2 Preliminaries

2.1 Biological Background

2.1.1 Cells

Cells are regarded as the fundamental structural units of life being capable of functioning independently. They emerge by cell division of preexisting cells. Any living organism is assembled from cells each of which contains the organism's complete hereditary information that is transmitted to the next generation of cells.

According to the internal cell layout one distinguishes two kinds of organisms. While cells of Eukaryotes contain membrane bound compartments, the organelles, the cells of Prokaryotes lack internal compartments. Prokaryotes are unicellular colonial organisms comprising eubacteria and archaeabacteria. Eukaryotic organisms are both unicellular as well as multicellular and include protozoa and fungi as well as plants and animals. Figure 2.1 schematically illustrates features of an animal cell. The diameter of such a cell ranges from 10 μm to 100 μm. About 50% of a cell's molecules are proteins.

2.1.2 Proteins

Life is manifested in proteins. They constitute most of our bodies' substances and cover a wide range of biological functions. For example, structural proteins form solid material, hormonal proteins coordinate complex body processes, receptor proteins detect biochemical signals, defensive proteins protect against pathogens, transport proteins serve to transport substances and enzymes catalyse biochemical reactions.

The functional diversity of proteins is reflected by an enormous diversity of protein structures. Proteins are macromolecules which are made up of one or more polypeptide chains which in turn are composed of amino acids. An amino acid consists of a central carbon atom (the alpha Carbon C_α) and an amino group (NH_2), a hydrogen atom (H), a carboxy group ($COOH$) and a side chain (R_i) bound to C_α (see Figure 2.2). The covalent bond between the Carbon atom of the Carboxy group of one amino acid and the nitrogen atom of another amino acid's amino group by dehydration is

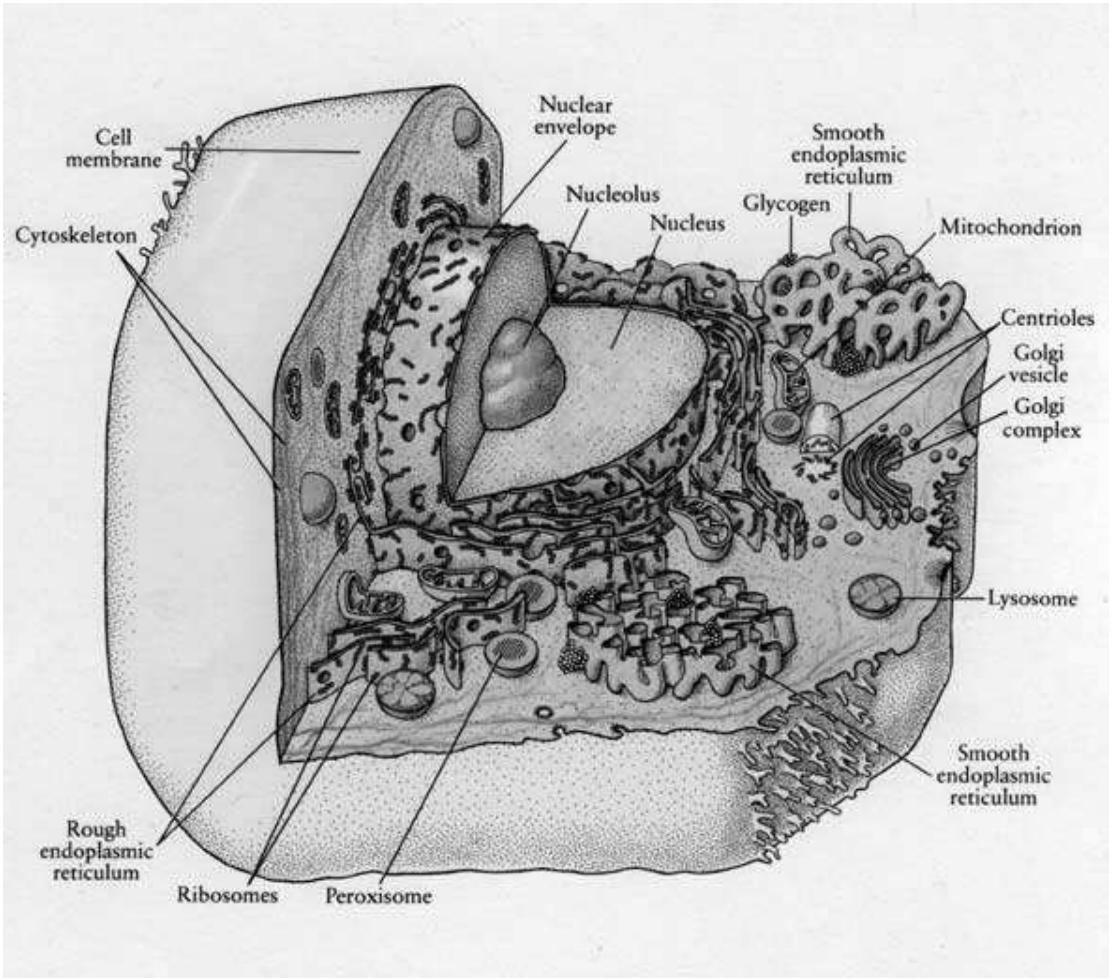


Figure 2.1: Schematic representation of an animal cell. The interior of the cell is called its cytoplasm. It consists of its insoluble constituents, including membrane bound organelles and the cytoskeleton, and the water rich cytosol. The cell membrane encloses the cytoplasm and mediates the transport of substances as well as communication via signal transduction proteins. The organelles perform well defined tasks. Genetic material (DNA) is stored in the double membrane bound nucleus while the nucleolus is the site where ribosomal RNA (rRNA) is produced. The endoplasmic reticulum (ER) constitutes a network of membranous tubules that are connected to the outer membrane of the nucleus and serve to transport proteins. Ribosomes are the sites where proteins are synthesized. They consist of ribosomal proteins and rRNA and are either suspended to the cytosol or attached to the rough ER. Lysosomes are digestive sacks being responsible for degrading substances. The Golgi complex processes ribosomal proteins and sorts them within vesicles. The cell's energy repository in the form of ATP (adenosine triphosphate) is produced by mitochondria.

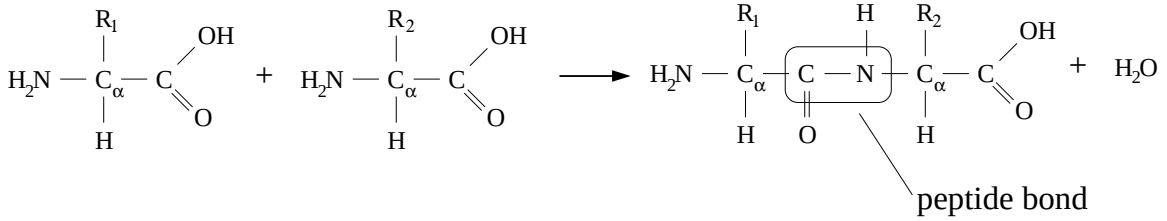


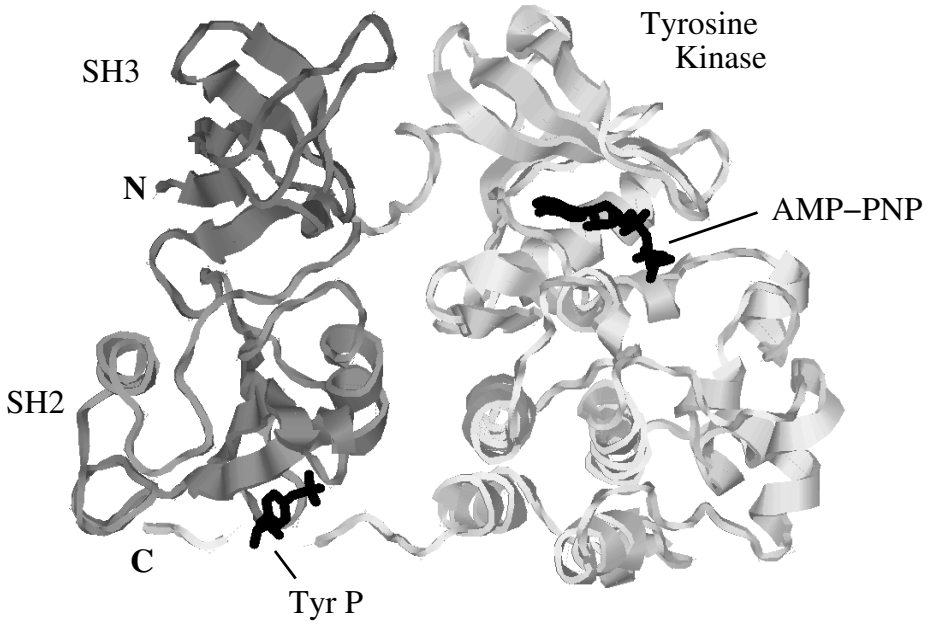
Figure 2.2: Two amino acids form a dipeptide.

called peptide bond. Amino acid residues connected by peptide bonds constitute a polypeptide chain. The backbone of the polypeptide is made up of the repeated sequence of three atoms of each residue in the chain: the amide N, the alpha Carbon C_α and the Carbonyl C. The existence of an amino group (N-Terminal) at one end of the chain and a carboxy group (C-Terminal) at the other end assigns a direction to the chain. Conventionally the beginning of a polypetide chain is its N-Terminal.

Different side chains R_i make up different amino acids with different physical and chemical properties (see appendix **??**). While the side chain of the amino acid glycine just consists out of one hydrogen atom the one of tryptophan contains two carbon rings. Twenty amino acids constitute naturally occuring peptides.

