1

Basic concepts of molecular evolution

Anne-Mieke Vandamme

1.1 Genetic information

The phenotype of living organisms is always a result of the genetic information that they carry and pass on to the next generation and of the interaction with the environment. The genome, carrier of this genetic information, is in most organisms deoxyribonucleic acid (***DNA***), whereas some viruses have a ribonucleic acid (***RNA***) genome. Part of the genetic information in DNA is transcribed into RNA, either mRNA, which acts as a template for ***protein*** synthesis; rRNA, which together with ribosomal proteins constitutes the protein translation machinery; or tRNA, which offers the encoded ***amino acid***. The genomic DNA also contains elements, such as ***promotors*** and ***enhancers***, that orchestrate the proper transcription into RNA. A large part of the genomic DNA of eukaryotes consists of genetic elements, such as introns, alu-repeats, the function of which is still not entirely clear. Proteins, RNA, and to some extent DNA, through their interaction with the environment, constitute the phenotype of an organism.

DNA is a double helix in which the two ***polynucleotide*** strands are antiparallelly oriented, whereas RNA is a single-stranded polynucleotide. The backbone in each DNA strand consists of deoxyriboses with a phosphodiester linking each 5′ carbon with the 3′ carbon of the next sugar. In RNA, the sugar moiety is ribose. On each sugar, one of the following four bases is linked to the 1′ carbon in DNA: the ***purines***, ***adenine*** (***A***) or ***guanine*** (***G***); or the ***pyrimidines***, ***thymine*** (***T***), or ***cytosine*** (***C***); in RNA, thymine is replaced by ***uracil*** (***U***). Hydrogen bonds and base stacking result in binding of the two DNA strands, with strong (triple) bonds between G and C, and weak (double) bonds between T/U and A (Figure 1.1). These hydrogen-bonded pairs are called ***complementary***. During DNA duplication or RNA transcription, DNA or RNA polymerase synthesizes a complementary 5′–3′ strand starting with the lower 3′–5′ DNA strand as template, such that the genetic information is preserved. This genetic information is represented by a one-letter code, indicating the 5′–3′ sequential order of the bases in the DNA or RNA

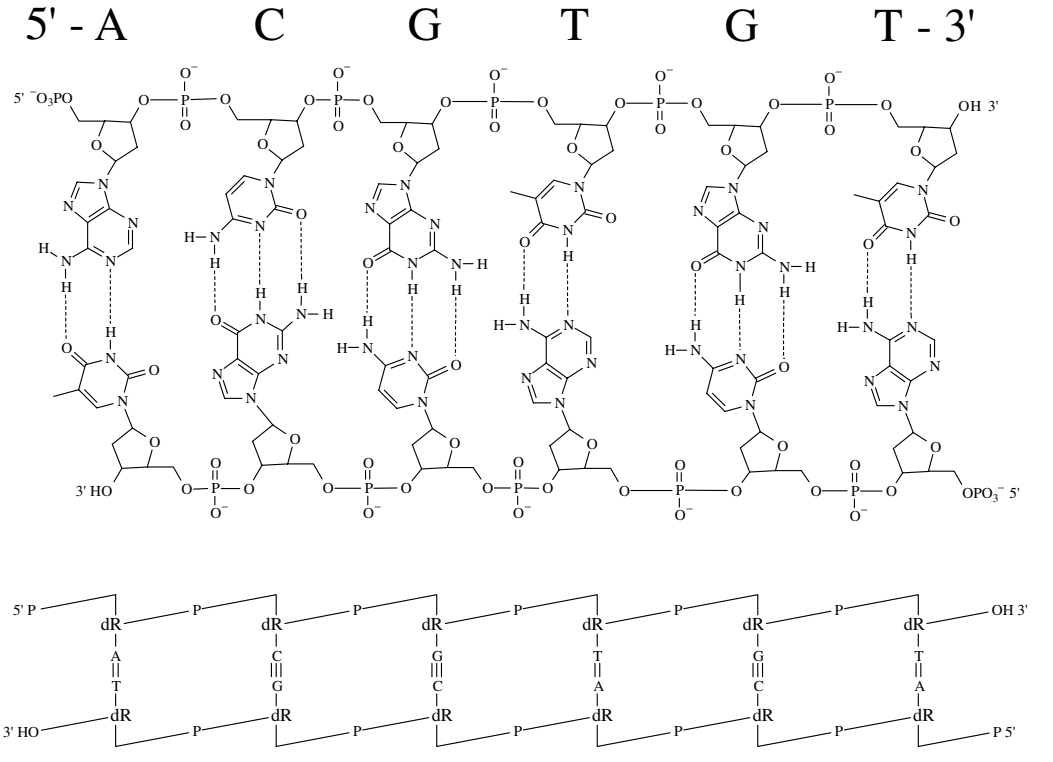

Figure 1.1 Chemical structure of double-stranded DNA. The chemical moieties are indicated as follows: dR, deoxyribose; P, phosphate; G, guanine; T, thymine; A, adenine; and C, cytosine. The strand orientation is represented in a standard way: in the upper strand 5′–3′, indicating that the chain starts at the 5′ carbon of the first dR, and ends at the 3′ carbon of the last dR. The one-letter code of the corresponding genetic information is given on top, and only takes into account the 5′–3′ upper strand. (Courtesy of Christophe Pannecouqe.)

(Figure 1.1). A nucleotide sequence is thus represented by a contiguous stretch of the four letters A, G, C, and T/U.

In the RNA strands that encode a protein, each triplet of bases is recognized by the ribosomes as a code for a specific amino acid. This translation results in polymerization of the encoded amino acids into a protein. Amino acids can be represented by a three- or one-letter abbreviation (Table 1.1). An amino-acid sequence is represented by a contiguous stretch of the 21 letters of the one-letter amino-acid abbreviation.

The ***genetic code*** is universal for all organisms, with only a few exceptions such as the mitochondrial code, and is usually represented as an RNA code because the RNA is the direct template for protein synthesis (Table 1.2). The corresponding DNA code can be easily reconstructed by replacing the U with a T. Each position of the triplet code can be one of four bases; hence, 4^3 or 64 possible triplets encode 20 amino

Table 1.1 Three- and one-letter abbreviations of 20 naturally encoded amino acids

Amino acid	Three-letter abbreviation	One-letter abbreviation
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

acids (61 ***sense*** codes) and 3 stop codons (3 ***non-sense*** codes). The genetic code is said to be degenerated, or redundant, because all amino acids except methionine have more than one possible code. The first codon for methionine ***downstream*** (or 3′) of the ribosome entry site also acts as the start codon for the translation of a protein. As a result of the triplet code, each contiguous nucleotide stretch has three reading frames in the 5′–3′ direction. The complementary strand encodes three other reading frames. A reading frame that is able to encode a protein starts with a codon for methionine and ends with a stop codon. These reading frames are called ***open reading frames*** (***ORFs***).

