# Section I

# Introduction

1

# Basic concepts of molecular evolution

Anne-Mieke Vandamme

## 1.1 Genetic information

It was the striking phenotypic variation of finches in the Galapagos Islands that inspired Darwin to draft his theory of evolution. His idea of a branching process of evolution was also consistent with the knowledge of fossil researchers who revealed phenotypic variation over long periods of time. Today, evolution can be observed in real time by scientists, with the fastest evolution occurring in viruses within months, resulting, for example, in rapid development of human immunodeficiency virus (HIV) drug resistance. The phenotype of living organisms is always a result of the genetic information that they carry and pass on to the next generation and its interaction with the environment. Thus, if we want to study the driving force of evolution, we have to investigate the changes in the genetic information.

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; tRNA, which offers the encoded amino acid; or small RNAs, some of which are involved in regulating expression of genes. The genomic DNA also contains elements, such as *promotors* and *enhancers*, which orchestrate the proper transcription into RNA. A large part of the genomic DNA of eukaryotes consists of genetic elements such as introns or alu-repeats, the function of which is still not entirely clear. Proteins, RNA, and to some extent DNA, constitute the phenotype of an organism that interacts with the environment.

DNA is a double helix with two antiparallel *polynucleotide* strands, whereas RNA is a single-stranded polynucleotide. The backbone in each DNA strand

*The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, Philippe Lemey, Marco Salemi, and Anne-Mieke Vandamme (eds.). Published by Cambridge University Press. 

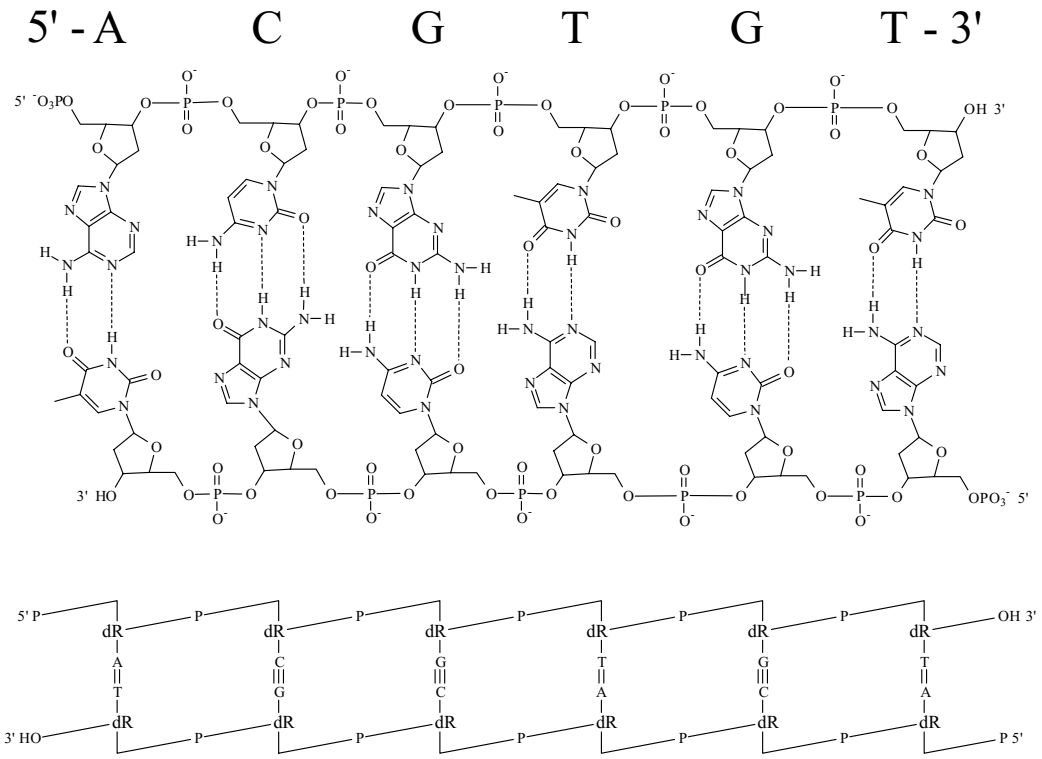


Fig. 1.1 Chemical structure of double-stranded DNA. The chemical moieties are indicated as follows: dR, deoxyribose; P, phosphate; G, guanine; T, thymine; A, adenine; and C, cytosine. The strand orientation is represented in a standard way: in the upper strand 5′–3′, indicating that the chain starts 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 Professor C. Pannecouque.)