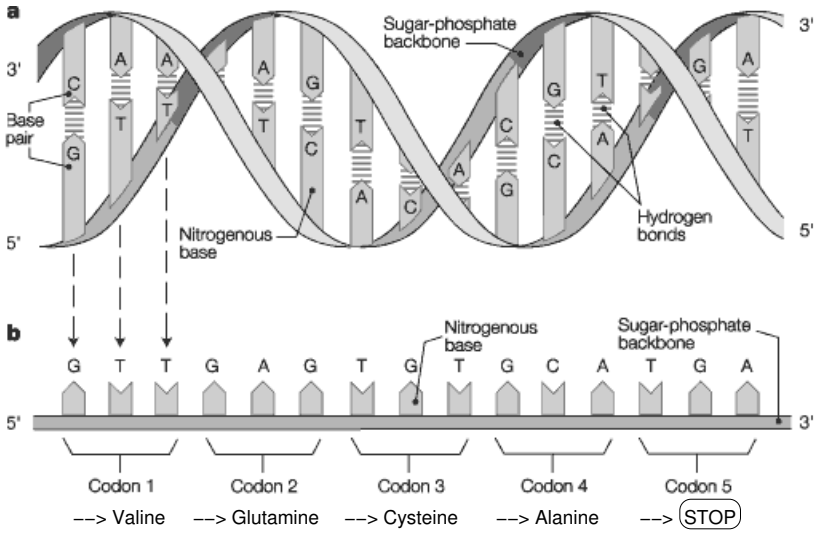
Proteins fold in three dimensions (see Figure 2.3). The term *primary structure* refers to the one-dimensional sequence of amino acid residues. *Secondary structure* refers to stable local spatial structures like α-helices, β-sheets and reverse turns. Those emerge when a lot of weak bonds between backbone atoms stabilize a specific conformation. In contrast, a random sequence of residues is expected to adopt a random coil structure. α-helices and β-sheets are formed by regular hydrogen bonds within a contiguous stretch of the polypeptide chain and are preferably located at the interior of the folded protein. Reverse turns are tight turns that reside at the surface of globular proteins and reverse the direction of the polypeptide chain. *Tertiary structure* describes the three-dimensional packing of secondary structures with respect to a whole polypeptide chain. Often a protein consists of several polypeptide chains. The spatial arrangement of the chains is called *quaternary structure*.

Specific sequences of tightly packed α-helices and β-sheets constitute *structural domains*. Domains are assumed to fold independently of the rest of the protein, they are the basic functional units of a protein. In a multidomain protein several domains are linked by flexible polypeptides. Interactions between secondary structural elements within a domain are stronger than interactions between domains.



The diagram was obtained from the PDB-entry 2SRC (`http://www.rcsb.org/pdb/`) using the RasMol visualization software (`http://www.openrasmol.org/`)

Figure 2.3: Tertiary structure of the inactive form of *proto-oncogene tyrosine-protein kinase Src*. In the so-called ribbon diagram alpha helices are represented by coiled ribbons and beta sheets are represented by arrows. Starting from the N-Terminal the Src protein is composed of an SH3 domain shown in dark gray, an SH2 domain (gray) and a catalytic Tyrosine Kinase domain (light gray). A phosphorylated tyrosine (Tyr P) in the C-terminal tail binds to the SH2 domain. The ATP analog (AMP-PNP) is bound in a cleft of the catalytic domain [Brown and Cooper, 1996; Abram and Courtneidge, 2000].



(Figure adapted from *Molecular Biology of the Cell, Garland Publishing Inc., New York 1994*)

Figure 2.4: a) The DNA double helix. b) Codons code for amino acids.

2.1.3 DNA and gene transcription

Each cell contains the complete genetic information needed to construct and maintain the living organism and in particular its constituting proteins. This "blueprint of life" is stored in the DNA (deoxyribonucleic acid). A DNA is a polymeric molecule composed of a sequence of monomeric subunits called nucleotides. The nucleotides consist of the sugar 2'-deoxyribose, a phosphate group and one out of four nitrogenous bases: adenosine (A) and guanosine (G) are both pyrines, and thymidine (T) and cytidine (C) are pyrimidines. Nucleotides are linked together by phosphodiester bonds where the $5'$ carbon of one phosphate residue binds to the $3'$ carbon of another. Repetitions of the sugar attached to the phosphate form the backbone the bases stick out from. Conventionally a DNA molecule, that is a nucleotide sequence, is defined to start at its $5'$ end and to finish at its $3'$ end (see Figure 2.4 a).

DNA in living cells is double stranded. Two nucleotide chains are wound around each other and form a double helix. The double helix is held together by characteristic hydrogen bonds between bases: while adenine pairs with thymine, cytosine pairs with guanine. Thus the two DNA strands are complementary. Eukaryotic cells commonly contain a couple of large DNA molecules packed in the nucleus within chromosome territories. During mitotic cell division, the DNA is condensed to microscopically visible chromosomes and the DNA is replicated by uncoiling the double helix. Hydrogen bonds are broken to separate the complementary strands each serving as a template for the DNA polymerase to synthesize new DNA strands. A *gene* corresponds to a region of the DNA that encodes for a regulatory function, for proteins or for RNA

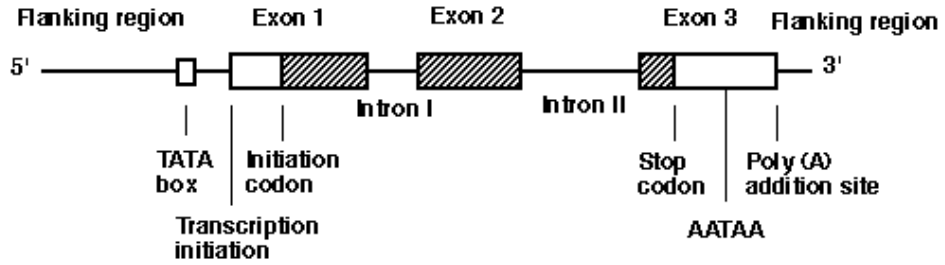


Figure 2.5: Schematic structure of a eukaryotic protein coding gene.

(ribonucleic acid) molecules.

Figure 2.5 shows the gene structure of a eukaryotic protein coding gene. The first step in the expression of a protein coding gene is called *transcription*, the procedure of copying the information contained in the DNA's nucleotide sequence to a single stranded mRNA molecule. Like DNA an mRNA molecule is composed of nucleotides where the sugar is ribose and thymine (T) is replaced by uracil (U). Within eukaryotes transcription requires binding of *transcription factors* to the gene's promoter region located at the $5'$ flanking region of the gene. Transcription factors mediate activation and attachment of the enzyme RNA polymerase II to the DNA. RNA polymerase II catalyzes the synthesis of the RNA strand by adding to the 3' end of the growing RNA molecule one nucleotide at a time. The resulting RNA-molecule is sometimes referred to as pre-mRNA as it is further processed in the nucleus. The transcribed region of a gene consists of so-called exons and introns. In particular the processing of pre-mRNA involves the excision of introns, a process which is referred to as splicing. However sometimes genes are alternatively spliced and certain exons are excised from the pre-mRNA, too. That is, alternative splicing gives rise to different mRNA molecules and thus to different polypeptide sequences. Finally, the mature mRNA migrates out of the nucleus destined for *translation*.

2.1.4 Protein synthesis and secretion

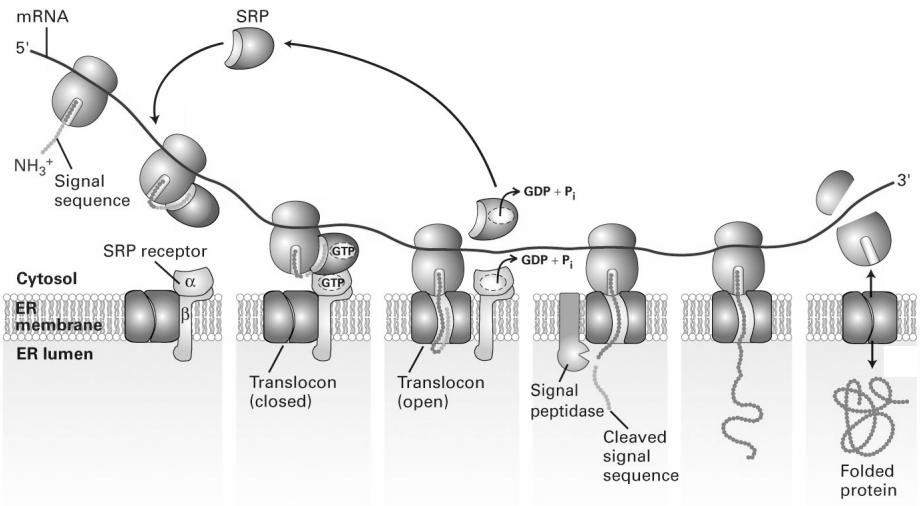
Mature mRNAs are translated to amino acid sequences on ribosomes. Ribosomes consist of a large and a small subunit that are made of ribosomal RNAs (rRNA) and ribosomal proteins. The mRNA specifies amino acids by triplets of nucleotides called *codons* (see Figure 2.4). The mapping of codons to amino acid residues, the so-called *genetic code*, is given in Table 2.1. There are $4^3 = 64$ possible nucleotide triplets. Three *stop-codons* are designated to terminate translation. Because 61 *sense codons* specify only 20 amino acids, the genetic code is said to be degenerate. Transfer RNAs (tRNA) constitute the interface between a codon and an amino acid. Translation of mRNA is initiated when the small ribosomal subunit binds to the *start codon*

First Position	Second Position: U	Second Position: C	Second Position: A	Second Position: G	Third Position
U	phenylalanin	serine	tyrosine	cysteine	U
					C
	leucine		STOP	STOP	A
				tryptophan	G
C	leucine	proline	histidine	arginine	U
					C
			glutamine		A
					G
A	isoleucine	threonine	asparagine	serine	U
					C
	methionine START		lysine	arginine	A
					G
G	valine	alanine	aspartic acid	glycine	U
					C
			glutamic acid		A
					G

Table 2.1: The genetic code.