During duplication of the genetic information, the DNA or RNA polymerase can occasionally incorporate a noncomplementary nucleotide. In addition, bases in a DNA strand can be chemically modified due to environmental factors such as UV light or chemical substances. These modified bases can potentially interfere with the synthesis of the complementary strand and thereby also result in a nucleotide incorporation that is not complementary to the original nucleotide. When these changes escape the cellular repair mechanisms, the genetic information is altered, resulting in what is called a ***point mutation***. The genetic code has evolved in such a

Table 1.2 Universal codon table

Codon	Amino acid[a]	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
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

[a] Amino acids are indicated by three-letter codes as indicated in Table 1.1.

way that a point mutation at the 3rd position rarely results in an amino-acid change (only in 30% of possible changes). A change at the second position always, and at the 1st position usually (96%), results in an amino-acid change. Mutations that do not result in amino-acid changes are called ***silent*** or ***synonymous mutations***. When a mutation results in the incorporation of a different amino acid, it is called nonsilent or ***nonsynonymous***. A site within a coding triplet is said to be ***fourfold degenerate*** when all possible changes at that site are synonymous; ***twofold degenerate*** when only two different amino acids are encoded by the four possible nucleotides at that position; and ***nondegenerate*** when all possible changes alter the encoded amino acid.

Incorporation errors replacing a purine with a purine and a pyrimidine with a pyrimidine are for steric reasons more easily made. The resulting mutations are called ***transitions***. ***Transversions***, purine to pyrimidine changes and the reverse, are less likely. When resulting in an amino-acid change, transversions often have a larger impact on the protein than transitions. There are four possible transition errors (A $\rightleftarrows$ G, C $\rightleftarrows$ T) and eight possible transversion errors (A $\rightleftarrows$ C, A $\rightleftarrows$ T, G $\rightleftarrows$ C, G $\rightleftarrows$ T); therefore, if a mutation would occur randomly, a transversion would be two times more likely than a transition. However, in many genes, transitions are twice as more likely to occur than transversions, which is used as default substitution

parameter in substitution models that can score transitions and transversions differently (see also Chapter 4).

Single nucleotide changes in a single codon often result in an amino acid with similar properties (e.g., hydrophobic), such that the tertiary structure of the encoded protein is not altered dramatically. Living organisms can therefore tolerate a limited number of nucleotide point mutations in their coding regions. Point mutations in noncoding regions are subject to other constraints, such as conservation of binding places for proteins or conservation of base pairing in RNA tertiary structures.

Errors in duplication of genetic information can also result in the ***deletion*** or ***insertion*** of one or more nucleotides, called ***indels***. When multiples of three nucleotides are inserted or deleted in coding regions, the reading frame remains intact, but one or more amino acids are inserted or deleted. When a single nucleotide or two nucleotides are inserted or deleted, the reading frame is disturbed and the resulting gene generally codes for an entirely different protein, with different amino acids and a different length than the original gene. Insertions or deletions are therefore rare in coding regions, but rather frequent in noncoding regions. When occurring in coding regions, indels can occasionally change the reading frame of a gene and make another ORF of the same gene accessible. Such mutations can lead to acquisition of new gene functions. Viruses make extensive use of this possibility. They often encode several proteins from a single gene by using overlapping ORFs.

When parts of two different DNA strands are recombined into a single strand, the mutation is called a ***recombination***. Recombinations have major effects on the affected gene. ***Splicing*** is the most common form of recombination. Eukaryotic genes are encoded by coding gene fragments called ***exons***, which are separated from each other by ***introns***. Joining of the introns occurs in the nucleus at the pre-mRNA level in dedicated spliceosomes. Mutations can result in altered splicing patterns. These usually destroy the gene function, but can occasionally result in the acquisition of a new gene function. Viruses have again used these possibilities extensively. By alternative splicing, sometimes in combination with the use of different reading frames, viruses are able to encode multiple proteins by a single gene. For example, human immunodeficiency virus (HIV) is able to encode two additional regulatory proteins using part of the coding region of the *env* gene by alternative splicing and overlapping reading frames. Another common form of recombination happens during ***meiosis***, when recombination occurs between ***homologous chromosomes***, shuffling the ***alleles*** for the next generation. Consequently, recombination has a major contribution to evolution of diploid organisms. In general, these recombinations occur in-between genes. However, if they occur within genes, they have deleterious effects on the affected gene, although sometimes genes with entirely different functions can be created.

A special case of recombination is ***gene duplication***. Gene duplication results in genome enlargement and can involve a single gene, or large genome sections (e.g., chromosome duplication or ***aneuploidy***). They can be partial, involving only gene fragments, or complete, whereby entire genes are duplicated. Genes in which a partial duplication took place, such as domain duplication, can potentially have a greatly altered function. An entirely duplicated gene can evolve independently. After a long history of independent evolution, duplicated genes can eventually acquire a new function. Duplication events have played a major role in the evolution of species. For example, complex body plans were possible due to separate evolution of duplications of the homeobox genes (Carroll, 1995).

1.2 Population dynamics

Mutations in a gene that are passed on to the offspring and that coexist with the original gene result in ***polymorphisms***. At a polymorphic site, two or more variants of a gene circulate in the population simultaneously. Population geneticists deal with the dynamics of the frequency of these polymorphic sites over time. In population dynamics, the evolution of these frequencies is investigated on a small time scale, covering a number of generations. The location in the genome where two variants coexist is called the ***locus***. The different variants are each called an allele. Virus genomes are flexible to genetic changes; RNA viruses can contain many polymorphic sites simultaneously in a single population. HIV, for example, has no single genome, but consists of a swarm of variants called a ***quasispecies*** (Eigen and Biebricher, 1988; Domingo et al., 1997). This is due to the rapid and error-prone replication of RNA viruses. ***Diploid*** organisms always carry two alleles. When both alleles are identical, the organism is ***homozygous*** at that locus; when the organism carries two different alleles, it is ***heterozygous*** at that locus. Heterozygous positions are polymorphic.

Evolution is always a result of changes in ***allele frequencies***, also called ***gene frequencies***. Whereby some alleles are lost over time and other alleles increase their frequency to 100 percent, they become ***fixed*** in the population (Figure 1.2). For RNA viruses, this evolution is reflected in the frequency of a variant in the quasispecies distribution. The long-term evolution of a species results from the successive fixation of particular alleles, which reflects fixation of mutations. The rate at which these mutations are fixed in the population is called the ***evolutionary rate***, or ***fixation rate***, and it is usually expressed as number of nucleotide (or amino acid) changes per site per year. This rate is dependent on the ***mutation rate***, the rate at which mutations arise at the DNA level, usually expressed as number of nucleotide (or amino acid) changes per site per replication cycle, on the ***generation***

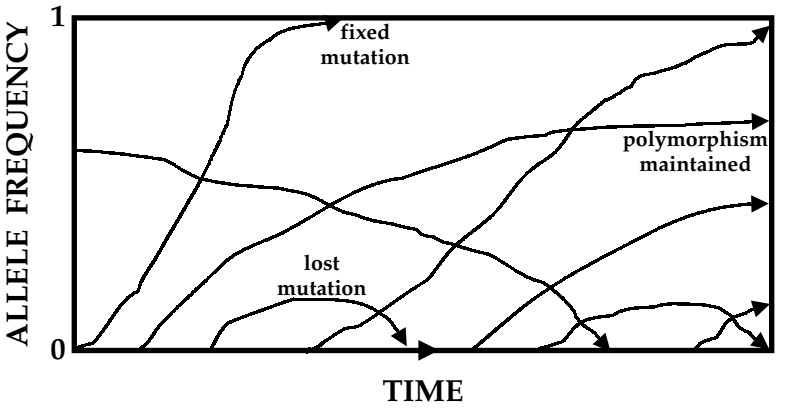

Figure 1.2 Loss or fixation of an allele in a population.

time, the time separating two generations, and on ***evolutionary forces***, such as the fitness of the organism carrying the allele or variant, positive and negative selective pressure, population size, genetic drift, reproductive potential, and competition of alleles.

