Institutionen för Cell- och Molekylärbiologi
Karolinska Institutet

Orthology and Protein Domain Architecture Evolution

VOLKER HOLLICH

Karolinska Institutet

Stockholm 2006

Institutionen för Cell- och Molekylärbiologi, Karolinska Institutet
171 77 Stockholm
ISBN 91-7140-783-9 SWEDEN

Tryck: Larserics Digital Print AB

Ich wünsche der nachfolgenden Generation
alles Gute im Umgang mit dem Computer.
Möge dieses Werkzeug Ihnen helfen,
die Probleme dieser Welt zu lösen,
die wir Alten Euch hinterlassen haben.
Konrad Zuse (1910-1995)

Abstract

A major factor behind protein evolution is the ability of proteins to evolve new domain architectures that encode new functions. Protein domains are widely considered to constitute the "atoms" of protein chains, acting as building blocks of proteins as well as evolutionary units. A small number of domains are found in many different domain combinations, while the majority of domains co-occur with very few types of other domains. Domain architectures are not necessarily created once only during evolution. Cases of convergent evolution show how a favourable domain architecture has evolved multiple times independently. A basic concept for understanding evolution on gene level is orthology. Two genes are orthologous if they have evolved from the same gene in the last common ancestor of the species and have thus been created by a speciation event. Paralogous genes result from a duplication event that produced two gene copies within the same species. The concept of orthology can be transferred from genes to protein domains and utilised to explain recombination of protein domains and the evolution of domain architectures.

The focus of this work is to augment the understanding of domain architecture evolution and its functional implications. We have examined, evaluated and improved existing methods as well as developed new approaches. The concept of orthology plays a major role in this work. Orthology is often inferred from phylogenetic trees that are based on pairwise distance estimations of protein sequences. The *Scoredist* protein sequence distance estimator has been developed as one part of this thesis. It combines robustness with low computational complexity and can be calibrated towards various evolutionary models. Accurate phylogenetic trees are crucial for many applications, hence the appropriate tree reconstruction algorithm should be chosen with care. The strengths and weaknesses of many current tree reconstruction algorithms were assessed, and findings underscore the value of the *Scoredist* estimator. The Pfam protein families database comprises a large number of protein families and domains. As part of this thesis it has been enhanced by search and query tools, such as PfamAlyzer or the browser-based domain query, that can be applied on whole domain architectures instead of individual domains only.

We have developed a Maximum Parsimony algorithm for the prediction of ancestral domain architectures. In contrast to previous approaches, it employs gene trees rather than species trees. The algorithm was a starting point for an extensive study of the domain architectures present in Pfam for 50 completely sequenced species. Sampling widely across the kingdoms of life, the study sought to find and analyse cases where a domain architecture had been created multiple times. The algorithm proved robust to potential biases from horizontal gene transfer. Convergent evolution of domain architectures was found more frequently than by previous approaches. No strong biases driving convergent evolution were found. It therefore seems to be a random process in much the same way evolution through duplication and recombination, yet less frequent.

Original publications

I. Hollich V, Storm CEV, Sonnhammer ELL (2002)
OrthoGUI: graphical representation of Orthostrapper results
Bioinformatics **18**:1272-3

II. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR (2004)
The Pfam protein families database
Nucleic Acids Res **32**:D138-41

III. Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer ELL, Bateman A (2006)
Pfam: Clans, Web Tools and Services
Nucleic Acids Res **34**:D247-51

IV. Hollich V, Sonnhammer ELL (2006)
PfamAlyzer: Domain-Centric Homology Search
submitted

V. Sonnhammer ELL, Hollich V (2005)
Scoredist: a simple and robust protein sequence distance estimator
BMC Bioinformatics **6**:108

VI. Hollich V, Milchert L, Arvestad L, Sonnhammer ELL (2005)
Assessment of Protein Distance Measures and Tree-Building Methods for Phylogenetic Tree Reconstruction
Biol Mol Evol **22**:2257-2264

VII. Hollich V, Henricson A, Sonnhammer ELL (2006)
Gene Tree based Analysis of Domain Architecture Evolution
submitted

Contents

Chapter 1

Introduction

Die Natur versteht gar keinen Spaß,
sie ist immer wahr, immer ernst, immer strenge;
sie hat immer recht, und die Fehler und Irrtümer
sind immer die des Menschen!
Johann Wolfgang von Goethe (1749-1832)

Grundlagenforschung ist, was ich tue,
wenn ich nicht weiß, was ich tue.
Wernher Freiherr von Braun (1912-1977)

Evolution has been described as "the mystery of mysteries the replacement of extinct species by others", by Johann Friedrich Herschel. In a letter to Charles Lyell he continued

> "He that on such quest would go must know not fear or failing
> To coward soul or faithless heart the search were unavailing."