(AUG) of the mRNA. Large and small ribosomal subunits then fit together to form a complex. While mRNA is shifted through the ribosome, tRNA-attached amino acids are released and catalytically added to the growing peptide chain. When the ribosome encounters a stop-codon, there is no tRNA that associates with it and translation is terminated. The mRNA and the polypeptide are released from the ribosome which in turn dissociates into its two subunits. Within eukaryotic cells, extra-cellular proteins are targeted for secretion, to the cell membrane or to lysosomes. Intracellular proteins are targeted for the cytoplasm, to the nucleus or to mitochondria.

Secreted and *transmembrane* proteins exhibit an N-terminal *signal peptide*. The signal peptide is made up of charged residues like Lysine and Arginine at the N-terminal followed by 7-15 hydrophobic residues (mainly Leu, Ile, Val, Ala, Phe) and about five polar residues. When the signal peptide is translated, it is recognized by and bound to the signal recognition particle (SRP). Further translation is inhibited and the SRP binds to the SRP receptor, a docking protein on the membrane of the *endoplasmic reticulum (ER)* (see also Figure 2.1). The part of the ER where the complex of the SRP and the ribosome binds is called the *rough ER*. Protein synthesis resumes and the growing peptide chain is cotranslationally translocated into the ER. A signal peptidase inside the ER cleaves the signal peptide at a specific cleave site. The completion of protein synthesis is followed by the dissociation of the ribosomal subunits. The rest of the peptide in the ER can either become secreted, be transferred to the Golgi



(Figure adapted from *Molecular Cell Biology, W.H. Freeman and Company, fifth edition, 2003* with permission)

Figure 2.6: The process of signal peptide recognition and translation of a peptide chain being released into the lumen of the endoplasmic reticulum.

apparatus or a lysosome, or be retained in the ER. Figure 2.6 gives a cartoon view of this process.

2.2 Molecular Evolution

2.2.1 Molecular Evolution

Life on earth is manifold. More than a million species are named and the number of unnamed species remains subject to speculation. Despite life's extraordinary diversity all living beings have many points in common. They are assembled from cells, they use DNA molecules to store the hereditary information and their protein synthesizing machineries exhibit the same basic principles and molecules. The similarities among diverse organisms feed the paradigm of evolutionary biology that all living organisms are related and have evolved from common ancestors by modification and diversification of the genetic material. The isolation of a population of organisms and the continuous process of evolutionary change bring about new species. In modern evolutionary theories, mutations on the molecular level of germline cells being inherited to the offspring are considered as the ultimate source of genetic variation. The underlying mechanisms of molecular mutations constitute *large scale mutations* on the chromosomal as well as *small scale* or *point mutations* on the nucleotide level.

2.2.2 Genome evolution

Large scale mutations on the chromosomal level affect a significant percentage of the chromosome. The most dramatic changes occur when chromosomes are unequally sorted into the two daughter cells during cell division. As a consequence whole chromosomes are duplicated or lost. Polyploidy or a genome duplication occurs when all chromosomes end up in one daughter cell. Regional changes on the chromosomal level are caused by mechanisms like unequal crossing over or retroposition. Genomes are said to be *rearranged* when pieces of chromosomes are moved or copied to another location.

2.2.3 Point Mutations

Changes in DNA molecules affecting single nucleotides are called point mutations. They are either caused by errors during DNA replication or by the influence of certain so-called natural mutagenic agents like DNA reactive chemicals or ultraviolet radiation. Point mutations may involve the deletion, the insertion or the substitution of nucleotides in a DNA molecule.

The most frequently occuring type of nucleotide substitions are *transitions* where a purine is replaced by another purine (A,G) or a pyrimidine is replaced by another pyrimidine (C,T). Substitutions of a purine by a pyrimidine or vice versa are called *transversions*.

Nucleotide substitutions in protein coding sequences are classified according to their effect on the translation-product, the protein. While *synonymous substitutions* do not cause a change of the specified amino acid sequence, *nonsynonymous substitutions* alter the amino acid sequence. As a consequence of a nonsynonymous substitution, an amino acid residue may be exchanged by another one in the protein, an effect which is called *amino acid replacement*. Another type of nonsynonymous substitutions are *nonsense mutations* changing a sense codon into a stop codon. As a consequence of a nonsense mutation the protein is truncated.

2.2.4 Amino Acid Replacement

According to Darwins theory of *evolution by natural selection*, mutations in individuals are positively selected and fixed in a population if they improve the individuals' fitness, their ability to survive and to reproduce. With a small number of amino acid sequences at hand, Kimura estimated in the late 1960ies that on average one amino acid replacement occurs every 28×10^6 years per 100 sites of a polypeptide [Kimura, 1968]. The fact that the inferred rate of amino acid replacements was too high to be explained by theories of natural selection prompted him to postulate the *neutral theory*

of evolution: The effect of random genetic drift, that is the random fixation of neutral mutations not affecting the individuals' fitness, cannot be neglected. Later, DNA sequencing revealed that the rate of synonymous substitutions is much larger than the rate of nonsynonymous substitutions and that Kimura even underestimated the rate of evolution acting at the molecular level. Today, modern evolutionary theories at the molecular level generally are consistent with both aspects of the evolutionary process, the selective and the neutral one.

When an amino acid undergoes a radical change into a chemically dissimilar one, the peptide may loose the ability to fold and to take a similar or modified function. As a consequence the fitness of the organism is reduced and the organism has a high chance to be removed from the population. Such a mutation is called *deleterious* and the type of selection is called *negative* or *purifying selection*. In contrast, synonymous substitutions are expected to be selectively neutral and exchanges of amino acids with similar chemical and physical properties are expected to be either neutral or advantageous. Dayhoff revealed the concrete patterns of amino acid replacement and tabulated frequencies and types of amino acid replacements in the 1970ies [Dayhoff *et al.*, 1972, 1978]. She found that the acceptance of an amino acid replacement does not primarily depend on the number of nucleotide substitutions that are required to interchange an amino acid into another. Preferentially physicochemically similar amino acids are exchanged. Her results reveal that natural selection is acting on amino acid replacements.

2.2.5 Evolution of protein function

Gene families

Homologous genes have diverged from an ancestral gene. Since they are related by evolution, they are usually grouped into a so-called *gene family* and the translated proteins are said to be members of a *protein family*. One can distinguish two types of homology: *orthology* and *paralogy* (see Figure 2.7). Two genes are orthologous if they diverged from an ancestral gene by a speciation event. Paralogous genes are descendants of an ancestral gene that has undergone one or more *gene duplications* [Fitch, 1970] where the gene was directly copied within the same genome, e.g., by a genome duplication. Gene families including paralogs are called *multigene families*.

Gene duplications

The evolution of multigene families plays a fundamental role for the emergence of new gene functions. For example Li writes [Li, 1983]: "Gene duplication is probably the most important mechanism for generating new genes and new biochemical processes

that have facilitated the evolution of complex organisms from primitive ones." The emergence of a new gene by a duplication event can be regarded as a mutation event being either advantageous, deleterious or neutral.

Ohno formulated the most popular hypothesis concerning the fate of a duplicated gene in 1973 [Ohno, 1973]: "The mechanism of gene duplication provides a temporary escape from the relentless pressure of natural selection to a duplicated copy of a functional gene locus. While being ignored by natural selection, a duplicated and thus redundant copy is free to accumulate all manner of randomly sustained mutations. As a result, it may become a degenerate, nonsense DNA base sequence. Occasionally, however, it may acquire a new active site sequence, therefore a new function and emerge triumphant as a new gene locus." Ohno denies the influence of natural selection on "redundant" gene copies. His hypothesis implies that one of the two gene copies is prone to a rapid accumulation of nonsynonymous substitutions and that functional diversification of genes within multigene families occurs by chance. Indeed the most typical scenario is that one of the copies of duplicated genes loses its functionality by deleterious mutations and becomes a so-called pseudogene. Yet analysis of nonsynonymous and synonymous substitution rates in case and in large-scale studies of functional duplicate genes provide only little support for an accelerated substitution rate in one of the copies [Lynch and Conery, 2000; Kondrashov *et al.*, 2002; Wagner, 2002; Raes and de Peer, 2003]. This may be due to the fact that once the duplicated gene adapted its new function, it becomes subject to purifying selection and tests based on nonsynonymous substitution rates can only be used in a relatively short time (about 50 millions years) after the genes were duplicated.

The model proposed by Ohno implies that one of the genes retains its original function while the other one acquires a new one. Case studies on the evolution of multigene families rather support the following view [Hughes *et al.*, 1994; Hughes, 2002]: A gene duplication leading to functional diversification of genes is preceded by a period of gene sharing where a single gene performs two distinct functions. After gene duplication the two gene copies specialize in different subfunctions of their ancestor. For example, subfunctionalization sometimes is achieved by a change in regulatory regions of the genes leading to the expression in restricted sets of tissues [Cresko *et al.*, 2003].