If a particular allele is more fit than its polymorphic allele in a particular environment, it will be subjected to ***positive selective pressure***; if it is less fit, it will be subjected to ***negative selective pressure***. An allele can be less fit when it is homozygous, but have an advantage as heterozygote. In this case, polymorphism is advantageous and can be selected; this is called ***balancing selection***. For example, humans who carry the hemoglobin S allele on both chromosomes suffer from sickle-cell anemia, whereas heterozygotes are to some extent protected against malaria (Allison, 1956). Fitness of a variant is always the result of a particular phenotype of the organism; therefore, in coding regions, selective pressure always acts on mutations that alter function or stability of a gene or the amino-acid sequence encoded by the gene. Synonymous mutations could at first sight be expected to be neutral because they do not result in amino-acid changes. However, this is not always true. For example, synonymous changes can change RNA secondary structure and influence RNA stability; also, they result in the usage of a different tRNA, which may be less abundant. Still, most synonymous substitutions can be considered to be selectively neutral.

Whether a mutation becomes fixed through ***deterministic*** or ***stochastic*** forces depends on the ***effective population size*** (***Ne***) of the organism. This can be defined as the size of an ideal, randomly making population that has the same gene frequency changes as the population being studied. The effective population size can be smaller than the overall ***population size*** (***N***) because some members of a population may produce no offspring and there may be some level of inbreeding. It is the effective population size that determines the allele frequencies over time. When

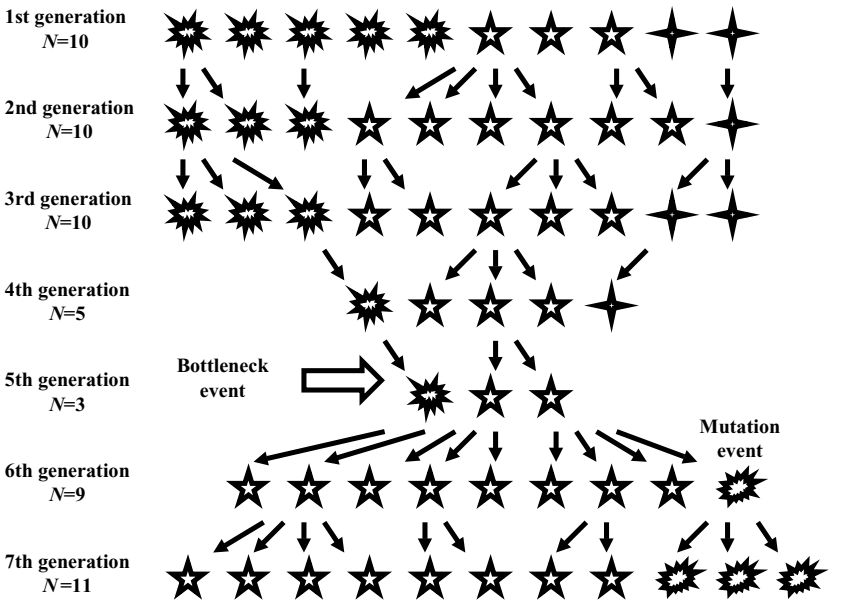

Figure 1.3 Population size (N), and the bottleneck effect.

the effective population size varies over multiple generations, the rates of evolution are notably influenced by generations with the smallest effective population sizes. This may be particularly true when population sizes are greatly reduced due to catastrophes or during migrations, etc. (Figure 1.3). These are called bottlenecks.

A ***deterministic model*** assumes that changes in allele frequencies or quasispecies distributions depend solely on the reproductive fitness of the variants in a particular environment and on the environmental conditions. In such a model, the gene frequencies can be predicted if the fitness and environmental conditions are known. In deterministic evolution, changes other than environmental conditions (e.g., chance events) do not influence allele frequencies or quasispecies distributions; therefore, this can only be true if the effective population size is infinite. ***Natural selection***, the effect of positive and negative selective pressure, accounts entirely for the changes in frequencies. When random fluctuations determine in part the allele frequencies, chance events play a role and allele frequencies or quasispecies distributions cannot be entirely predicted. In such a ***stochastic model***, one can only determine the probability of frequencies in the next generation. These probabilities still depend on the reproductive fitness of the variants in a particular environment and on the environmental conditions; however, in this case, chance events – due to the limited ***population size*** – also play a role. Allele frequencies or quasispecies distributions can only be predicted approximately. ***Random genetic drift***, therefore, contributes significantly to changes in frequencies under the stochastic model. The

smaller the effective population size, the larger the effect of chance events and the more the mutation rate is determined by ***genetic drift*** rather than by selective pressure.

Evolution is never entirely deterministic or entirely stochastic. Depending on the effective population size, allele frequencies and quasispecies distributions evolve more due to natural selection or random genetic drift. Although genetic changes are always random, an ***adaptive change*** under positive selective pressure will increase its frequency and become fixed after fewer generations than a neutral change, provided the effective population size is large enough. A mutation under negative selective pressure can become fixed due to random genetic drift when it is not entirely deleterious, but this requires more generations than for a neutral change. ***Nonsynonymous*** mutations result in a change in the phenotype of an organism, changing the interaction of that organism with its environment, and are thus subject to selective pressure. Synonymous substitutions, when not under constraints other than their coding potential, are neutral and therefore only become fixed due to random genetic drift. The effect of positive and negative selective pressure can be investigated by comparing the ***synonymous and nonsynonymous substitution rate*** (see also Chapter 11).

Darwin realized that the factors that shaped evolution were an environment with limited resources, inheritable variations among organisms that influenced fitness, competition between organisms, and natural selection. In his view, the survival of the fittest was the result of these factors and the major force behind the origin of species (Darwin, 1859). Only in the twentieth century, after the rediscovery of Mendelian laws, was it realized that the sources of variation on which selection could act were random mutations. In ***neo-Darwinism***, random mutations result in genetic variation, on which natural selection acts as the dominant force in evolution. Advantageous changes become fixed due to positive selective pressure, changes that result in a disadvantage are eliminated, and neutral changes result in polymorphisms that are maintained in a population. Changes in the environment can change the fate of neutral changes into advantageous or disadvantageous changes, resulting in subsequent fixation or elimination. Polymorphism also can be selected through balancing selection. Neo-Darwinism corresponds to a rather deterministic approach. In neo-Darwinism, a gene substitution is always the result of a positive adaptive process. The surviving organisms increase their fitness and become increasingly more adapted to the environment. This is called ***adaptive evolution***.