It remains unknown if Charles Darwin was aware of this quote when he started his research into this subject. However in the preface of his work *"On the origin of species"* (1859; 1872) he cited Herschel even if not in name. In the 6th edition, Darwin sketched the difficulties of his research

> "We possess no pedigrees or armorial bearings; and we have to discover and trace the many diverging line of descent in our natural genealogies, by characters of any kind which have long been inherited."

At first, only more or less obvious properties such as morphology were at hand (Haeckel, 1866). Nowadays molecular biology makes available the blueprint of organisms contained in the deoxyribonucleic acid and ribonucleic acid sequences. Sequencing projects of the human genome (Lander *et al.*, 2001; Venter *et al.*, 2001)

and numerous others research efforts constantly deliver more and more data to be analysed and understood.

1.1 Homology, orthology and paralogy

The concept of homology (ὁμος same, equal, similar; λόγος word, speech, principle) describes relationships between individuals that are involved in evolutionary processes. According to the modern definition, features are homologous if they share a common evolutionary origin (Reeck *et al.*, 1987). The general idea is that homologous proteins will have similar properties, such as localisation or function. Homology is often imposed on nucleic sequences, however this is no neccessity and the concept can also be looked at from a broader angle.

An early definition of homology was given by Owen (1843) as "the same organ under every variety of form and function". This definition is even older than Darwin's first publication on evolution (1859) and it therefore completely lacks an evolutionary perspective. Transferring Owen's definition into today's world is not completely straightforward. "The same organ" can be interpreted from a functional, morphological or evolutionary point of view. Occasionally the different perspectives will coincide in their result. Although, in a larger number of cases they will diverge and lead to different classifications (Patterson, 1988; Abouheif *et al.*, 1997).

The current molecular definition of homology states that two nucleic acid sequences are homologous if they have evolved from a common ancestor. The terms orthology and paralogy have been introduced to extend the definition of homology (Fitch, 1970).

> "Where the homology is the result of gene duplication so that both copies have descended side by side during the history of an organism, (for example, alpha and beta hemoglobin) the genes should be called paralogous (para = in parallel). Where the homology is the result of speciation so that the history of the gene reflects the history of the species (for example alpha hemoglobin in man and mouse) the genes should be called orthologous (ortho = exact)."

Applied to figure 1.1, the chicken α gene is an ortholog to both the mouse and human α gene, since they are separated by speciation. Likewise is the chicken β gene an ortholog to both the mouse and human β gene. The two hemoglobin genes of one species are separated by a duplication event. Thus, the human α and β gene are paralogs.

Orthology and paralogy have a number of interesting properties (Fitch, 2000):

- Orthology is not transitive. If the pairs A,B and B,C are known orthologs, it may not be concluded that A,C are orthologs as well. Transitivity is a

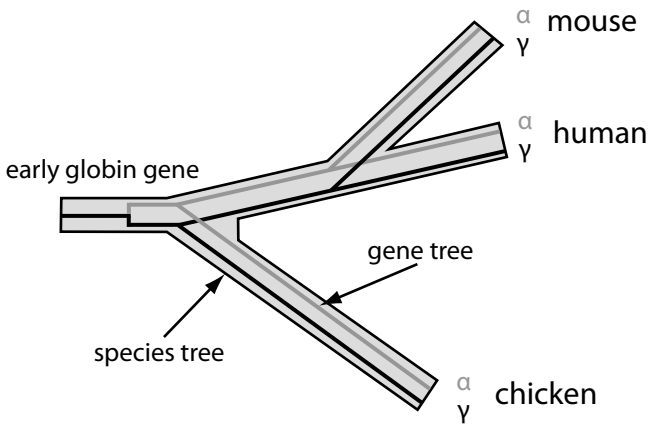

Figure 1.1: The evolution of hemoglobin in selected species. The gene tree for the hemoglobin genes is arranged in the species tree. The ancestor at the tree root carried only an early globin gene. This was duplicated prior to the speciation event that separated chicken from human and mouse, giving rise to an α and β copy in all the species.

property of many algebras. Orthology, in contrast to simple homology, does not show this property.
The sequence α_A in figure 1.2 is separated from α_B and β_B by the speciation event s_1. Following the definiton of orthology, the sequence pairs α_A, α_B and α_A, β_B are orthologs. Nevertheless, α_B and β_B are paralogs since they are separated by the duplication event d_1 .

- Orthology can be an one-to-many or many-to-many relationship.
  An example of a many-to-many relationship in figure 1.2 are the two proteins α_C and γ_C that are homologs to α_A,α_B and α_D.

- The phylogeny of orthologous sequences precisely reflects the phylogeny of the involved species. This property is unique to orthologs.
  The black α proteins are all orthologs to each other. They mirror exactly the species tree.