One of the most important processes acting on multigene families is *concerted evolution*: Genetic mechanisms like unequal crossing-over and gene conversion transfer DNA sequences between paralogous genes of a genome. As a consequence mutations are spread to several paralogous genes that evolve together. Concerted evolution imposes a problem when paralogous sequences within one organism are used to estimate the time point of a duplication event because the level of sequence divergence does not necessarily reflect the time having passed since duplication.

Domain shuffling and protein diversity in Metazoa

Nucleotide substitutions accumulating in a duplicated gene may alter the regulation of the gene or the structure of the protein product. If the modifications are advantageous for the organism, the mutations are selected for. An alternative more radical possibility for the evolution of novel protein functions emerges when preexisting genes or exons are joined or rearranged. The outcome of such a rearrangement eventually results in a novel and useful combination of protein domains and becomes fixed, a process called *domain shuffling*.

Domain shuffling is a common mechanism in the evolution of novel protein function. About 65% of all prokaryotic proteins and 80% among eukaryotic proteins are supposed to be composed of more than one domain. There are many domains occuring in one or two combinations and few that occur in many combinations [Chothia *et al.*, 2003]. In particular during metazoan evolution, extensive domain shuffling was prevalent and gave rise to the enormous protein diversity of animals and the metazoan radiation [Doolittle, 1995]. Figure 4.2 shows architectures of typical animal multidomain proteins. The EGF-like domain, the *immunoglobulin* (Ig) and the *fibronectin type III* (Fn3) domain were abundantly reused and combined with others to novel proteins to regulate multicellular aspects. Other domains like the *collagen triple helix repeat* or the *sushi* domain occur as long tandem repeats though in various combinations with other domains.

2.3 Computational Molecular Biology

2.3.1 Pairwise sequence alignment

A sequence alignment is a scheme of writing the characters of one sequence on top of another. Characters are either nucleotides or amino acid residues and a vertical column or a site of the alignment holds characters which are deemed to have a common evolutionary origin. If the same character occurs in both sequences then this position may have been conserved in evolution. If the characters differ, the two sequences supposedly have derived from an ancestral character. Homologous sequences not necessarily have the same length which is generally explained by insertions or deletions depicted by pairing of characters with dashes or gaps in the alignment. For example an alignment of the two sequences `RDISLVKNAGI` and `RNILVSDAKNVGI` is

```
... R D I S L V - - - K N A G I ...
... R N I - L V S D A K N V G I ...
```

In the similarity framework one distinguishes between the different possible mismatches and also among different kinds of matches. For amino acids *scoring matrices* have been derived that assign a score to each possible pair of amino acids. The most famous scoring matrices are the PAM and the BLOSUM series of matrices [Dayhoff *et al.*, 1978; Henikoff and Henikoff, 1992]. Like the PAM matrices, the variable time (VT) series matrices are derived from probabilistic models of amino acid replacement [Müller and Vingron, 2000] (see Sections 2.4.5 and 2.4.5). For every matrix one needs to specify appropriate penalties for gaps. The idea is to assign an alignment a similarity score as the sum of site scores and to find the alignment being optimal with respect to the similarity score. Dynamic programming algorithms generate a two-dimensional matrix with values corresponding to optimal similarity scores for sequence prefixes. Then an optimal alignment can be identified with a runtime being proportional to the product of the two sequence lengths. Needleman and Wunsch first introduced a dynamic programming algorithm to align two sequences over their entire length to obtain a *global alignment* [Needleman and Wunsch, 1970]. In many cases sequences share local rather than global similarities. Smith and Waterman adapted the dynamic programming principle for computing pairwise *local alignments* [Smith and Waterman, 1981].

2.3.2 Fast heuristic alignment, BLAST and FASTA

Traditionally, the first step in the analysis of an uncharacterized protein sequence is to search the protein databases for similar sequences. If the search yields significant similarities, information from the proteins in the search results can be inferred to apply to the query sequence. Often the biologist is interested in finding local similarities. Heuristic database searches approximate the search for optimal local alignments and are used to speedup the database search. The two most popular program packages are FASTA [Pearson and Lipman, 1988] and BLAST (Basic Local Alignment Search Tool) [Altschul *et al.*, 1990]. Each hit in a database search comes with a raw score and an *E-value* (Expect value) that describes the number of hits with such a score one can "expect"to see just by chance when searching a database of a particular size. Raw scores of different database searches are not comparable since they do not incorporate knowledge of the scoring system. Therefore it is preferable to cite the *bit score* which is normalized by the statistical parameters of the scoring system and has a standard set of units.

2.3.3 Profiles, PSI-BLAST

While BLAST and FASTA compare a single sequence against a database of sequences, the *profile* concept allows to incorporate more knowledge about the query or the database and to improve the sensitivity and the specificity of a search. A *profile* defined

by Gribskov *et al.* [1987] represents a set of related sequences as a position-dependent scoring matrix being constructed from a multiple alignment of the sequences. The rows of the profile represent the sites of the alignment and the columns correspond to the 20 amino acid residues. Matrix values $M(p, a)$ give the score of amino acid a at position p

$$M(p,a) = \sum_{b=1}^{20} W(p,b)S(a,b) \ ,$$

where $W(p, b)$ holds the relative frequency of amino acid b at position p and $S(a, b)$ is the similarity score from a scoring matrix.

Profiles can be aligned against a sequence using the dynamic programming algorithm. The similarity score for a residue in the sequence and the profile is taken from the column of the position dependent scoring matrix corresponding to the residue and the row representing the position in the profile.

PSI-BLAST (Position-Specific Iterated BLAST) [Altschul *et al.*, 1997] is an extension to BLAST. In one iteration of PSI-BLAST a BLAST search of a profile against a database of protein sequences is performed. A new profile is automatically constructed from significant hits of the search and used to search the database in the next iteration. The initial query to PSI-BLAST is either a single sequence or a multiple alignment. PSI-BLAST has proven to be sensitive in the detection of remote homologies.

2.3.4 Hidden Markov Models

Profile Hidden Markov Models

Hidden Markov Models (HMMs) are statistical, probabilistic and generative models. They contain a doubly embedded stochastic process. While the states emitting output symbols remain hidden, the series of output states is observed.

A *profile HMM* is a stochastical representation of a sequence family and was devised in the 1990ies [Brown *et al.*, 1993; Krogh *et al.*, 1994]. For each column in a multiple alignment there are match, insertion and deletion hidden states to model the evolutionary divergence of the sequences. Match states are emitted according to the symbol distribution in a column of the multiple alignment as given in a row of the profile.

Automatic learning algorithms adapt transition probabilities to best characterize the family members. The model generates sequences by emitting different states. Thus, each sequence has a certain probability to have been generated by the model. Profile HMM are commonly used to search sequence databases to identify additional family members.

Sequences are aligned to profile HMMs in much the same way as they are aligned to a profile. Yet the interpretation of the procedure is different. Aligning a sequence to an HMM is to delineate the most likely series of states having produced the sequence.

HMMER

The HMMER package [Eddy, 1998] provides implementions of profile HMM tools. Programs being mentioned later in the thesis include "hmmpfam"to search an HMM database for matches to a query sequence, "hmmsearch"to search a sequence database for matches to an HMM and "hmmalign"to align sequences to an existing model.

Pfam

Pfam is short for "Protein families database of alignments and HMMs"[Sonnhammer *et al.*, 1997, 1998a; Bateman *et al.*, 2004]. Pfam is both, accurate as well as comprehensive. It contains more than 7500 profile HMMs being built on curated alignments. 74% of the publicly available protein sequences have at least one match to these domains. In this thesis, we refer to Pfam release 12.0.

SMART

SMART (Simple Modular Architecture Research Tool) [Schultz *et al.*, 1998] is a web service allowing the identification and annotation of genetically mobile domains. The SMART release 3.7 we refer to in the thesis includes profile HMMs of 617 domains. The collection of SMART domains provides a lower coverage of the sequence space than Pfam. Yet SMART is based on high-quality human crafted alignments. Domain models come with extensive annotations as well as specific E-value cutoffs for the detection of the lowest scoring true positive and the highest scoring false positive hits in HMM searches. With respect to subcellular localization, SMART domains are categorized as either being “signalling”, “nuclear”, “extra-cellular” or “others”.

TMHMM

TMHMM2 (Transmembrane HMM) is the state-of-the-art program to predict the existence and to detect the topology of transmembrane helices [Sonnhammer *et al.*, 1998b]. The underlying HMM contains different states for the helix core, caps, loops on the cytoplasmic and non-cytoplasmic side and for globular domains in the middle of loops. TMHMM makes a prediction for each residue as being either inside the membrane, outside the membrane or a part of a transmembrane helix.

SignalP

SignalP predicts signal peptides [Nielsen *et al.*, 1997; Nielsen and Krogh, 1998]. The version 2.0 of the program that was applied in this thesis consists of two different predictors based on neural networks and Hidden Markov Models. Several neural networks are trained to evaluate whether an N-terminal residue is part of a signal sequence or a cleavage site. The HMM contains submodels for the N-terminal charged region, the hydrophobic region, and the region around the cleavage site as well as for signal anchors. The latter submodel improves the prediction of signal peptides since it allows to discriminate between signal peptides and uncleaved signal anchors.

2.3.5 Clustering homologous sequences

Sequence clustering

With the rapid growth of biological sequence databases one faces the need for procedures to group evolutionary related proteins. Expert driven family databases classify proteins according to secondary structures [Conte *et al.*, 2000] or by dissecting them into domains, e.g., Pfam domains. Classifying sequences on the basis of such resources inherently requires the previously derived knowledge of the resources to apply to the sequences. A comprehensive clustering of a sequence space including previously uncharacterized sequences still requires automatic procedures. Common sequence clustering procedures are based on pairwise sequence comparisons.