The ***neutral theory of evolution*** follows a more stochastic approach. Kimura (1983) advocated that the majority of gene substitutions were the result of random fixation of neutral or nearly neutral mutations. Positive selection does operate, but the effective population size is generally so small in comparison with the magnitude of the selective forces that the contribution of positive selection to evolution is too

weak to shape the genome. According to the neutral theory, only a small minority of mutations become fixed because of positive selection. Organisms are generally so well adapted to the environment that many nonsynonymous changes are deleterious and, therefore, quickly removed from the population by negative selection. Stochastic events predominate and substitutions, which are fixed mutations, are mainly the results of random genetic drift.

To what extent natural selection or neutral evolution acts on an organism or gene can be investigated with specific tools that are explained in detail in Chapter 11.

1.3 Data used for molecular phylogenetic analysis

To investigate the evolution and relationships among genes and organisms, different kinds of data can be used. The classical way of estimating the relationship between species is by comparing their morphological characters (Linnaeus, 1758). Taxonomy is still mainly based on morphology. The molecular information that is increasingly becoming available, such as nucleotide or amino-acid sequences and restriction fragment length polymorphism (RFLP), also can be used to infer phylogenetic relationships, based on the concepts of natural selection and neutral evolution. Whether the morphological or molecular approach is preferable for any particular evolutionary question has been hotly debated during the last *decennia* (Patterson, 1987). However, the use of molecular data for inferring ***phylogenetic trees*** has now gained considerable interest among biologists of different disciplines, and it is often used in addition to morphological data to study relationships in further detail. For extinct species, it is difficult or impossible to obtain molecular data, and using morphological characteristics of mummies or fossils is usually the only way to estimate their relationships. However, organisms such as viruses do not leave fossil records. The only way to study their past is through the phylogenetic relationships of existing viruses. In this book, we introduce the concepts, mathematics, and techniques to infer phylogenetic trees from molecular data and, in particular, from nucleotide and amino-acid sequences. Therefore, all applications described in this book restrict themselves to the use of sequence data.

According to the evolutionary theory, all organisms evolved from one common ancestor, going back to the origin of life. Different mechanisms of acquiring variation have led to today's biodiversity. These mechanisms include mutations, duplication of genes, reorganization of genomes, and recombination. Of all these sources, only mutations (i.e., point mutations, insertions, and deletions) are used by the different molecular phylogenetic methods to infer relationships between genes. To perform these evaluations, the similarity of the genes is considered, assuming that they are homologous (i.e., they share a common ancestor). Although it is assumed that all organisms share a common ancestor, over time the similarity in two

genes can be eroded such that the sequence data themselves do not carry enough information on the relationship between the two genes and they have accumulated too much variation. Therefore, the term ***homology*** is used only when the common ancestor is recent enough such that sequence information has retained enough similarity to be used in phylogenetic analysis. Thus, genes are either homologous or they are not. Consequently, there does not exist such an expression as 95% homology; rather, one should speak of 95% ***similarity***.

When two sequences are compared, one can always calculate the percentage similarity by counting the amount of identical nucleotides or amino acids, relative to the length of the sequence. This can be done even if the sequences are not homologous. DNA is composed of four different types of residues: A, G, C, and T. If gaps are *not* allowed, on average, 25% of the residues in two randomly chosen aligned sequences would be identical. If gaps *are* allowed, as much as 50% of the residues in two randomly chosen aligned sequences can be identical, resulting in a 50% similarity. For proteins, with 21 different types of codons (i.e., twenty amino acids and one terminator), it can be expected that two random protein sequences – after allowing gaps – can have up to 20% identical residues. In general, the higher the similarity, the more likely that the sequences are homologous.

Taxonomic comparisons show that the genes of closely related species usually only differ from one another by point mutations. These are usually found in the third (i.e., redundant) codon positions of ORFs such that the 3rd codon position has a faster evolutionary rate than the 1st and 2nd codon positions. The redundancy of the genetic code ensures that nucleotide sequences usually evolve more quickly than the proteins they encode. The sequences also may have a few inserted or deleted nucleotides (i.e., indels). Genes of more distantly related species differ by a greater number of changes of the same type. Some genes are conserved more than others, especially those parts encoding, for example, catalytic sites or the core of proteins. Other genes may have little or no similarity. Distantly related species often have discernible sequence relatedness only in the genes that encode enzymes or structural proteins. These similarities, when found, can be very distant and involve only short segments (i.e., motifs) interspersed with large regions with no similarity and of variable length, which indicates that many mutations and indels have occurred since they evolved from their common ancestor. Some of the proteins from distantly related species may have no significant sequence similarity but clearly have similar secondary and tertiary structures. Primary structure is lost more quickly than secondary and tertiary structure during evolutionary change. Thus, differences between closely related species are assessed most sensitively by analysis of their nucleotide sequences. More distant relationships, between families and genera, are best analyzed by comparing amino-acid sequences and may be revealed only by parts of some genes and their encoded proteins (see also Chapter 8).

```
VI557     TTAAATGCATGGGTAAAAGTAGTAGAAGAGAAGGCTTTTAGCCCAGAAGT 50
VI69      TTAAATGCATGGGTAAAGGTGATAGAAGAGAAGGCTTTTAGTCCAGAAGT 50
BZ162     TTAAATGCATGGGTAAAGGTGATAGAAGAGAAGGCTTTTAGCCCAGAAGT 50
VI313     TTGAATGCGTGGGTAAAAGTAATAGAGGAGAAGGCTTTCAGCCCGGAGGT 50
UG268     TTGAATGCATGGGTAAAAGTAATAGAGGAAAAGGCTTTCAGCCCAGAGGT 50
DJ259     TTGAATGCATGGGTAAAAGTAATAGAGGAGAAGGCTTTCAGCCCAGAAGT 50
K112      TTGAATGCATGGGTAAAAGTAATAGAAGAAAAGGCTTTCAGCCCAGAAGT 50
CI20      TTGAATGCATGGGTGAAGGTAATAGAGGAAAAGGCTTTCAGCCCAGAAGT 50
GAG46     TTAAATGCATGGGTAAAAGTAGTAGAAGAAAAGGCTTTCAGCCCAGAAGT 50
LAV       TTAAATGCATGGGTAAAAGTAGTAGAAGAGAAGGCTTTCAGCCCAGAAGT 50
HAN       TTAAATGCATGGGTAAAAGTAGTGGAAGAGAAGGCTTTCAGCCCAGAAGT 50
BZ121     TTAAATGCATGGGTCAAAGTAGTAGAAGAGAAGGCTTTCAGCCCAGAAGT 50
LBV217    TTAAATGCATGGGTAAAAGTAGTAGAAGAAAAGGCCTTCAGTCCAGAAGT 50
HIVBL     TTGAATGCATGGGTAAAAGTAGTAGAAGAAAAGGCCTTCAGTCCAGAAGT 50
VI191     TTGAATGCATGGGTAAAAGTAATAGAAGAAAAGAACTTCAGTCCAGAAGT 50
VI174     TTAAATGCATGGGTAAAGGTGATAGAAGAGAAAGCTTTTAGCCCAGAAGT 50
VI525     TTAAATGCATGGGTAAAAGTAGTAGAAGAAAAGGCTTTTAGCCCAGAAGT 50
Z2Z6      TTGAACGCATGGGTAAAAGTAATAGAAGAAAAGGCTTTCAGCCCAGAAGT 50
NDK       TTGAACGCATGGGTAAAAGTAATAGAAGAAAAGGCCTTCAGCCCGGAAGT 50
VI203     TTGAACGCATGGGTAAAAGTAATAGAGGAAAAGGCTTTCAATCCAGAAGT 50
```