A further differentiation of paralogy is the introduction of in- and outparalogs (Sonnhammer and Koonin, 2002). The distinction between in- and outparalogs is the point at which the duplication took place relative to the speciation event. For inparalogs duplication happened after the speciation event whereas for outparalogs the reverse is true. In consequence, inparalogs are automatically orthologs whereas outparalogs are not.

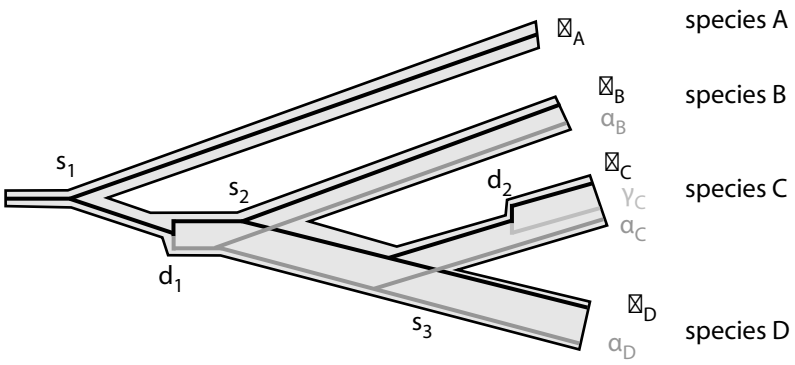

Figure 1.2: Example of speciation and duplication events in the evolution of four species.

When comparing species B and C in figure 1.2, the proteins α_C and γ_C are inparalogs whereas α_C and β_B are outparalogs.

Horizontal gene transfer describes the insertion of genetic material from any organism but its ancestors. It is known to be common within bacteria (Jain *et al.*, 1999; de la Cruz and Davies, 2000) but signs of horinzontal gene transfer between other organisms have also beeen observed (Storm and Sonnhammer, 2003). The term xenology has been created to describe phylogenies that arose from horizontal gene transfer (Gray and Fitch, 1983; Novozhilov *et al.*, 2005).

1.2 Protein domains and architectures

Protein domains are parts of proteins or they can constitute a smaller protein. They are considered the atoms of evolution that undergo recombination to create new function (Bornberg-Bauer *et al.*, 2005; Janin and Chothia, 1985; Riley and Labedan, 1997; Vogel *et al.*, 2004a; Doolittle, 1995). It has been estimated that two thirds of prokaryotic proteins and 80% of eukaryotic proteins have more than one protein domain (Liu and Rost, 2004). One of the first described domains is a NADH-binding domain named after its discoverer the Rossman domain (Rossmann *et al.*, 1974). Protein domains can be defined either by sequence similarity or by using the three-dimensional structure. It is believed that there is an upper bound for the number of protein domains (Wolf *et al.*, 2000; Chothia, 1992). The evolution of protein domains has been described by stochastical models (Karev *et al.*, 2003, 2002; Koonin *et al.*, 2002).

1.2.1 Protein domain databases

Various protein domain databases have been created that use different classifications of protein domains. The SCOP database defines protein domains by structure and groups domains on four hierarchical levels of similarity (family, superfamily, fold, class) (Murzin *et al.*, 1995; Andreeva *et al.*, 2004). The CATH database uses a semi-automatic procedure to classify structures into four hierarchical levels (protein class, architecture, topology and homologous superfamily) (Orengo *et al.*, 1997, 2003; Pearl *et al.*, 2003). It has been shown that SCOP and CATH agree on the majority of classifications (Shakhnovich and Harvey, 2004; Hadley and Jones, 1999).

The Pfam database (**Paper II, III**; Sonnhammer *et al.*, 1997, 1998) defines protein domains based on sequence similarity. Protein domains are manually curated building a multiple seed for each domain familiy alignment. The seed alignment is used to train a profile-Hidden Markov Model (HMM) (Eddy, 1998, 1996; Durbin *et al.*, 1998) built by the hmmer software package (Eddy, 2006). The obtained model it subsequently used to find other members of the domain family. This is achieved by running the UniProt sequence database (Bairoch *et al.*, 2005) against the profile-HMM. The found domains construct the full dataset that can be several orders of magnitude larger than the seed. This approach combines the quality of manually curated data with the quantitative advantages of automatic procedures. Until recently Pfam only consisted of domain families and lacked a hierarchy. The 2006 article (**Paper III**) introduced clans as a new hierarchical level. Some but by far not all domain families are manually grouped together to form a clan. Besides these high quality Pfam-A families, Pfam-B families are of lower quality and completely automatically created using the Domainer algorithm (Sonnhammer and Kahn, 1994).

The SUPERFAMILY database applies the same principle as Pfam using a manually curated seed dataset to train profile-HMMs (Gough, 2002; Gough and Chothia, 2002; Madera *et al.*, 2004). One of the differences to Pfam is that SUPERFAMILY uses the SAM program package to create profile-HMMs (Hughey and Krogh, 1996). The properties, strengths and weaknesses of hmmer and SAM have been studied in detail (Madera and Gough, 2002; Wistrand and Sonnhammer, 2005).