Orthology detection

Orthologous proteins are related by speciation and are supposed to occupy the same functions in different organisms. The inference of orthologous relationships among sequences therefore is of central importance. Further, orthology is a prerequisite for inferring organismal phylogeny from genes or proteins.

Ideally, for a given set of proteins one would like to assign each protein to a protein family and further to have a phylogenetic tree for each family with all homology relations being resolved into paralogy and orthology. Yet automated detection of putative orthologs is restricted to *complete proteomes*, that is to organisms where the sequences of all peptides are known.

Consider the tree shown in Figure 2.7. An ancestral gene was duplicated prior to the divergence of species A and B. Bβ is paralogous to Aα and orthologous to Aβ. The procedure to identify orthologous relationships is based on the following rationale: When comparing the protein Aβ to all proteins of species B, the ortholog Bβ is expected to exhibit the largest similarity. Vice versa Aβ is expected to provide the

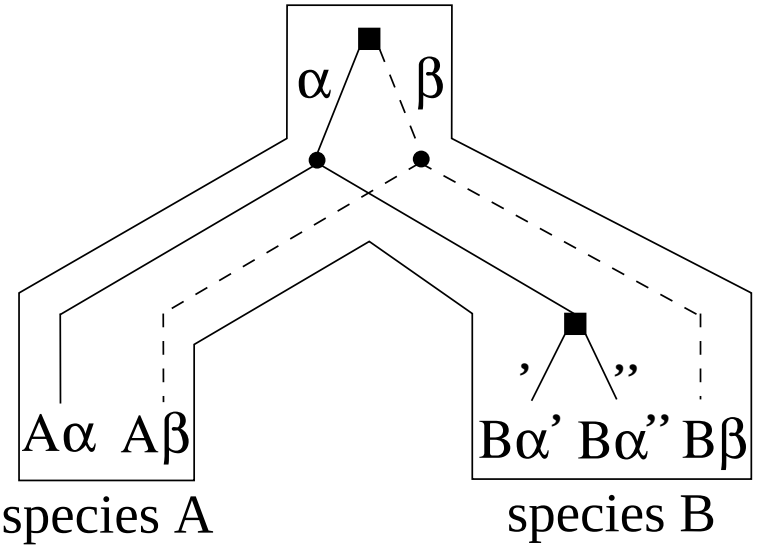


Figure 2.7: Evolutionary history of an ancestral gene embedded in the species tree for species A and species B. Black squares represent duplications, circles represent speciations. Aα is orthologous to Bα' and Bα'', Aβ is orthologous to Bβ. Aα is out-paralogous to Bα'. Bα' is in-paralogous to Bα''.

best match when comparing Bβ to all proteins of species A. In practice, the analysis is performed by BLAST searches between two proteomes. Each protein of one organism is searched against all proteins of the other organism and vice versa. A putative pair of orthologs is detected as the so-called *reciprocal best blast hit*, that is two putative orthologs mutually and significantly find each other as best hits. If the compared proteomes are not complete (for example consider that Aα and Bβ in Figure 2.7 are missing), reciprocal best blast hits may link paralogous sequences.

INPARANOID, COG

The reciprocal best blast hit yields one-to-one relationships. Sonnhammer and Koonin [2002] differentiate paralogous relationships with respect to a speciation and introduced a new terminology. They refer to paralogs as *in-paralogs* if they arose from a duplication after the speciation and as *out-paralogs* if the duplication predates the speciation. Hence the genes Bα' and Bα" in the tree of Figure 2.7 are in-paralogs. Both of them are orthologous to Aα and all genes can be divided into two orthologous groups, one containing Aα, Bα' and Bα" and the other one containing Aβ and Bβ. Clearly all genes in one orthologous group are out-paralogous to all genes in the other orthologous group.

Tatusov *et al.* established a concept to detect in-paralogous relationships and to derive orthologous groups called COGs (Clusters of orthologous groups of paralogs) [Tatusov

et al., 1997]. They require the members of a COG-entry to be present in at least three species. Here, large clusters including multidomain proteins need to be manually resolved into smaller clusters.

A fully automated procedure for finding orthologs and in-paralogs from two species is implemented in the INPARANOID software [Remm *et al.*, 2001]. The authors employ bit scores of BLAST searches between species as well as of searches within the same species. First, reciprocal best blast hits are detected. The two putative orthologs are called *main orthologs* and form the core of an orthologous group. Other sequences are marked to be in-paralogous to one of the main orthologs and eventually added to the orthologous group if the score to one of the main orthologs is larger than or equal to the score between the main orthologs. The user specifies a confidence value ranging from 0% to 100% that shows "how orthologous"a given sequence is. 100% is assigned to a main ortholog and 0% is assigned if the bit score of the putative in-paralogous sequence to the main ortholog from the same species equals the score between the main orthologs.

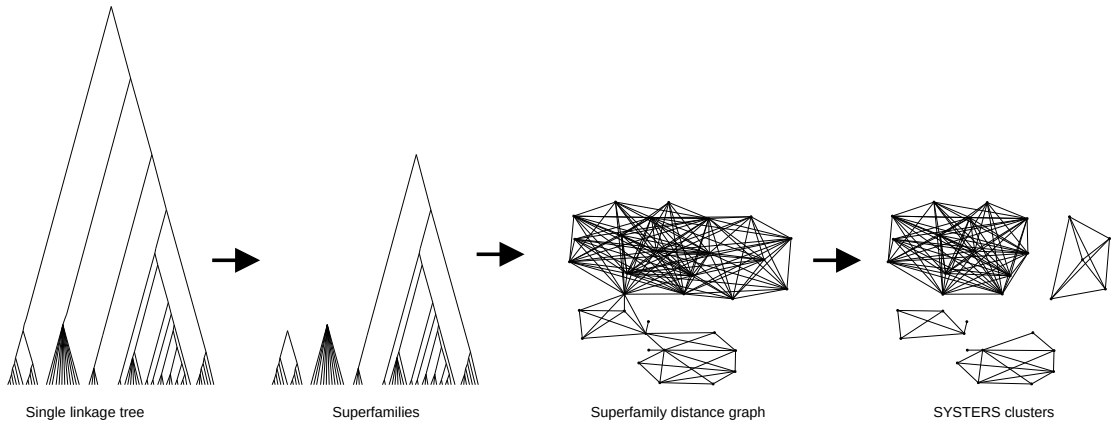
SYSTERS protein families

COG and INPARANOID identify sets of orthologs. These methods do not necessarily assign all sequences to an orthologous group. Further, distant or out-paralogous relationships are not resolved. The SYSTERS clustering procedure is designed to take a whole protein sequence database as input and to hierarchically assign each sequence to a superfamily and a family cluster [Krause and Vingron, 1998; Meinel *et al.*, 2005; Krause *et al.*, 2005]. Superfamily and family clusters are supposed to provide a meaningful partitioning of the whole sequence space.

Meaningful clusters cannot be automatically obtained through simple database searches. Consider a query protein and a database search to identify a group of proteins being related to the query protein. Hits with an E-value lower than 10^{-20} may be considered significant and hits with an E-value larger than 0.01 generally are discarded. E-values in between form the so-called *twilight zone*. Hits within the twilight zone cannot automatically be classified as being or being not related to the query.

The original idea behind the SYSTERS procedure (SYSTERS is short for "SYSTEmatic Re-Searching") is the iteration of database searches. Consider a seed sequence for which a set of related sequences shall be found and an E-value cutoff for significant hits. The initial database search is performed with the seed as query. Significant hits are first included in the set of sequences being related to the seed and second used as queries for subsequent database searches which in turn yield significant hits. The procedure is iterated until the set of related sequences is not extended any more.

Practically SYSTERS clustering first constructs a single linkage hierarchy built on E-values of pairwise all-against-all comparisons. Figure 2.8 gives a schematic overview of



(Figure kindly provided by Antje Krause)

Figure 2.8: Schematic overview of the SYSTERS clustering procedure.

the clustering procedure. Since the proteins evolve at different rates, the application of a static cutoff to derive sets of related proteins does not yield a biologically reasonable partitioning of the data set. Instead, the internal branching structure of the tree is used to automatically determine superfamily specific E-value cutoffs. Here, traversing the single linkage tree from leaf to root and comparing the sizes of merging subtrees is the key to elucidate the twilight zone. The superfamily holds the same members as obtained by the above described "re-searching"procedure. For each superfamily a superfamily distance graph is built from E-values. The superfamily distance graph in turn is cut at weak connections to resolve the SYSTERS family clusters.

In this thesis we refer to the SYSTERS Release 4. The clustering was performed on $1,168,542$ sequences taken from Swiss-Prot (release 41.0) and TrEMBL (release 23.0) [Gasteiger *et al.*, 2003] and from the peptides of several completely sequenced organisms present in Ensembl [Hubbard *et al.*, 2002]. The non-redundant sequence space contains $546,538$ sequences. Those fall into $147,796$ superfamilies and $158,153$ SYSTERS family clusters. $37,488$ SYSTERS clusters contain more than one non-redundant sequence.

2.3.6 Multiple sequence alignment

A multiple alignment of homologous sequences extends the pairwise alignment concept. Subtle similarities being missed in pairwise comparisons become visible when observed simultaneously among many sequences. The construction of a multiple sequence alignment is closely related to reconstructing the evolutionary history of homologous sequences and all methods to reconstruct phylogenetic trees rely on a multiple alignment

of the sequences under study (see Section 2.4.1). Conversely the task of aligning homologous sequences of varying degree of divergence ideally relies on the knowledge of evolutionary relationships among the sequences [Vingron and von Haeseler, 1997].