Figure 1.4 Nucleotide sequence alignment of a fragment of the HIV-1 *gag* sequence. This alignment is part of the alignment used to draw the tree in Debyser et al. (1998).

Phylogenetic analysis estimates the relationship between genes or gene fragments by inferring the common history of the genes or gene fragments. To do this, it is essential that homologous sites be compared with each other (i.e., ***positional homology***). For this reason, the homologous sequences under investigation are aligned such that homologous sites form columns in the alignment (Figure 1.4). Obtaining the correct alignment is easy for closely related species and can even be done manually using a word processor. The more distantly related the sequences are, the trickier it is to find the best alignment. Therefore, alignments are usually constructed with specific software packages using particular algorithms. This topic is extensively discussed in Chapter 3.

Most algorithms start by comparing the sequence similarity of all sequence pairs, aligning first the two sequences with the highest similarity. The other sequences, in

order of similarity, are added progressively. The alignment continues in an iterative fashion, adding gaps where required to achieve positional homology, but gaps are introduced at the same position for all members of each growing cluster. Alignments obtained in this way are optimal for clusters of sequences, as there is no global optimization of the total alignment. When several gaps have been added to clusters of sequences, the total alignment often can be improved by manual editing. Obtaining a good alignment is one of the most crucial steps toward a good phylogenetic tree. When the sequence similarity is so low that an alignment becomes too ambiguous to be confident that homologous sites are aligned correctly, it is better to delete that particular gene fragment from the alignment so as not to distort the phylogenetic tree. Gaps at the beginning and end of a sequence, representing missing sequence data for the shorter sequences, have to be removed to consider equal amounts of data for all sequences. Often, columns in the middle of the sequence with deletions and insertions for the majority of the sequences are also removed from the analysis (see Chapter 3). The best alignment possible is the data that phylogenetic software packages use to construct phylogenetic trees.

For a reliable estimate of the phylogenetic relationship between genes, the entire gene under investigation must have the same history. Therefore, recombination events within the fragment under investigation, which distort this common history, also will distort a phylogenetic tree. Recombination outside the fragment of interest does not disturb the tree; however, knowledge of the recombination event is necessary when the two fragments are both investigated.

Genes originating from a duplication event recent enough to reveal their common ancestry at the nucleotide or amino-acid level are called ***paralogous***. Comparing such genes by phylogenetic analysis will result in information on the duplication event. Homologous genes in different species that have started a separate evolution because of the speciation are called ***orthologous***. Comparing such genes by phylogenetic analysis will result in information on the speciation event. Therefore, when performing phylogenetic analysis on homologous genes, it is important to know whether the genes are orthologous or paralogous. This prevents making the wrong conclusions on speciation events by comparing paralogous genes instead of orthologous genes.

Evolution of nonhomologous genes under similar selective pressures can result in ***parallel*** or ***convergent evolution***. When two enzymes evolve to have a similar function, the similar functional requirements can result in a similar active site consisting of the same or similar amino acids. This effect can result in the two sequences having higher than expected similarity, which can be mistaken for homology. Other events can result in a higher similarity of two sequences than the similarity expected from their evolutionary history. ***Sequence reversals*** occur when a mutation reverts back to the original nucleotide; ***multiple hits*** when a mutation has occurred several times at the same nucleotide, resulting in the same

nucleotide at homologous positions in two divergent sequences; and ***parallel substitutions*** when the same substitution happened in two different lineages. All these events disturb the linear relationship between the time of evolution and sequence divergence. This effect is called ***homoplasy*** (see Chapter 4).

Presently, sequence information is stored in databases such as the National Center for Biotechnology Information (NCBI), the National Library of Medicine (NLM), the European Molecular Biology Organization (EMBO), and the DNA Database of Japan (DDJ). A search for homologous sequences in individual databases can be done in various ways, based on scoring the similarity between sequences. Some organizations provide a search service via the international computer network (e.g., BLAST). However, no search method is perfect and related sequences may be missed. Information on search engines is provided in Chapter 2.

1.4 What is a phylogenetic tree?

Evolutionary relationships among genes and organisms can be elegantly illustrated by a phylogenetic tree, comparable to a pedigree showing which genes or organisms are most closely related. Phylogenetic trees are described this way because the various diagrams used for depicting these relationships resemble the structure of a tree (Figure 1.5), and the terms referring to the various parts of these diagrams (i.e., ***root***, ***stem***, ***branch***, ***node***, and ***leaf***) are also reminiscent of trees. ***External*** (***terminal***) ***nodes***, the ***extant*** (***existing***) ***taxa***, are often called ***operational taxonomic units*** (***OTUs***), a generic term that can represent many types of comparable ***taxa*** (e.g., a family of organisms, individuals, or virus strains of a single species; a set of related genes; or even gene regions). Similarly, ***internal nodes*** may be called ***hypothetical taxonomic units*** (***HTUs***) to emphasize that they are the hypothetical progenitors of OTUs. A group of taxa that belong to the same branch have a ***monophyletic*** origin and is called a ***cluster***. In Figure 1.5, the taxa A, B, and C form a cluster, have a common ancestor H, and, therefore, are of monophyletic origin. C, D, and E do not form a cluster without including additional strains; thus, they are not of monophyletic origin. The branching pattern – that is, the order of the nodes – is called the ***topology*** of the tree.