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 four following 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 the two DNA strands binding together, with strong (triple) bonds between G and C, and weak (double) bonds between T/U and A (Fig. 1.1). These hydrogen-bonded pairs are called *complementary*. During DNA duplication or RNA transcription, DNA or RNA polymerases synthesize a complementary 5′–3′ strand starting with the lower 3′–5′ DNA strand as template, in order to preserve the genetic information. This genetic information is represented by a one letter code, indicating the 5′–3′ sequential order of the bases in the DNA or RNA (Fig. 1.1). A nucleotide sequence is thus represented by a contiguous stretch of the four letters A, G, C, and T/U.

**Table 1.1** Three- and one-letter abbreviations of the 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 |

In RNA strands encoding 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 generally represented by a contiguous stretch of one-letter amino acid abbreviations (with 20 possible letters).

The **genetic code** is universal for all organisms, with only a few exceptions such as the mitochondrial code, and it 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 by a T. Each position of the triplet code can be one of four bases; hence, $4^3$ or 64 possible triplets encode 20 amino acids (61 *sense* codes) and 3 stop codons (3 *non-sense* codes). The genetic code is said to be degenerated, or redundant, since all amino acids except methionine have more than one possible triplet 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

**Table 1.2** The universal genetic code

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

The first nucleotide letter is indicated on the left, the second on the top, and the third on the right side. The amino acids are given by their three-letter code (see Table 1.1). Three stop codons are indicated.

contiguous nucleotide stretch has three reading frames in the 5′–3′ direction. The complementary strand encodes another three 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** or **ORFs**.

During duplication of the genetic information, the DNA or RNA polymerase can occasionally incorporate a non-complementary 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 way that a point mutation at the third codon position rarely results in an amino acid change (only in 30% of possible changes). A change at the second codon position always, and at the first codon position mostly (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 non-silent or ***non-synonymous***. A site within a coding triplet is said to be *fourfold degenerate* when all possible changes at that site are synonymous (for example "CUN"); *twofold degenerate* when only two different amino acids are encoded by the four possible nucleotides at that position (for example, "UUN"); and *non-degenerate* when all possible changes alter the encoded amino acid (for example, "NUU").

Incorporation errors replacing a purine (A, G) with a purine and a pyrimidine (C, T) with a pyrimidine occur more easily because of chemical and steric reasons. 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 usually have a larger impact on the protein than transitions, because of the more drastic changes in biochemical properties of the encoded amino acid. There are four possible transition errors (A $\leftrightarrow$ G, C $\leftrightarrow$ T), and eight possible transversion errors (A $\leftrightarrow$ C, A $\leftrightarrow$ T, G $\leftrightarrow$ C, G $\leftrightarrow$ T); therefore, if a mutation occurred randomly, a transversion would be two times more likely than a transition. However, the genetic code has evolved in such a way that, in many genes, the less disruptive transitions are more likely to occur than transversions.

Single nucleotide changes in a particular codon often change the amino acid to one 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 non-coding regions are subject to other constraints, such as conservation of binding places for proteins, conservation of base pairing in RNA tertiary structures or avoidance of too many homopolymer stretches in which polymerases tend to stutter.

Errors in duplication of genetic information can also result in the **deletion** or **insertion** of one or more nucleotides, collectively referred to as *indels*. When multiples of three nucleotides are inserted or deleted in coding regions, the reading frame remains intact and one or more amino acids are inserted or deleted. When one 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 from the original protein. The consequence of this change depends on the position in the gene where the change took place. Insertions or deletions are therefore rare in coding regions, but rather frequent in non-coding 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. Having small genomes, viruses make extensive use of this possibility. They often encode several proteins from a single gene by using overlapping ORFs. Another type of

mutation that can change reading frames or make accessible new reading frames is mutations in splicing patterns. Eukaryotic proteins are encoded by coding gene fragments called **exons**, which are separated from each other by **introns**. Joining the introns is called **splicing** and occurs in the nucleus at the pre-mRNA level through dedicated spliceosomes. Mutations in splicing patterns usually destroy the gene function, but can occasionally result in the acquisition of a new gene function. Viruses have used these mechanisms 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, 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.