1.2.2 Domain architectures and recombination

Protein domains are often referred to as the building blocks or "atoms" of evolution. The majority of known protein sequences consist of multiple domains. It is, however, not completely understood how domains are recombined throughout evolution. Yet some progress has been achieved. Apic *et al.* (2001) used the SCOP database to analyse domain architectures. They found that most domain families combine with only one or two other domain families. A small number of domain families are very versatile in their combination behaviour and can have various neighbouring domain

families. The distribution of the domain family versatilities follows a power law. The graph of domain combinations thus forms a scale-free network (Wuchty, 2001).

The duplication of genes is an important factor in protein evolution. It has been argued that although most of the additional copies are silenced later, duplication is the main source for speciation (Lynch and Conery, 2000). Having two copies of a gene gives way for one copy to mutate freely while the original function is still preserved by the other copy. Protein evolution on the domain level is believed to take place through fusion and fission events. Single-domain proteins are fused to multi-domain proteins; multi-domain proteins are separated through fission. The domain families that occur together with many other domain families are duplicated more often than expected by chance (Apic *et al.*, 2003). From the correlation between versatility and abundance it has been concluded that domain recombination occurs randomly (Vogel *et al.*, 2005). Single domains do not always recombine independently. Combinations of two or three domains can act as evolutionary units, supra-domains. The domains within this unit have a particular functional and spatial relationship (Vogel *et al.*, 2004b).

The rate of fusion and fission events has been studied for 131 completely sequenced species (Kummerfeld and Teichmann, 2005). The authors concluded that fusion was four times more common than fission. It has been found that domain architectures are seldomly created de novo. Only 0.4 to 4% of the sequences from 62 species were considered to be involved in convergent evolution (Gough, 2005). This could also explain, the tendency of domains to occur in only one combination. Of all observed combinations of two domains A and B, the majority is solely found as AB; only 2% also exist in the inverted sequential order BA (Bashton and Chothia, 2002).

Another mechanism in protein evolution is the rearrangement of protein domains, which does not necessarily occur as a results of fusion/fision events. Besides fusion/fission, two other mechanism have been proposed (Ponting and Russell, 1995; Jeltsch, 1999). Using examples of cicular permutation of protein domains the frequency of the proposed mechanisms has been assessed (Weiner *et al.*, 2005; Weiner and Bornberg-Bauer, 2006). The authors found examples for all three mechanisms. However, the fusion/fission-mechanism seemed to be much more common.

1.3 Protein sequence analysis

A common task of protein sequence analysis is to determine the relatedness between a set of given sequences and to estimate the evolutionary distances between them. Such values are needed for the reconstruction of phylogenetic trees and homology analysis that provide the basis for more advanced methods. Unfortunately, non-homologous sequences can be very similar and are likely to be annotated as homologs by mistake (Spang and Vingron, 1998). However, the measured difference between two sequences always depends on the underlying evolutionary model.

Usually only substitutions of single sites (point mutations) are considered for estimating evolutionary distances. The substitution of a site is assigned a certain cost and the combination of the costs for all dissimilar sites gives the evolutionary distance. Insertion and deletions (indels) of sites are more cumbersome to process. One reason for this is the difficulty of assigning costs for indel events. It is also not clear how different indel lengths should be treated. Another problem in dealing with indels is that it is no longer obvious which sites from two sequences match. This task is in most cases solved by alignment algorithms based on dynamic programming as introduced by Needleman and Wunsch (1970) as well as Smith and Waterman (1981). When aligning a large number of sequences, dynamic programming is not feasible due to the complexity of $O(n^m)$ for m sequences of length n. Current multiple sequence alignment algorithms use a variety of methods to find suitable trade-offs between performance, quality and generality (Katoh *et al.*, 2005; Edgar, 2004; Lee *et al.*, 2002a; Notredame *et al.*, 2000; Subramanian *et al.*, 2005; Lassmann and Sonnhammer, 2002).