An ***unrooted*** tree only positions the individual taxa relative to each other without indicating the direction of the evolutionary process. In an unrooted tree, there is no indication of which node represents the ancestor of all OTUs. To indicate the direction of evolution in a tree, it must have a root that leads to the common ancestor of all the OTUs in it (see Figure 1.5). The tree can be ***rooted*** if one or more of the OTUs form an ***outgroup*** because they are known as, or are believed to be, the most distantly related of the OTUs (i.e., ***outgroup rooting***). The remainder then forms the ***ingroup***. The root node is the node that joins the ingroup and the

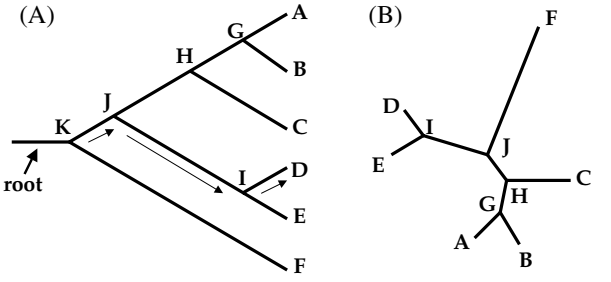

Figure 1.5 Structure of rooted (A) and unrooted (B) phylogenetic trees. Both trees have the same topology. A rooted tree is usually drawn with the root to the left. A, B, C, D, E, and F are external nodes or OTUs. G, H, I, J, and K are internal nodes or HTUs, with K as root node. The unrooted tree does not have a root node. The lines between the nodes are branches. The arrow indicates the direction of evolution in the rooted tree (e.g., from root K to external node D). The direction of evolution is not known in an unrooted tree.

outgroup; therefore, it must represent the common ancestor of both the outgroup and the ingroup. It is still possible to assign a root even when it is not known which OTU to use as the outgroup. Assuming that the rate of evolution in the different lineages is similar, the root will then lie either at the midpoint of the path joining the two most dissimilar OTUs, or at the mean point of the paths that join the most dissimilar OTUs connected through a single edge (i.e., ***midpoint rooting***).

When trying to root a tree, do not choose an outgroup that is distantly related to the ingroup taxa. This may result in serious topological errors because sites may have become saturated with multiple mutations, by which information may have been erased. Also, do not choose an outgroup that is too closely related to the taxa in question; in this case, it may not be a true outgroup. The use of more than one outgroup generally improves the estimate of tree topology. As noticed previously, midpoint rooting could be a good alternative when no outgroups are available, but only in case of approximately equal evolutionary rates over all branches of the tree.

Various styles are used to depict phylogenetic trees. Figure 1.6 demonstrates the same tree as in Figure 1.5, but in a different style. Branches at internal nodes can be rotated without altering the topology of a tree. Both trees in Figure 1.6 have identical topologies. Compared with tree (A), tree (B) was rotated at nodes J and H.

Phylogenetic trees illustrate the relationship among the sequences aligned; therefore, they are always ***gene trees***. Whether these gene trees can be interpreted as representing the relationship among species depends on whether the genes provided to the alignment are orthologous or paralogous genes. When the ancestral gene A is duplicated into A1 and A2 within the same species, then the relationship between A1 and A2 will give information on the duplication event. Suppose the speciation into species C and D – with C1 and D1 being the descendant of gene A1, and C2 and D2 descendant from A2 – occurs after the gene duplication. Comparing C1

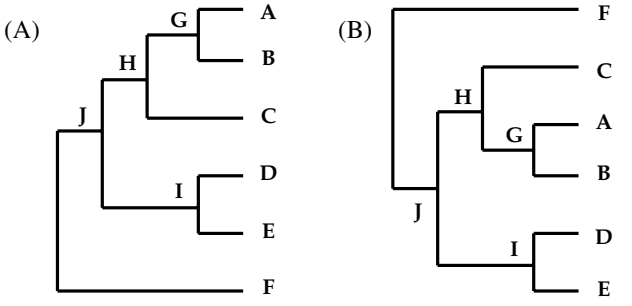

Figure 1.6 This is the same tree as in Figure 1.5, but in a different style. Both trees (A) and (B) have identical topologies, with some of the internal nodes rotated.

with D2 (or C2 with D1) will give information on the duplication event, whereas comparing C1 with D1 (or C2 with D2) will give information on the speciation event and the tree can be considered a ***species tree***. When dealing with a gene that has polymorphic sites in that species, the nodes in the gene tree never really indicate the speciation event. Because some sequence variation existed before speciation – represented by the different alleles (or quasispecies variation for RNA viruses) – the gene tree is a population tree and the nodes represent the separation of the different alleles, which precede speciation. Alternatively, some alleles may have become extinct after speciation and the separation of the different alleles may follow speciation.

The ***coalescence time*** is the time when the ***most recent common ancestor*** (***MRCA***) of the extant alleles still existed. When trying to acquire information on the origin of a species population by analyzing the sequence variability of different alleles of a particular gene, the coalescence time depends on the extinction of alleles after speciation. In Figure 1.3, individuals of the 7th generation have one common ancestor in the 4th generation; therefore, the coalescence time is later than the first generation. Thus, when the effective population size is small and alleles are lost after speciation, the coalescence time for the different alleles within a species in a population tree is later than the speciation time. For example, the coalescence time of human mitochondrial DNA, which is inherited through the female line, is calculated to be around two hundred thousand years ago (Vigilant et al., 1991; Ingman et al., 2000). The coalescence time for the Y chromosome is around seventy thousand years ago (Dorit et al., 1995; Thomson et al., 2000), yet human speciation was not at a different time for women than for men. The estimated dates are the coalescence times for the two different genes analyzed, whose polymorphic origins do not necessarily have to be simultaneous. When the origin of polymorphism predated speciation, the coalescence time of the existing alleles of a species can even precede speciation. Whether the coalescence time of existing alleles precedes or follows speciation is dependent on the effective population size.

To estimate coalescence times of genes, alleles, or quasispecies variants, a specific assumption – that sequence divergence increases over time – always has to be made. Time runs only in one evolutionary direction; therefore, even if the morphology of a species has not changed, its sequence divergence will almost always have increased. The time since speciation is related to the extent of sequence divergence in a species. The easiest way to calculate divergence times is to assume that sequence divergence accumulates linearly over time; this is called a ***molecular clock***. When the molecular clock holds, all lineages in the tree have accumulated substitutions at the same rate; the evolutionary rate is constant (see also Chapter 10). However, the evolutionary rate is dependent on many factors, including the metabolic rate in a species, the generation time, bottleneck events, and selective pressure. Therefore, an absolute molecular clock does not exist. There is always some difference in evolutionary rate along the branches of a tree; this is especially true for viruses that have high replication rates, change hosts – and thus selective pressure environment – and frequently go through bottleneck events. However, statistical tests can be performed, as is explained in Chapter 10, that provide an idea of how different the evolutionary rates along the branches in a tree are from a uniform rate. In many situations, these differences are so small that a molecular clock can be safely assumed to calculate coalescence times (or divergence times when starting from ancestral nodes) in a tree. To apply a molecular clock to a tree, the ancestor has to be known; that is, the direction of time in a tree has to be known and the tree has to be rooted.

Such a tree – in which the direction of time is known, the molecular clock holds, and the taxa are organisms – represents a ***cladogram***. A cladogram maps the ancestor–descendant relationship between organisms or groups of organisms. A ***phenogram*** simply represents the relationships among a group of taxa. Because of the effects, of a population tree, a species tree, and a gene tree, and because an absolute molecular clock does not exist, a cladogram will never be identical to a phenogram. Although the topology can be identical, the branch lengths may differ slightly. Cladograms can be drawn based on morphological characters of fossils, and the branches can be calculated from independent dating methods such as radiocarbon dating. A cladogram also can be based on a phylogenetic tree; a phenogram is always based on a phylogenetic tree.