When parts of two different DNA strands are combined into a single strand, the genetic exchange is called **recombination**. Recombination has a major effect on the genetic make-up of organisms (see Chapter 15). The most common form of recombination happens in eukaryotes during ***meiosis***, when recombination occurs between *homologous chromosomes*, shuffling the *alleles* for the next generation. Consequently, recombination contributes significantly to evolution of diploid organisms. More details on the process and consequences of recombination are provided in Chapter 15.

Another form of genetic exchange is lateral gene transfer, which is a relatively frequent event in bacteria. A dramatic example of this is the origin of eukaryotes arising from bacteria acquiring other bacterial genomes that evolved into organelles such as mitochondria or chloroplasts. The bacterial predecessor of mitochondria subsequently exchanged many genes with the "cellular" genome. Substantial parts of mammal genomes are "littered" with endogenous retroviral sequences, with the "fusion" capacity of some retroviral envelope genes at the origin of the placenta. Every retroviral infection results in lateral gene transfer, usually only in somatic cells.

Genetic variation can also be caused by **gene duplication**. Gene duplication results in genome enlargement and can involve a single gene, or large genome sections. They can be partial, involving only gene fragments, or complete, whereby entire genes, chromosomes (*aneuploidy*) or entire genomes (*polyploidy*) are duplicated. Genes experiencing partial duplication, 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), and especially in plants, new species are frequently the result of polyploidy.

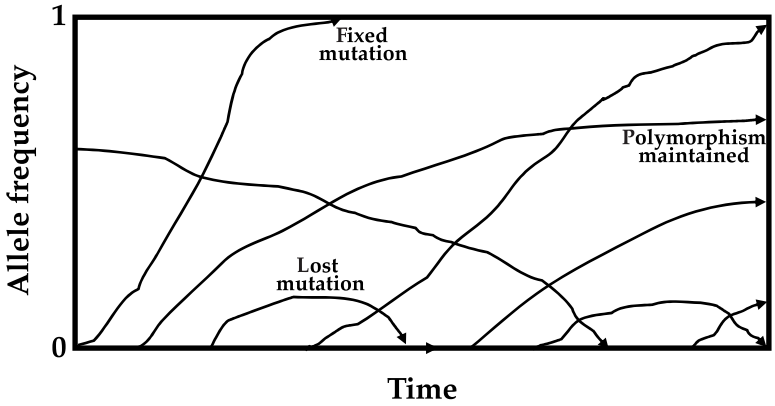


Fig. 1.2 Loss or fixation of an allele in a population.

## 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 typically study the dynamics of the frequency of these polymorphic sites over time. The location in the genome where two or more variants coexist is called the *locus*. The different variants for a particular locus are called *alleles*. Virus genomes, in particular, are very flexible to genetic changes; RNA viruses can contain many polymorphic sites in a single population. HIV, for example, does not exist in a single host as a single genomic sequence, but consists of a continuously changing swarm of variants sometimes referred to as a ***quasispecies*** (Eigen & Biebricher, 1988; Domingo *et al.*, 2006). Although this has become a standard term for virologists, the quasispecies theory has specific mathematical formulations and to what extent virus populations comply with these is the subject of great debate. The high genetic diversity mainly results from 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, while other alleles sometimes increase their frequency to 100%, they become *fixed* in the population (Fig. 1.2). The rate at which this occurs is called the ***fixation rate***. The long-term evolution of a species results from the successive fixation of particular alleles, which reflects fixation of mutations. Terms like ***fixation rate***, ***mutation rate***, ***substitution rate*** and ***evolutionary rate*** have been used interchangeably by several authors, but they

can refer to markedly different processes. This is particularly so for mutation rate, which should preferably be reserved for the rate at which mutations arise at the DNA level, usually expressed as the number of nucleotide (or amino acid) changes per site per replication cycle. Fixation rate, substitution rate, and the rate of molecular evolution are all equivalent when applied to sequences representing different species or populations, in which case they represent the number of new mutations per unit time that become fixed in a species or population. However, when applied to sequences representing different individuals within a population, the interpretation of these terms is subtly altered, because not all observed mutational differences among individuals (polymorphisms) will eventually become fixed in the population. In these cases, fixation rates are not appropriate, but substitution rate or the rate of molecular evolution can still be used to represent the rate at which individuals accrue genetic differences to each other over time (under the selective regime acting on this population). To summarize this from a phylogenetic perspective, the differences in nucleotide or amino acid sequences between taxa are generally called substitutions (although recently generated mutations can be present on terminal branches of trees). If these taxa represent different species or populations, the substitutions will be equivalent to fixation events. If the taxa represent different individuals within a population, branch lengths measure the genetic differences that accrue within individuals, which are not, but ultimately may lead to, fixation events.

