positions of genes and the genome GC content in 11 different bacterial species covering a broad range of genome GC content. At third codon positions, the GC content of genes is nearly equal to the genome GC content, suggesting that the effect of mutation pressure is very strong. At second positions, however, the slope of the linear relationship with the genome GC content is much lower than that at third positions. This suggests that the effect of mutation pressure is less important at second positions and that the GC content is determined largely by purifying selection due to the functional constraints of the genes, as mentioned above. As expected, the slope of the relationship between the gene and genome GC contents for first positions is intermediate between the slopes for third and second positions. These observations support the view that the codon usage is controlled by both mutation pressure and purifying selection.

The fact that the GC content varies extensively among different groups of bacteria indicates that the pattern of nucleotide substitution is not the same for all groups of bacteria. This introduces complications in the study of phylogenetic relationships of these organisms (Galtier and Gouy 1995, 1998). The different groups of bacteria considered here probably diverged more than one billion years ago, so one might think that this problem is not important for the study of evolution of higher organisms, which evolved more recently. In practice, however, there is evidence that even in a relatively shorter period of evolutionary time the pattern of nucleotide substitution may change (Moriyama and Powell 1998).

In contrast to single-cell organisms, animals and plants are known to have a narrow range of GC content when the total genome is considered (Sueoka 1962). In particular, vertebrate animals all have a GC content of 40–45%. However, codon usage bias is still observed in many genes of higher organisms. In some invertebrates such as *Drosophila* the bias is quite strong, and this bias is apparently caused by the relative abundance of isoaccepting tRNAs, as in the case of microorganisms (Shields et al. 1988; Akashi 1994; Moriyama and Powell 1997).

In vertebrates, this issue is somewhat complicated, because gene expression is tissue-dependent and the genome is heterogeneous in terms of GC content. Bernardi et al. (1985, 1988) have shown that the vertebrate genomes are a mosaic composed of GC-rich regions and GC-poor regions and that some GC-rich regions have a GC content of about 60% and some GC-poor regions about 30%. Each of these GC-rich or GC-poor regions may be as long as 300 kb and contain many functional genes. These GC-rich or GC-poor regions are called **isochores.** Interestingly, the GC content at third codon positions of genes within an isochore is generally close to the GC content of the entire isochore. There are four major groups of isochores (two GC-rich and two GC-poor isochores) in warm-blooded vertebrates such as mammals and birds, but in cold-blooded vertebrates GC-rich isochores are rare or nearly absent. The boundary between the GC-rich and GC-poor isochores is known to be quite narrow (Ikemura and Aota 1988).

The origin of isochores in vertebrates has been a subject of controversy, and no consensus has been reached at the present time. The reader who is interested in this subject may refer to Wolfe et al. (1989), Holmquist and Filipski (1994), and Bernardi (1995). However, it is important to note that the genes located in different isochores are expected to have different patterns of codon usage biases, and since codon usage bias affects the rate of nucleotide substitution (Shields et al. 1988; Sharp et al. 1989), they may evolve at different rates.

### *Statistical Measures of Codon Usage Bias*

The numbers of occurrence (frequencies) of different codons for a given amino acid as those shown in Figure 1.4 clearly indicate the extent of codon usage bias if there is any. However, the absolute frequencies of codons are not convenient for comparing the bias for different genes or for different organisms, because the total number of codons examined is not necessarily the same. In this case, a more useful measure of codon usage bias is the **relative synonymous codon usage** (*RSCU*), which is defined as the observed frequency of a codon divided by the expected frequency under the assumption of equal codon usage (Sharp et al. 1986). For a given amino acid, *RSCU* is given by

$$RSCU = X_i/\bar{X} \tag{1.1}$$

where $X_i$ is the observed number of the *i*-th codon for the amino acid, and $\bar{X}$ is the average of $X_i$ over all codons, that is, $\bar{X} = \Sigma_i X_i/m$, where $m$ is the number of different codons for the amino acid.

As an example, let us consider the codon usage for the RNA polymerase genes, *rpo B* and *D,* of *E. coli* in Figure 1.4. For amino acid proline (Pro) there are four codons ($m = 4$), the observed number of a codon ($X_i$) varies from 0 to 55, and the average frequency becomes $\Sigma_i X_i/m = 75/4 = 18.75$. Therefore, *RSCU* is 0.48 for codon CCU, 0 for CCC, 0.59 for CCA, and 2.93 for CCG. *RSCU*s for other codons can be obtained in the same way, and they are given in Figure 1.4. These values can now be used for comparing the codon usage patterns for different genes. Table 1.5 shows the *RSCU*s for four amino acids for the genes from bacteria, yeasts, fruit flies, and humans. There is conspicuous codon usage bias for all organisms, but the pattern of the bias varies considerably with organism.

A number of authors have proposed statistical methods for measuring the extent of codon bias for the entire sequence of a gene or a genome. The measures include the **codon adaptation index** (Sharp and Li 1987), **scaled $\chi^2$ measure** (Shields et al. 1988), and **effective number of codons** (Wright 1990). Each measure has advantages and disadvantages, but at this moment it is not clear which method is most useful (Comeron and Aguade 1998).