1.5 Methods to infer phylogenetic trees

Reconstructing the phylogeny from gene or amino-acid sequence alignments is, unfortunately, not as straightforward as one might hope, and it is rarely possible to verify that one has arrived at the true conclusion. The reconstruction results in an inferred phylogenetic tree, which may or may not differ from the true phylogenetic

tree. There are no uniquely correct methods for inferring phylogenies, and many methods are used.

The methods for constructing phylogenetic trees from molecular data can be grouped first according to whether the method uses ***discrete character*** states or a ***distance matrix*** of pairwise dissimilarities, and second according to whether the method clusters OTUs stepwise, resulting in only one best tree, or considers all theoretically possible trees.

Character-state methods can use any set of discrete characters, such as morphological characters, physiological properties, restriction maps, or sequence data. When comparing sequences, each position in the aligned sequences is a "character," and the nucleotides and amino acids at that position are the "states." All characters are analyzed separately and usually independently from each other. Character-state methods retain the original character status of the taxa and, therefore, can be used to reconstruct the character state of ancestral nodes.

In contrast, distance-matrix methods start by calculating some measure of the dissimilarity of each pair of OTUs to produce a pairwise distance matrix, and then estimate the phylogenetic relationships of the OTUs from that matrix. These methods seem particularly well suited for analyzing sequence data. Although it is possible to calculate distances directly from pairwise aligned sequences, more consistent results are obtained when all sequences are aligned. Distance-matrix methods allow for scoring multiple hits. When two sequences are divergent, it is likely that at a certain position, two or more consecutive mutations have occurred. These multiple events result in two sequences being more distantly related than can be deduced from the percentage difference in sequence. The more divergent the sequences, the bigger the impact of multiple events. Mathematical models allow for correcting the percentage difference between sequences. This is called the ***genetic*** or ***evolutionary distance***, which is always bigger than the distance calculated by direct comparison of sequences (see Chapter 4). Distance methods discard the original character state of the taxa; as a result, the information to reconstruct character states of ancestral nodes is lost. The major advantage of distance methods is that they are much less computer-intensive, which is important when many taxa have to be compared.

Exhaustive-search methods are tree-evaluation methods that examine the theoretically possible tree topologies for a given number of taxa, using certain criteria to choose the best one. In particular, maximum-likelihood methods (discussed later in this chapter and in Chapter 6) share the main advantage of producing a large number of different trees and estimate for each tree the conditional probability that it represents the true phylogeny, given the data (i.e., the aligned sequences) and a specific evolutionary model (see also Chapters 6 and 7). This allows the investigator to compare the support for the best tree with the support for the second

Table 1.3 Number of possible rooted and unrooted trees for up to 10 OTUs

Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025

best, and to estimate the confidence in the tree obtained. Unfortunately, the number of possible trees, and thus the computing time, grows quickly as the number of taxa increases; the number of bifurcated rooted trees for n OTUs is given by $(2n-3)!/(2^{n-2}(n-2))!$ (Table 1.3). This means that for a data set of more than 10 OTUs, only a subset of possible trees can be examined. Thus, various strategies are used to search the "tree space," but there is no algorithm that guarantees that the best possible tree was examined.

The ***stepwise-clustering*** methods avoid this problem by examining local subtrees first. They are tree-construction methods because they follow specific algorithms to construct a single tree. Typically, the two most closely related OTUs are combined to form a cluster. The cluster is then treated like a single OTU representing the ancestor of the OTUs it replaces; therefore, the complexity of the data set is reduced by one OTU. This process is repeated, clustering the next closest related OTUs, until all OTUs are combined. The various stepwise-clustering algorithms differ in their methods of determining the relationship of OTUs and in combining OTUs into clusters. They are usually fast and can accommodate large numbers of OTUs. Because they produce only one tree, the confidence estimators of the exhaustive search methods are not available, although various other statistical methods have been developed to estimate the confidence in the correctness of a tree obtained. The majority of distance-matrix methods use stepwise clustering to compute the "best" tree, whereas most character-state methods adopt the exhaustive-search approach.

Table 1.4 lists the currently most used phylogenetic tree construction and tree-analysis methods, classified according to the strategy used: character state or distance matrix, exhaustive search or stepwise clustering. All methods use particular evolutionary assumptions, which do not necessarily apply to the data set. Therefore, it is important to realize which assumptions were made when evaluating the best tree given by each method. The methods themselves and their assumptions are extensively explained in the following chapters.

Table 1.4 Most used phylogenetic analysis methods and their strategies

	Exhaustive search	Stepwise clustering
Character State	Maximum parsimony (MP) Maximum likelihood (ML)	
Distance Matrix	Fitch-Margoliash	UPGMA Neighbor-joining (NJ)

Maximum parsimony (***MP***) aims to find the tree topology for a set of aligned sequences that can be explained with the smallest number of character changes (i.e., mutations). The MP algorithm starts by considering a tree with a particular topology. It then infers the minimum number of character changes required to explain all nodes of the tree at every sequence position. Another topology is then evaluated. When all reasonable topologies have been evaluated, the tree that requires the minimum number of changes is chosen as the best tree (see Chapter 7).

Maximum likelihood (***ML***) is similar to the MP method in that it examines every reasonable tree topology and evaluates the support for each by examining every sequence position. In principle, the ML algorithm calculates the probability of expecting each possible nucleotide (amino acid) in the ancestral (internal) nodes and infers the likelihood of the tree structure from these probabilities. The likelihood of all reasonable tree topologies is searched in this way, and the most likely tree is chosen as the best tree. The actual process is complex, especially because different tree topologies require different mathematical treatments, so it is computationally demanding (see Chapters 6 and 7).

UPGMA is the acronym for ***unweighted pair group method with arithmetic means***. This is probably the oldest and simplest method used for reconstructing phylogenetic trees from distance data. Clustering is done by searching for the smallest value in the pairwise distance matrix. The newly formed cluster replaces the OTUs it represents in the distance matrix. The distances between the newly formed cluster and each remaining OTU are then calculated. This process is repeated until all OTUs are clustered. In UPGMA, the distance of the newly formed cluster is the average of the distances of the original OTUs. This process of averaging assumes that the evolutionary rate from the node of the two clustered OTUs to each of the original OTUs is identical. The whole process of clustering thus assumes that the evolutionary rate is the same in all branches, meaning that no one strain has accumulated mutations faster than any other strain. This assumption is almost never true. Therefore, UPGMA tends to give the wrong tree when evolutionary rates are different along the branches (see also Chapter 5).