The rate at which populations genetically diverge over time is dependent on the underlying mutation rate, the ***generation 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 others in a particular environment, it will be subject to ***positive selective pressure***; if it is less fit, it will be subject to ***negative selective pressure***. An allele can confer a lower fitness to the homozygous organism, while heterozygosity of both alleles at this locus can be an advantage. In this case, polymorphism is advantageous and will be maintained; this is called ***balancing selection*** (heterozygote is more fit than either homozygote). For example, humans who carry the hemoglobin S allele on both chromosomes suffer from sickle-cell anaemia. However, this allele is maintained in the human population because 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 since they do not result in amino acid changes. However, this is not always true. For example, synonymous changes can alter

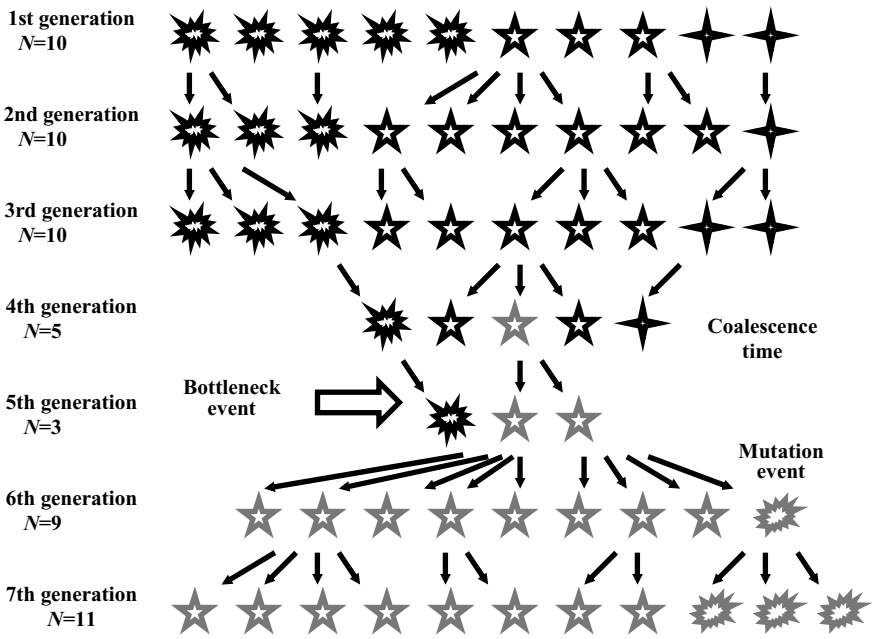


Fig. 1.3 Population dynamics of alleles. Each different symbol represents a different allele. A mutation event in the sixth generation gives rise to a new allele. The figure illustrates fixation and loss of alleles during a bottleneck event, and the concept of coalescence time (tracking back the time to the most recent common ancestor of the grey individuals). *N*: population size.

RNA secondary structure and influence RNA stability; also, they result in the usage of a different tRNA that may be less abundant. Still, most synonymous mutations can be considered selectively neutral.

The rate at which a mutation becomes fixed through *deterministic* or *stochastic* forces depends on the ***effective population size*** ($N_e$) of the organism. This can be defined as the size of an idealized population that is randomly mating and that has the same gene frequency changes as the population being studied (the "census" population). The effective population size is smaller than the overall population size ($N$), when a substantial proportion of a population is producing no offspring, when there is inbreeding, in cases of population subdivision, and when selection operates on linked viral mutations. The effective population size is a major determinant of the dynamics of the allele frequencies over time. When 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 if population sizes are greatly reduced due to catastrophes, or during migrations, etc. (Fig. 1.3). Such events can significantly affect genetic diversity and are called ***genetic bottlenecks***. Two individual lineages merging into