Sum of pairs- (SP) score, MSA and DCA

The most prominent scoring scheme for a multiple alignment is the *sum of pairs* (SP) score where the scores of the corresponding pairwise alignments contained in the multiple alignment are added. The MSA (Multiple Sequence Alignment) program generalizes the global alignment algorithm for two sequences to optimize the SP-score for more than two sequences [Lipman *et al.*, 1989]. Each sequence to align makes up a dimension in the dynamic programming matrix. And since the search space and the memory requirements are multiplied by the length of every sequence to align, the algorithm can only be run for a moderate number of sequences. For more than three sequences algorithms have been developed that reduce the search space by considering only alignments with a cost below some threshold only while still optimizing the given scoring function. An alternative approach is used by *DCA* (Divide and Conquer Alignment) [Stoye *et al.*, 1997; Stoye, 1998]. According to a Divide and Conquer strategy the sequences are cut several times, the shortened sequences are multiply aligned and the obtained alignments are concatenated. The user can specify a parameter called "recursion stop size" defining the minimal length of a sequence that is not further cut. DCA provides near-to-optimal results for sufficiently homologous sequences. The complexity of the algorithm still depends on the number of sequences to be aligned.

Progressive profile alignment, CLUSTAL

The most common remedy to overcome the demanding time and space requirements of a SP-optimal multiple alignment is to apply a progressive alignment strategy. Progressive alignment relies on an initial estimation of evolutionary relationships among the sequences. For example, distances derived from pairwise sequence alignments can be used to obtain a hierarchical guide tree. Initially the most closely related sequences are aligned. "Progressively"more distantly related sequences or groups of sequences are added to the initial alignment. An already computed alignment on a subset of sequences is interpreted as a profile and "frozen"for the remaining computation. A generalized scoring scheme for profiles uses average scores (see Section 2.3.3). At each internal node of the guide tree a pairwise profile alignment using the dynamic programming algorithm together with the scoring scheme is produced.

The above alignment strategy was outlined in [Feng *et al.*, 1985], and implemented in Higgins program CLUSTAL [Thompson *et al.*, 1994; Higgins *et al.*, 1996]. In CLUSTALW ("W"= weighting), the most widely used program for multiple sequence alignment, contributions of single sequences are weighted such that the information in

the multiple alignment is adequately collected. The major problem with progressive alignment is that the initial guide tree not necessarily reflects the true evolutionary relationships and that errorneous assumptions in the guide tree are propagated to the multiple alignment.

2.4 Molecular Phylogenetics

2.4.1 Inferring molecular phylogenies

The prerequisite for inferring molecular phylogenies on present day sequences is a historically correct multiple sequence alignment in which the aligned positions share a common ancestry. The aim of inferring molecular phylogenies is to reconstruct the evolutionary history of the sequences in the multiple alignment. The evolutionary history is generally modeled by a binary tree. Present day sequences are the leaves of the tree, internal nodes represent ancestral sequences and edge lengths reflect the amount of evolutionary change or the number of substitutions between sequences. The tree that is shown in Figure 2.7 is a *rooted* tree. Sequences at leaves are supposed to have evolved from the common ancestral sequence at the root node through a series of speciation and duplication events. In such a tree, all edges can be assigned a direction with respect to time.

2.4.2 The molecular clock, ultrametric and additive trees

Based on the observation that the number of substitutions between haemoglobins is roughly proportional to divergence times, Zuckerkandl and Pauling have put forward the *molecular clock hypothesis* in the early 1960s [Zuckerkandl and Pauling, 1962, 1965]. The molecular clock assumes that substitutions accumulate in all lineages at constant rates. Assuming validity of the molecular clock is equivalent to reconstructing a rooted *ultrametric tree* where all leaves are equally distant to the root.

The molecular clock assumption rarely holds. Phylogenetic methods therefore reconstruct unrooted *additive trees* that allow evolutionary rates to vary among lineages. Sometimes one intentionally adds a homologous yet distantly related sequence, a so-called *outgroup* sequence, to the data set. Then, subsequent to the tree reconstruction, a root node can be placed at the edge to the outgroup. A rooted ultrametric tree with edges reflecting divergence times and an additive tree are shown in Figure 3.3.

2.4.3 Character- and distance based methods

Phylogenetic tree reconstruction methods can be classified according to the type of input data they use.

Character-based methods use a set of discrete characters. In molecular phylogenetics the characters are nucleotides or amino acid residues at a specific site of the multiple alignment. *Maximum Parsimony* tries to find the tree that explains the evolution of present day sequences by a minimal number of substitutions along the edges of the tree. *Maximum Likelihood* methods evaluate the probablility to observe an alignment under a probabilistic model of sequence evolution (see Section 2.4.5). In principle, character based methods evaluate each tree topology. Yet the tree search space grows super-exponentially in the number of leaves. Computing the Maximum Parsimony or the Maximum Likelihood tree for many sequences becomes time demanding or untractable.

Distance-based methods use a distance matrix as input. A distance matrix can be obtained from a multiple alignment by computing pairwise evolutionary distances (see Section 2.4.5). *Neighbor-Joining* (NJ) is the most commonly used method to reconstruct phylogenetic trees [Saitou and Nei, 1987]. NJ is a clustering method that subsequently identifies neighbors in the tree. Its time complexity is cubic in the number of sequences.

Sometimes prior knowledge of the relatedness among certain subgroups of sequences is available. Then distances can also be computed between profiles and put into the NJ algorithm. *Profile Neighbor Joining* (PNJ) trees have been proven to be more robust and accurate than other reconstruction methods [Müller *et al.*, 2003, 2004].

2.4.4 Non-parametric bootstrapping

Non-parametric bootstrapping is the most commonly used method to obtain a quantity that tells us something about the uncertainty in reconstructed tree topologies [Felsenstein, 1983]. The columns of a multiple alignment provide a sample of the phylogeny to be estimated. The order of alignment columns is irrelevant for the outcome of the tree estimate. Non-parametric bootstrapping generates new aligments, the bootstrap replicates, by randomly drawing columns from the original alignment with replacement. Typically $100 - 1000$ bootstrap replicates are produced and the tree estimation is applied to all bootstrap replicates in turn. *Bootstrap values* (or the *bootstrap support*) are associated to the edges of the tree estimated on the original alignment. They correspond to the relative frequency at which a bipartition of the set of taxa (that results from cutting the tree at the edge) occurs in bootstrap replicates. Boostrap values provide a measure of how robust an estimted tree topology is, when the data is “disturbed”.

2.4.5 Markovian modeling of sequence evolution

Maximum Likelihood and Maximum Parsimony

Assessing the likelihood of sequence data under one or several models is benefitial compared to Maximum Parsimony. Maximum Parsimony relies on William of Ockham's principle that entities ought not to be multiplied unnecessarily and does not provide a model of the underlying process. Indeed Maximum Parsimony trees and Maximum Likelihood trees agree when the sequences' evolutionary rate is small and the probability for a site to have changed more than once is marginal [Felsenstein, 1973]. Yet in many data sets we see evolutionary rates that are not small. Then Maximum Parsimony is prune to reconstructing a wrong topology. Since distances between sequences having evolved at larger rates are underestimated, the taxa at the tips of long branches in a model tree show the tendency to get attracted in the MP tree (*long branch attraction*) [Felsenstein, 1988].

In contrast to assessing the parsimony score computing the likelihood is based on probabilistic modeling of sequence evolution. Here phylogeny estimation implicitely becomes a method of statistical inference. Model parameters are the distribution of characters, substitution rates and the edge lengths and topology of the tree. Different model assumptions are reflected in differences in the parameter space. Comparing phylogenetic hypotheses is subject to a likelihood ratio test statistic.

Probability, likelihood and the likelihood ratio

The aim of a maximum likelihood estimation is to find parameter values that maximize the probability to observe the data. The *probability concept* assumes that we know model parameters Θ which govern the probability $\Pr(\mathcal{X}|\Theta)$ of a random experiment's outcome $\mathcal{X}$. The sum of probabilities over all possible outcomes or data sets $\mathcal{X}$ adds up to one, $\sum_{\mathcal{X}} \Pr(\mathcal{X}|\Theta) = 1$. The *likelihood concept* implies that we observe the outcome of a random experiment and assess the probability $\mathcal{L}(\Theta|\mathcal{X})$ of model parameters Θ when the observed data set $\mathcal{X}$ is fixed:

$$\mathcal{L}(\Theta|\mathcal{X}) = \Pr(\mathcal{X}|\Theta)$$

The ML estimate are parameter values $\theta \in \Theta$ where $\mathcal{L}(\theta)$ assumes its maximum.

If competing hypotheses can be described by restricting some set of model parameters, a likelihood ratio test (LRT) can be carried out [Felsenstein, 1981]. For example one can test whether nucleotide data are better fit under the assumption that transitions occur more frequently than transversions, or whether the reconstruction of an ultrametric tree instead of an additive one is judged, or if there act two different selective

regimes along a sequence rather than one. The task is to compare two nested models of sequence evolution. The simpler model with a smaller number of model parameters represents the null-hypothesis. Its likelihood $\mathcal{L}_0$ is smaller than or equal to the likelihood $\mathcal{L}_1$ of the alternative model, $\mathcal{L}_0 \leq \mathcal{L}_1$. The likelihood ratio statistic Δ is a difference of log likelihoods $\Delta = -2 \log \frac{\mathcal{L}_0}{\mathcal{L}_1} = 2(\log \mathcal{L}_1 - \log \mathcal{L}_0)$. Under the null-hypothesis Δ is (asymptotically) distributed as a chi-square variable with degrees of freedom equal to the difference in the numbers of free parameters in the two models [Ewens and Grant, 2001].