1.3.1 Simple models of evolution

Once two sequences are aligned, the evolutionary distance can be determined by examining the dissimilar sites (excluding indels). The most simple way to measure evolution is by calculating the p distance (see Nei and Kumar, 2000). It is defined as the proportion

$$\hat{p} = n_d/n,$$

where n is the total number of sites and n_d the number of dissimilar sites. One major disadvantage is that the p distance does not account for multiple substitutions at one site. Mulitple substitutions cannot be discovered retrospectively, as the real evolutionary distance will be bigger than the measured difference. The Poisson corrected distance corrects the p distance to

$$\hat{d} = -\ln(1-\hat{p}).$$

The evolution of deoxyribonucleic acid (DNA) sequences has been modelled by Jukes and Cantor (1969). They assumed the probability of a site to mutate within some time unit to some given nucleotide to be α. The evolutionary distance is then given as

$$\hat{d} = -\frac{3}{4}\ln\left(1-\frac{4}{3}\hat{p}\right).$$

Other models for DNA evolution use different rates for transitional and transversional nucleotide subsitution (Kimura, 1980) or GC content as the HKY model (Hasegawa *et al.*, 1985). The latter two models are designated to DNA sequences only. The Jukes-Cantor model can be adapted to amino acid sequences. It suffices

to exchange the original factor $\frac{3}{4}$ with $\frac{19}{20}$ (Takezaki *et al.*, 1995), leading to

$$\hat{d} = -\frac{19}{20} \ln \left(1 - \frac{20}{19} \hat{p} \right).$$

1.3.2 Models based on collected sequence data

Assuming the substitution probability α for all transitions from one amino acid to another is very ad hoc and does not model reality well. Amino acids are often substituted with an amino acid of similar biochemical properties, e.g. polarity, volume (Dayhoff, 1972). An evolutionary model can be based on the actually observed substitutions on multiple sequence alignments. Dayhoff *et al.* (1978) used data from 71 families of closely-related proteins. Within each family, two sequences differed at the most in 15% of the sites. From this data, they extrapolated substitutions for larger evolutionary distances using a Markov chain model. Dayhoff and co-workers modelled evolution in two steps. At first, amino acid substitutions occur as a result of a mutated in the underlying DNA sequence. In a second step this change is accepted by evolution if it is advantageous or at least not harmful. The evolutionary distance of two aligned protein sequences is measured as Percent Accepted (point) Mutation (PAM). An evolutionary distance of 150 PAM corresponds to 1.5 substitutions per site on average. Since multiple substitutions will be experienced at some sites, two sequences of a distance of 150 PAM will still have some preserved sites. In fact, even for 250 PAM 20% of the positions remain unchanged. The crux of evolutionary distances is that multiple substitutions cannot be observed directly and need to be estimated.

Evolution modeled as Markov chain

The Dayhoff model regards evolution as a Markov chain with the typical property that only the current state determines the probability distribution of the next transition. Let $X = X(t) \geq 0$ be a familiy of probability variables. The Markov property states that a process X is a Markov chain, if

$$\mathbb{P}[X(t_n) = j | X(t_1) = i_1, \ldots, X(t_{n-1}) = i_{n-1}] = \mathbb{P}[X(t_n) = j | X(t_{n-1}) = i_{n-1}],$$

$\forall$ states $j, i_1, \cdots, i_{n-1}$, $\forall$ points in time $t_1 < t_2 < \ldots < t_n$. The Dayhoff model assumes time homogeneity for the process, i.e. the transition probability $\mathbb{P}[X(s + t) = j | X(s) = i]$ is independent of time s (see Müller, 2001; Ewens and Grant, 2001). The transition probability matrix $P(t) = (p_{ij}(t))$ denotes all transition probabilities after time t has passed. Each matrix element is given as $p_{ij}(t) = \mathbb{P}[X(s + t) = j | X(s) = i]$. The transition probability matrix $P(t)$ has some important properties:

- $P(0) = I$

- $p_{ij}(t) \geq 0$
- $\sum_j p_{ij}(t) = 1$
- $P(s+t) = P(s)P(t) \quad \text{for } s,t \geq 0$

With the differentiability assumption this renders possible to calculate the rate matrix Q as the limes

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

The rate matrix Q can be used to easily compute the transition probability matrix $P(t)$ for all $t \geq 0$

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

In the long run the model will have the same distribution of amino acid occurance independent of the distribution at the start. This stationary distribution is called π, with $\lim_{t \to \infty} p_{ij}(t) = \pi_j > 0$ the frequency of amino acid j independent of i. The property of the stationary distribution is that it is not affected by evolution:

$$\forall t \geq 0: \quad \pi P(t) = \pi \quad \text{and} \quad \pi Q = 0$$

A common task is to assign a similarity measure to two amino acids. One application is the alignment to sequences where it is essential to know which sites are more related and are thus more likely to have developed from a common ancestor. The similarity of amino acids is measured as log-odds score (or lods score). This is a comparison of two models. The first model M assumes a common ancestor from which the current amino acids have developed. The second model U states that independent evolution has led to the two amino acids. The score $s(i,j)$ for the two amino acids i and j is in general given as

$$s(i,j) = \log \frac{P(i,j|M)}{P(i,j|U)}$$

and for the Dayhoff model for some evolutionary distance $t \geq 0$ as

$$s(i,j|t) = \log \frac{p_{ij}(t)}{\pi_j}.$$

Despite the popularity of its overall approach, the Dayhoff substitution matrix suffers from two major problems. The data was gathered from a few closely-related proteins (up to 17 PAM distance). It is questionable whether the substitutions observed between more distantly-related sequences are just a magnification of the substitutions seen on closely-related proteins (Gonnet *et al.*, 1992). Additionally, possible small errors in the data can become severe when extrapolated to larger