The ***neighbor-joining*** (***NJ***) ***method*** constructs the tree by sequentially finding pairs of neighbors, which are the pairs of OTUs connected by a single interior node. The clustering method used by this algorithm is quite different from the one

described previously, because it does not attempt to cluster the most closely related OTUs, but rather minimizes the length of all internal branches and thus the length of the entire tree. So it can be regarded as parsimony applied to distance data. The NJ algorithm starts by assuming a bush-like tree that has no internal branches. In the first step, it introduces the first internal branch and calculates the length of the resulting tree. The algorithm sequentially connects every possible OTU pair and finally joins the OTU pair that yields the shortest tree. The length of a branch joining a pair of neighbors, X and Y, to their adjacent node is based on the average distance between all OTUs and X for the branch to X, and all OTUs and Y for the branch to Y, subtracting the average distances of all remaining OTU pairs. This process is then repeated, always joining two OTUs (neighbors) by introducing the shortest possible internal branch (see also Chapter 5).

The ***Fitch-Margoliash method*** is a distance-matrix method that evaluates all possible trees for the shortest overall branch length, using a specific algorithm that considers the pairwise distances.

There have been some reports of comparisons of different sets of algorithms using different sets of data. However, it is difficult to decide which method or methods are best, perhaps because different data sets seem to favor different algorithms. The reason is that different tree-making algorithms are based on different assumptions. If these assumptions are met by the data, the algorithm will perform well. The use of statistical methods helps to estimate the reliability of certain clusters (i.e., tree topologies) and/or branch lengths. However, they are also dependent on the phylogeny method used and suffer from the same bias. There is no evidence that any one method is superior to others, so it is advisable to employ more than one method with each set of data. The ML method intrinsically estimates the standard error on the branch length and therefore already gives some statistical support for each branch length and for the entire tree. The most used tree evaluation method is the ***bootstrapping resampling method***, which is explained in detail in the next chapters.

1.6 Is evolution always tree-like?

The algorithms discussed in the previous section usually generate ***strictly bifurcating trees*** (i.e., trees where any internal node is always connected to only three other branches; see Figure 1.5). This is the standard way of representing evolutionary relationships among organisms through a phylogenetic tree, and it presumes that the underlying evolutionary processes are therefore ***bifurcating*** (i.e., during the course of evolution, any ancestral sequence [internal nodes in the tree] can give rise to only two separate lineages [leaves]). However, there are phenomena in nature, such as the explosive evolutionary radiation of HIV or HCV, that might be best modeled by a ***multifurcating tree***, such as the one shown in Figure 1.7A, or by a ***nonstrictly bifurcating tree*** that allows for some multifurcations, such as the one

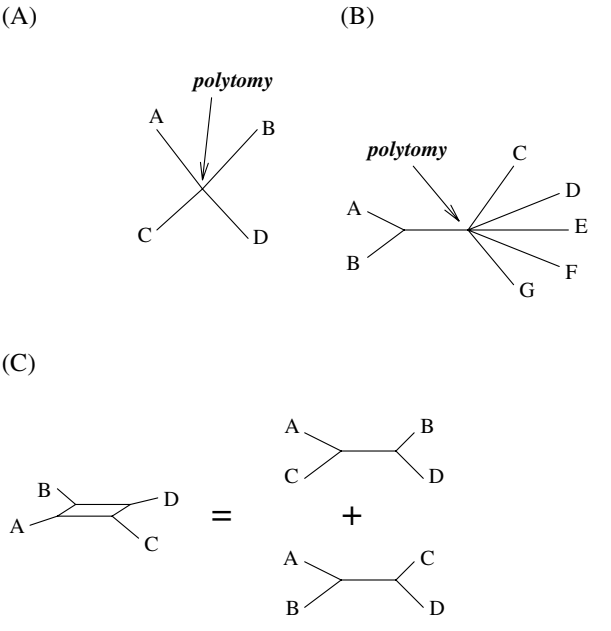

Figure 1.7 Nonbifurcating trees and networks; arrows indicate polytomy. (A) Star-like (or multifurcating) tree. (B) Tree with an internal politomy. (C) Networks representation: the network on the left is one way of displaying simultaneously the two conflicting tree topologies on the right.

in Figure 1.7B. Multifurcations on a phylogenetic tree are also known as ***polytomies***, and can be distinguished as ***hard polytomies*** and ***soft polytomies***. Hard polytomies are meant to represent explosive radiation in which a single common ancestor gave rise to multiple distinct lineages at the same time. Hard polytomies are difficult to prove and it is even questionable as to whether they actually do occur (cf. Li, 1997, and Page and Holmes, 1998, for a detailed discussion). Soft polytomies, in contrast, represent unresolved tree topologies. They reflect the uncertainty about which branching pattern precisely describes the data. Finally, there are situations – for example, in the case of recombination – in which the data seem to support equally well two or more different tree topologies. In such cases, the sequences under investigation may be better represented by a network, such as the one depicted in Figure 1.7C. These topics are covered in Chapters 6, 12, and 14.

REFERENCES

Allison, A. C. (1956). Sickle cells and evolution. *Scientific American*.

Carroll, S. B. (1995). Homeotic genes and the evolution of arthropods and chordates. *Nature*, 376, 479–485.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. London: Murray.

Debyser, Z., E. Van Wijngaerden, K. Van Laethem, K. Beuselinck, M. Reynders, E. De Clercq, J. Desmyter, and A.-M. Vandamme (1998). Failure to quantify viral load with two of the three commercial methods in a pregnant woman harbouring an HIV-1 subtype G strain. *AIDS Research and Human Retroviruses*, 14, 453–459.

Domingo, E., L. Menendezarias, and J. J. Holland (1997). RNA virus fitness. *Review of Medical Virology*, 7, 87–96.

Dorit, R. L., H. Akashi, and W. Gilbert (1995). Absence of polymorphism at the ZFY locus on the human Y chromosome. *Science*, 268, 1183–1185.

Eigen, M. and C. Biebricher (1988). Sequence space and quasispecies distribution. In: *RNA Genetics*, vol. 3, eds. E. Domingo, J. J. Holland, and P. Ahlquist, pp. 211–245. CRC Press, Boca Raton, FL.

Ingman, M., H. Kaessmann, S. Paabo, and U. Gyllensten (2000). Mitochondrial genome variation and the origin of modern humans. *Nature*, 408, 708–713.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Li, W.-H. (1997). *Molecular Evolution*. Sunderland: Sinauer Associates.

Linnaeus, C. (1758). *Systema Naturae*, 10th ed. Stockholm.

Page, R. D. M. and E. C. Holmes (1998). *Molecular Evolution: A Phylogenetic Approach*. Oxford: Blackwell Science.

Patterson, C., ed. (1987). *Molecules and Morphology in Evolution: Conflict or Compromise?* Cambridge: Cambridge University Press.

Thomson, R., J. K. Pritchard, P. Shen, P. J. Oefner, and M. W. Feldman (2000). Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proceedings of the National Academy of Sciences of the USA*, 97(13), 7360–7365.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes, and A. C. Wilson (1991). African populations and the evolution of human mitochondrial DNA. *Science*, 253(5027), 1503–1507.