Time-continuous Markov processes

In the following the probabilistic model of sequence evolution is discussed. Assuming that point mutations accumulate according to a stochastic process sequence evolution is modeled by a *Markov process* acting independently on the sites of the sequence. We shortly review the properties of *Evolutionary Markov processes* (EMP) [Dayhoff *et al.*, 1978; Müller and Vingron, 2000; Müller *et al.*, 2002].

Let $\mathcal{A}$ be a finite set of states. In the context of molecular evolution, $\mathcal{A}$ holds the 20 amino acid residues, the 61 amino acid coding codons, or the 4 nucleotides. A *time-continuous Markov process* on $\mathcal{A}$ is a sequence of $\mathcal{A}$-valued random variables $(X_t)_{t \geq 0}$ such that X_0 is distributed as the initial distribution of states π^0 ($\pi^0_i = \Pr(X_0 = i)$) and such that the *Markov property*

$$\Pr(X_{t_n} = a_n | X_{t_{n-1}} = a_{n-1}, ..., X_{t_0} = a_0) \;=\; \Pr(X_{t_n} = a_n | X_{t_{n-1}} = a_{n-1})$$

for times $t_0 < t_1 < ... < t_n$ and states $a_1, a_2, ..., a_n$ holds. The probability distribution of future states only depends on the current state and does not depend on past states.

The Markov process is *homogeneous* if there exists a stochastic *transition matrix*

$$P_{ij}(t) = \Pr(X_{s+t} = j | X_t = i) \qquad \text{for all } s, t \geq 0, \quad i, j \in \mathcal{A} \;,$$

independent of s.

The transition matrix in turn satisfies the *Chapman-Kolmogorov equation*:

$$P(s + t) = P(s)P(t) \;. \tag{2.1}$$

The rate matrix

We assume that the transition matrix $P(t)$ of a time continuous Markov chain is continuous and differentiable from the right at $t = 0$. That is the limit

$$Q = \lim_{t \searrow 0} \frac{P(t) - I}{t}$$

exists, where I denotes the identity matrix. Q is known as the *rate matrix* or the *generator* of the Markov chain. For small time periods $h > 0$, transition probabilities are approximated by

$$\begin{aligned} P(h) &\approx I + hQ \\ P_{ij}(h) &\approx hQ_{ij}, \qquad i \neq j. \end{aligned} \tag{2.2}$$

From equation (2.2) we see that off-diagonal entries of Q are substitution rates: The probability for state i to reach another state j for a small time perid h is proportional to h and Q_{ij}.

From the Chapman-Kolmogorov equation we get

$$\begin{aligned} \frac{d}{dt}P(t) &= \lim_{h \searrow 0} \frac{P(t+h) - P(t)}{h} \\ &= \lim_{h \searrow 0} \frac{P(t)P(h) - P(t)I}{h} \\ &= P(t) \lim_{h \searrow 0} \frac{P(h) - P(0)}{h} \\ \frac{d}{dt}P(t) &= P(t)Q = QP(t)\ . \end{aligned} \tag{2.3}$$

The differential equation can be solved and yields

$$P(t) = \exp(tQ) = \sum_{k=0}^{\infty} \frac{Q^k t^k}{k!}$$

under the initial condition $P(0) = I$.

Q provides an infinitesimal description of the process. Transition probabilities for any $t > 0$ are computed from the matrix Q. From P being a stochastic matrix and with (2.2) we gather some properties of Q for small $h > 0$:

$$P_{ij}(h) \geq 0 \ \Rightarrow \ Q_{ij} \geq 0 \quad \text{for} \quad i \neq j$$

$\sum_j P_{ij}(h) = 1$ and $Q_{ij} \geq 0,\ i \neq j \ \Rightarrow\ Q_{ii} \leq 0$

$1 = \sum_j P_{ij}(h) = 1 + h \sum_j Q_{ij} \ \Rightarrow \sum_j Q_{ij} = 0,\ Q_{ii} = -\sum_{j \neq i} Q_{ij}$

The Markov chain is given by (Q, π^0).

Stationarity and reversibility

The distribution of states π is called *stationary*, if the probability to observe a certain state remains the same for all time points, that is

$$\pi_j = \sum_{i \in \mathcal{A}} \pi_i P_{ij}(t) \quad \text{or} \quad \pi = \pi P(t)$$

for all $t \geq 0$.

For small $h > 0$ and stationarity,

$$\begin{aligned} \pi_j &= \sum_i \pi_i P_{ij}(h) = \pi_j + h \sum_i \pi_i Q_{ij} \\ \text{and} \quad &\sum_i \pi_i Q_{ij} = 0 \quad \text{or} \quad \pi Q = 0\ . \end{aligned} \tag{2.4}$$

The rate matrix Q is called *irreducible* if $P_{ij}(t) > 0$ for any two states $i, j \in \mathcal{A}$ and for some $t > 0$. For irreducible Markov chains the limes $\lim_{t \to \infty} P_{ij}(t) = \pi_j$ exists.

In general, π is not uniform. When reconstructing molecular phylogenies on present day sequences we model the evolution of state i reaching state j in time t by the same process as the evolution of state j reaching state i in time t (*reversibility*, see Figure 2.9):

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t) \tag{2.5}$$

Equation (2.5) is called the *detailed balance*.

The stationary distribution of a reversible Markov process is also called the *equilibrium distribution*. When tree estimation under a reversible model is concerned, reversibility implies that we are ignorant about the root position and the direction of time along the edges.

Calibration

The stationary distribution can be obtained from Q by solving equation (2.4). That is Q provides all model parameters of a Markov process with stationary distribution π. Multiplication of Q with any number yields a rate matrix describing the same process. Explicit model parameters fix the speed of the process.

The rate matrix of an EMP is calibrated to PAM-units (PAM - point accepted mutations). 1 PAM is the evolutionary distance where one substitution event per 100 sites is expected to have occured. That is we calibrate Q by requiring the expected number E of substituion events per time unit to be equal to $\frac{1}{100}$:

$$E \;=\; \sum_i \pi_i \sum_{j \neq i} Q_{ij} \;=\; -\sum_i \pi_i Q_{ii} \;=\; \frac{1}{100}\ .$$

Evolutionary Markov Process (X_t) with stationary distribution

We have discussed the properties of a Markov process being suited to describe the substitution process at a site of a molecular sequence. The definition of *Evolutionary Markov Process* (X_t) with stationary distribution π (π-EMP) summarizes these properties [Müller and Vingron, 2000; Müller *et al.*, 2002, 2004]:

1. (X_t) is time homogeneous,

   $P_{ij}(t) = \Pr(X_{s+t} = j | X_s = i) = \Pr(X_t = j | X_0 = i)$.

2. (X_t) is stationary w.r.t. π,
   $\pi_j = \sum_i \pi_i P_{ij}(t),\ \pi = \pi P(t)\ \ \forall\ t$.

3. (X_t) is reversible, $\pi_i P_{ij}(t) = \pi_j P_{ji}(t)$.

4. (X_t) is calibrated to 1 PAM, the evolutionary distance where one substitution event per 100 sites is expected to have occured.

Nucleotide substitution models

There are two main approaches to assess the rate matrix of an EMP, the *parametric* and the *empirical* one.

Nucleotide substitution models commonly are parametric models. Parameters specifying substitution rates and nucleotide frequencies are adapted from the data set under study. The most simple one-parametric model is the *Jukes-Cantor* model which assumes that nucleotides are uniformly distributed and substituted by each other at equal rates [Jukes and Cantor, 1969]. The *Kimura 2-parameter* model incorporates

the transition-transversion ratio as an additional parameter [Kimura, 1980], the *Felsenstein 81* model assumes non-uniformly distributed nucleotides [Felsenstein, 1981] and the *General time reversible* model does not impose constraints on substitution rates and the nucleotide distribution at all [Lanave *et al.*, 1984]. Likelihood ratio tests serve to check whether a parameter rich model fits the data significantly better than a simpler model.

The codon substitution model

Another example for a parametric model is the *codon substitution model* introduced by Goldman and Yang [Goldman and Yang, 1994]. The codon substitution model was designed to estimate the ratio $\omega = d_N/d_S$ of nonsynonymous to synonymous substitutions in protein coding nucleotide sequences. If the sequences are subject to strictly neutral evolution, nonsynonymous substitutions are fixed at the same rate as synonymous substitutions and $\omega \approx 1$ is expected. A ratio $\omega < 1$ indicates negative or purifying selection on deleterious amino acid changes whereas $\omega > 1$ indicates positive selection. The codon substitution model has a simple structure. The states of the model are the 61 sense codons. The substitution rate from codon i to codon j is given as

$$Q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ at more than one position} \\ \mu\pi_j & \text{if } i \text{ and } j \text{ differ by a synonymous transversion} \\ \mu\kappa\pi_j & \text{if } i \text{ and } j \text{ differ by a synonymous transition} \\ \mu\omega\pi_j & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion} \\ \mu\omega\kappa\pi_j & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition.} \end{cases}$$

Parameters in the rate matrix Q are the rate μ for a nucleotide changing into another, the stationary codon distribution π, the transition-transversion ratio κ and ω.