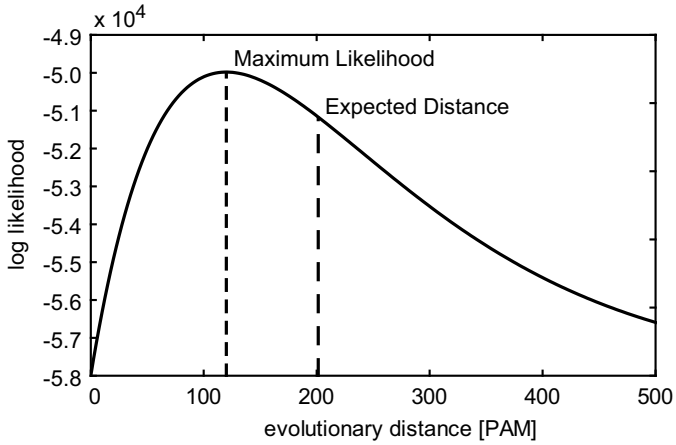

Figure 1.3: Typical distribution of the likelihood function $\mathcal{L}(t|F, Q, \mathbb{A})$. The Maximum Likelihood gives the single most likely value. The Expected Distance integrates over all likelihoods. For a typical likelihood distribution, this gives higher values for the Expected Distance estimator.

distances. Subsequent substitution matrices have tried to overcome these problems. Jones *et al.* (1992) used a larger dataset from many different proteins. Special adaptions also exist for transmembrane proteins (Jones *et al.*, 1994; Ng *et al.*, 2000). Müller and Vingron (2000) used a resolvent method (Ma and Röckner, 1992) to calculate substitution matrices. For this, they considered alignments between 1 and 300 PAM from the SYSTERS database (Krause and Vingron, 1998). Whelan and Goldman (2001) used a Maximum Likelihood phylogenetic trees instead of the Maximum Parsimony approach chosen by Dayhoff *et al.* and Jones *et al.*

Evolutionary distances from an alignment of two sequences can be estimates with the Dayhoff model in two different ways. The Maximum Likelihood estimates gives the distance of the substitution matrix that fits the alignment differences best. The Expected Distance estimator integrates over the whole range of likelihoods and in most cases gives higher estimates than the more popular Maximum Likelihood method (Agarwal and States, 1998). The likelihood for the distance t given the alignment $\mathbb{A}$, the rate matrix Q and the diagonal matrix F with statonary distribution values π is given as

$$\mathcal{L}(t|F, Q, \mathbb{A}) = \sum_{i,j} n_{ij} \log(Fe^{tQ})_{ij}.$$

The data matrix $N = (n_{ij})$ is generated from the alignment. The entry n_{ij} denotes the number of aligned amino acid pairs i and j. The Maximum Likelihood estimate $\hat{t}_{ML}$ is obtained as the solution of

$$0 = \frac{d}{dt}\mathcal{L}(t|F, Q, \mathbb{A}) = \sum_{i,j} n_{ij} \frac{d}{dt} \log(Fe^{tQ})_{ij}.$$

The Expected Distance is the result of integration of the whole range of likelihood estimates

$$\hat{t}_{ED} = \int t\, \mathcal{L}(t|F, Q, \mathbb{A})\, dt.$$

The BLOSUM evolutionary model

All the above mentioned models are based on a Markov chain and include the assumption of evolutionary time t. The Blosum score matrices come with a lighter statistical load. The BLOCKS database (Henikoff and Henikoff, 1991) comprises aligned, ungapped regions of homologous sequences. From this database clusters of sequences above some level of identity $L\%$ were derived. The frequency of observing the amino acid i in one cluster aligned to amino acid j in another cluster was calculated as A_{ij} correcting for the cluster size. The frequency of each amino acid is then obtained as

$$q_i = \frac{\sum_j A_{ij}}{\sum_{kl} A_{kl}}.$$

The fraction of pairing between the amino acids i and j is expressed as

$$p_{ij} = \frac{A_{ij}}{\sum_{kl} A_{kl}}.$$

These two estimates are enough to calculate the score as

$$s(i, j) = \log \frac{p_{ij}}{q_i q_j},$$

that constitute the BLOSUM matrices (Henikoff and Henikoff, 1992). Different L values give different score matrices. Contrasting to the Dayhoff evolutionary model the BLOSUM concept lacks some mathematical relationship between different score matrices. Still, matrices for high L values vaguely correspond to a short evolutionary time span. The BLOSUM62 matrix is widely used for alignment and database search purposes thanks to it general applicability. However, the BLOSUM50 matrix has been proposed for gapped alignments (Pearson, 1996). Since the BLOSUM matrix has no rate matrix, evolutionary distance estimation is not completely straightforward. One approach is to estimate a rate matrix from the BLOCKS database (Veerassamy *et al.*, 2003), thus bypassing the existing BLOSUM

matrices. A second approach is to calibrate the BLOSUM score matrix to fit the Dayhoff model (**Paper V**).

The detection of homology between protein sequences is a task that requires a model of protein evolution in order to distinguish related from unrelated sequences. Choosing score matrices built from the Dayhoff or the BLOSUM model is a natural choice for this. The Basic Local Align Search Tool BLAST (Altschul *et al.*, 1990) uses the BLOSUM62 matrix as default. BLAST seeks databases of either DNA or protein sequences for similar sequences according to the given score matrix. A stochastic model determines the significance, E-value, of the found matches (Karlin and Altschul, 1990).

The orginal published version supported only ungapped alignments. However later publications describe the inclusion of gaps in the theory (Altschul *et al.*, 1997). BLAST is often used to find homologs based on the whole sequence of a given protein. A different approach is to highlight protein domains and build homology detection on domain prediction results (**Paper IV**).

1.4 Phylogenetic trees

Phylogenetics (φύλο race; γεννήση birth) strives to display and explain the evolutionary relatedness of organisms. The descendence of sequences of whole species is often represented by phylogenetic trees that display similarities and differences between a selected set of attributes or characters.

A graph theoretical definition of a tree is a graph $G = (V, E)$ in which for every pair of vertices $v_1, v_2 \in V$ there is one unique path in G from v_1 to v_2. A path from v_1 to v_2 is a sequence of distinct vertices $v_1, v_2, \ldots, v_k$ such that for all $i \in \{1, 2, \ldots, k-1\}$ the edge $e = \{v_i, v_{i+1}\} \in E$. The number of vertices one vertex v is connected with is called degree of v, $d(v)$. Many applications in phylogeny focus on binary trees, that only consist of leaves of degree 1 and interior vertices of degree 3. Additionally they may have one additional node of degree 2, called root (see Semple and Steel, 2003).

The leaves of phylogenetic trees are labelled with elements from the set of sequence or species names X. Interior vertices typically remain unlabelled. A set of $n \geq 3$ sequences can give rise to $1 \times 3 \times 5 \times \ldots \times (2n-5) = (2n-5)!!$ different unrooted or $1 \times 3 \times 5 \times (2n-3) = (2n-3)!!$ different rooted trees. The range of tree shapes from one set of sequences may vary substantially. Two trees can be compared by examining which vertices are connected by edges. An edge can be represented by a bipartition that shows the two sets of leaves that would be obtained if the edge was removed, thus splitting the graph into two connected units. The bipartition is called X-split and denoted as $A|B$ with A,B as two non-empty, non-overlapping subsets of the set of sequences $A, B \subset X$, $A \cap B = \emptyset$, $A \cup B = X$ (Buneman, 1971). The similarity of two trees with the label set X can be measured by the Robinson-Foulds topological distance (RF distance) that compares the X-splits (Robinson

and Foulds, 1981). The distance equals the minimum number of elementary tree operations (merging and splitting of nodes) needed to transform one tree into the other. It is also equal to two times the number of dissimilar X-splits between the two trees. The RF distance ranges between 0 for identical trees and $4n - 6$ if all leaves are positioned differently.

Phylogenetic tree construction algorithms for sequences fall into two classes. The first ones analyse alignments of the DNA or protein sequences directly. Examples of such approaches are Maximum Likelihood or Maximum Parsimony algorithms. Methods of the second class use pairwise distances as input data. These pairwise distances can be obtained from the estimators as discussed in the previous section. Algorithms from the second class are very popular thanks to their speed. They are much faster than members of the first class while retaining an acceptable quality (see Nei and Kumar, 2000).

1.4.1 Maximum Likelihood and Maximum Parsimony

Maximum Likelihood (ML) is often considered the best approach (Zhang and Nei, 1997). Here, each tree topology is assigned a likelihood, summing over all possible ancestral sequences (Felsenstein, 1981; Kishino *et al.*, 1990). This is repeatedly carried out for all tree topologies and the tree with the highest likelihood is finally chosen. The major drawback of maximum likelihood is its poor scalability. Already with a small number of sequences, it becomes unfeasible to examine every possible tree topology.

The Maximum Parsimony (MP) approach is a general concept and can be used for various tasks in sequence analysis. It is frequently justified by "Ockham's razor", which states that the most simple solution should always be chosen. In the context of phylogenetic trees, Maximum Parsimony can either be used to assign a cost to a given tree or it can search through the tree space to find the tree with the lowest costs. Phylogenetic trees need not neccesarily to be constructed from DNA or protein sequences, but may also include other data (Fitch; Sankoff and Cedergren, 1983). The Maximum Parsimony can be used to predict ancestral sequences at inner nodes (**Paper VII**).