Models of amino acid replacement (Dayhoff, Müller-Vingron)

When modeling the substitution of the 20 amino acid residues, we observe that their distribution is far from being uniform and the same holds for the replacement frequencies. Further, the number of model parameters specifying transitions between amino acids amounts to 209 and parametric amino acid substitution models therefore are either simplistic or quite complex. Therefore the empirical approach has become generally accepted: The rate matrix is estimated by considering a large set of aligned sequences from a database and the obtained fixed parameter set is supposed to apply to other datasets.

Dayhoff proposed her pioneering and prominent model of amino acid replacement in the 1970ies from which she derived the PAM family of amino acid similarity matrices

[Dayhoff *et al.*, 1972, 1978]. The model is based on global alignments of closely related sequences and the reconstruction of phylogenetic trees followed by the estimation of ancestral sequences. Within the trees she counts the frequency of residues and residue pairs which are used to set up the 1-step transition matrix $P(1)$ of a time-discrete Markov chain. Transition matrices for larger evolutionary distances are obtained from multiples of $P(1)$, that is by extrapolating the observed replacement frequencies between close sequences.

A drawback of the Dayhoff-model is the exclusion of divergent alignments when replacement frequencies are estimated. Consequently a lot of information being available in public databases is discarded and large evolutionary distances are not adequately fit by the model. Müller and Vingron present and apply a Maximum Likelihood (ML) estimator for an EMP which accounts for alignments of varying degree of divergence [Müller *et al.*, 2002]. Estimating the rate matrix Q relies on evolutionary distance estimates of aligned sequences. Distances in turn depend on the rate matrix. Müller and Vingron follow an iterative strategy: they cycle between distance estimation and updating the current rate matrix. The ML approach is computationally demanding. Müller and Vingron provide an alternative empirical approach to estimate Q, the so called *resolvent method* [Müller and Vingron, 2000; Müller *et al.*, 2002] which approximates the ML estimate. As with the ML approach divergent alignments are taken into account. In contrast to the ML approach the resolvent method is considerably faster.

Similarity scores for amino acids

Searching a database with a query protein commonly involves a similarity measure between amino acid residues (see Section 2.3.1). The rationale of deriving scoring matrices from the rate matrix of an EMP is the following: Similar residues are replaced by each other more frequently than less similar residues. The measure of similarities between amino acids reverses this relation [Dayhoff *et al.*, 1978]. The *similarity score* for a pair of amino acids (i, j) is defined as

$$S_{ij}(t) := \log \frac{\pi_i P_{ij}(t)}{\pi_i \pi_j} \ .$$

The score is positive if the pair (i, j) frequently occurs in the alignments that were used to estimate the rate matrix of the EMP.

Maximum Likelihood estimation of evolutionary distances

Homologous sequences have diverged from an ancestral sequence. To measure the degree of their evolutionary divergence or their evolutionary distance is a fundamental

data set $\mathcal{X}$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$t_0 = 0$	F	P	H	V	L	E	V	S
$t_1 = t$	L	Q	H	V	L	E	Y	G

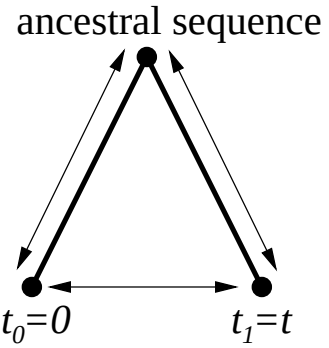


Figure 2.9: Two homologous amino acid sequences having evolved from an ancestral sequence. The reversibility of the evolutionary Markov process allows interpreting the substitution pattern as the result of the process acting on the sites of one sequence.

task. There are many distance measures, see e.g. [Zharkikh, 1994] for a review. Among those is the popular *paralinear* or *LogDet*-distance measure which applies to cases where the character distribution is not assumed to be homogeneous in different regions of the tree [Barry and Hartigan, 1987; Lake, 1994; Lockhart *et al.*, 1994]. Here we present the ML approach which is described, e.g., in [Felsenstein, 1993; Adachi and Hasegawa, 1996; Müller and Vingron, 2000].

Consider an amino acid sequence comprising N sites and let the sequence evolve according to a π-EMP. Let $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ denote the observed data set where x_i is a 2-dimensional vector holding the states or amino acid residues of site i at times $t_0 = 0$ and at time $t_1 = t$ (see Figure 2.9).

The probability that the data set $\mathcal{X}$ is generated by the π-EMP in time t is assessed from the set of parameters $\Theta = \{Q, t\}$:

$$\Pr(\mathcal{X}|t, Q) \;=\; \prod_{i=1}^{N} \Pr(x_i|t, Q) \;=\; \mathcal{L}(t, Q|\mathcal{X}) \tag{2.6}$$

For example, $\Pr((F, L)|Q, t) = \pi_F \cdot P_{FL}(t) = \pi_F \cdot [\exp(Qt)]_{FL}$.

Now consider that we observe two homologous sequences having evolved from an ancestral sequence and want to estimate the evolutionary distance of the sequences if they have evolved according to the π-EMP (see Figure 2.9). Due to the reversibility of the process the likelihood of observing the sequences with an evolutionary distance t is given by $\Pr(\mathcal{X}|t, Q)$ (see equation (2.6)). The Maximum Likelihood estimate of the evolutionary distance $\hat{t}$ is the t where the likelihood function $\mathcal{L}(t)$ assumes its maximum. The logarithm of the product in equation (2.6) is computationally easier

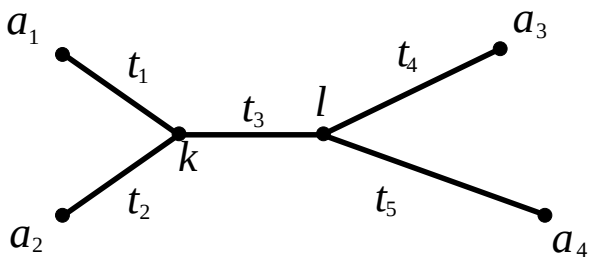


Figure 2.10: An additive tree for four taxa.

to assess. And since the logarithm is strictly increasing, maximizing $\log \mathcal{L}(t)$ instead of $\mathcal{L}(t)$ yields the same distance

$$\hat{t} \;=\; \underset{t}{\operatorname{argmax}} \;\; \log \mathcal{L}(t) \;=\; \underset{t}{\operatorname{argmax}} \;\; \textstyle\sum_{i=1}^{N} \log \Pr(x_i|t, Q).$$

Practically, $\hat{t}$ is computed by numerical optimisation.

With respect to the ML distance estimation, we interpret the sites of the alignment as samples or as outcomes of an EMP having acted on an ancestral sequence. The limited number of samples prevents the ML estimate to assume the true or the model distance. According to the asymptotic theory, the distances are normally distributed and centered around the true distance with a variance equal to the inverse Fisher information [Lindgren, 1993; Müller, 2001; Müller *et al.*, 2002].

Maximum Likelihood estimation of phylogenetic trees

The ML estimate of an evolutionary distance is a ML estimate of a phylogenetic tree for two taxa. The likelihood calculation for a tree with more taxa simply extends the above calculation. Time-reversibility of the model again implies that we cannot infer a root.

Consider four sequences and that we observe the states $x_i = \{a_1, a_2, a_3, a_4\}, a_i \in \mathcal{A}$, at site i of an alignment. In Figure 2.10 the states are placed at the tips of an additive tree according to its topology T. Edge lengths of the tree are denoted by $t_1, ..., t_5$. First assume that states $k, l \in \mathcal{A}$ at internal nodes are known. We arbitrarily choose a root node, e.g., the one with state k. The probability to observe these states is obtained by evolving the state of the root node according to the tree under the model:

$$Pr(x_i, k, l) \;=\; \pi_k P_{ka_1}(t_1)\; P_{ka_2}(t_2)\; P_{kl}(t_3)\; P_{la_3}(t_4)\; P_{la_4}(t_5)$$

Since we are ignorant about states at internal nodes, we sum over each possible state at internal nodes and obtain the probability to observe site i given T, edge lengths $t_1, ..., t_5$ and the substitution model Q:

$$\Pr(x_i|t_1, ..., t_5, T, Q) = \sum_{k \in \mathcal{A}} \pi_k P_{ka_1}(t_1)\ P_{ka_2}(t_2) \sum_{l \in \mathcal{A}} P_{kl}(t_3)\ P_{la_3}(t_4)\ P_{la_4}(t_5)$$

Sites are modeled independently of each other and the log likelihood to observe an alignment $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ of four sequences and the tree topology T is

$$\begin{aligned} \log \mathcal{L}(t_1, ..., t_5, T) &= \log \Pr(\mathcal{X}|t_1, ..., t_5, T, Q) \\ &= \sum_{i=1}^{N} \log \Pr(x_i|t_1, ..., t_5, T, Q). \end{aligned} \tag{2.7}$$

The *maximum likelihood tree* is the one with toplology $\hat{T}$ and edge lengths $\hat{t}_1, ..., \hat{t}_5$ maximizing $\log \mathcal{L}(t_1, ..., t_5, T)$ when the substitution model is fixed.

The log likelihood of a site is efficiently computed by recursion [Felsenstein, 1981]. A rooted tree topology is traversed from the leaves to the root. To each internal node k a conditional likelihood $\mathcal{L}_{s,k}$ is assigned as the likelihood of the subtree rooted at k, given that node k has state $s \in \mathcal{A}$.

ML estimation of phylogenetic trees is computationally the most expensive tree reconstruction method. A very fast and widely used heuristic to reduce the tree search space is *Quartet Puzzling* [Strimmer and von Haeseler, 1996; Schmidt *et al.*, 2002]. The optimal tree for all quartets, that is for all subsets of sequences consisting only of four sequences, is computed. Subsequently, the quartet trees are combined into a larger tree for all sequences.