Algorithms based on the Maximum Likelihood or Maximum Parsimony principle are computationally very expensive if they are used to inspect the whole tree space (Foulds and Graham, 1982). Even today's available speed improved versions (Ronquist and Huelsenbeck, 2003; Huelsenbeck and Ronquist, 2001; Schmidt *et al.*, 2002; Swofford, 1996) are only suitable for alignments of few sequences (Williams and Moret, 2003).

1.4.2 Phylogenetic trees from pairwise distances

One of the oldest and most simple algorithms among the distance methods is

the unweighted pair group method using arithmetic averages (UPGMA), originally developed by Sokal and Michener (1958). Actually, UPGMA is more a cluster method than a tree construction algorithm. The idea is to aggregate in each step the two closest clusters or single objects. Subsequently, distances between the new cluster and all other objects have to be calculated. UPGMA assigns the same weight to all objects by summing over all distances in the cluster and multiplying by the cluster's cardinality. All leaves of the resulting tree are equidistant to the root. This kind of tree is called ultrametric (see Semple and Steel, 2003). Applied to protein sequences this assumes a constant mutation rate for all proteins and all species. In most cases this molecular clock assumption is wrong. Particularily sequences from distant-related species or genes under a high selective pressure are likely to have varying mutation rates. In such cases the UPGMA tree can be very misleading, sharing not one single non-trivial X-split with the correct tree (see Durbin *et al.*, 1998).

UPGMA trees are fast and easy to generate, but do not prove suitable if the evolutionary rate is variable. The validity of the molecular clock hypothesis originally attributed to Zuckerkandl is still heavily debated. It has been assumed that essential genes which are exposed to a high evolutionary pressure would evolve slower than nonessential proteins. This has been reported for bacteria (Jordan *et al.*, 2002), however it could not be verified for eukaryotes (Hurst and Smith, 1999; Hirsh and Fraser, 2001).

The neighbour-joining algorithm (NJ) developed by Saitou and Nei (1987) can even be applied if the evolutionary rate is not fixed. It is not built around the assumption of a molecular clock, rather it assumes additivity. Additivity states that the distance between all pairs of leaves is given by the sum of edge lengths of the path between them. The neighbour-joining algorithm seeks the closest neighbour to a sequence. It does not just pick the sequence with the shortest distance but evaluates the surrounding of the other sequence. If the pairwise lengths are additive, neighbour-joining guarantees to find the correct tree. Pairwise distances computed from real alignments are seldom fully additive. This does not hinder the use of NJ and it often returns a tree very similar to the correct tree. Atteson (1997) showed that if the error of the distance estimates is at most half the length of the shortest branch in the underlying phylogeny, then NJ always returns the correct tree. It has been demonstrated that the topology of the NJ tree is close to that of the Minimum Evolution (ME) tree (Saitou and Imanishi, 1989). Fast versions of NJ have been published (Howe *et al.*, 2002; Mailund and Pedersen, 2004), in which heuristics are used to avoid unnecessary recomputations.

Several modifications have been applied to the original NJ algorithm. One of them is BIONJ that has been proposed by Gascuel (1997). In standard NJ the interior node has the same distance to the joined nodes. BIONJ uses a simple model of the sampling noise (variance) of evolutionary distances. During each step of the clustering process, nodes are selected for joining so that the variance of the new distance matrix is minimized. It thus takes into account the fact that long

distances present a higher variance than short ones. Another modification of NJ is the Weighbor algorithm, or "weighted neighbor joining" (Bruno *et al.*, 2000). Here the selection of nodes to join is based on additivity and positivity properties, which are estimated using Maximum Likelihood. It has been reported to achieve tree accuracies comparable to exhaustive ML, yet at much lower computational costs. Besides NJ and its various modifications, other attempts to fulfill the ME criterion have been taken by Desper and Gascuel (2002). Their Greedy Minimum Evolution algorithm is used to calculate a tree, which is further improved by Nearest Neighbor Interchange. The authors presented unweighted and weighted versions of their approach, both implemented in their program FastME. In a large study of distance methods BIONJ showed the highest general accuracy. However, the accuracy differed only little between the evaluated methods (**Paper VI**).

The significance of phylogenetic trees can be measured by bootstrapping (Felsenstein, 1985; Efron and Tibshirani, 1993). Bootstrapping generates a new alignment from the orginal alignment by randomly picking columns with replacement. The phylogenetic tree obtained from this alignment is compared to the original tree. For each shared X-split, the bootstrap value for the X-split is increased. This procedure is typically repeated at least 100 times. The final result is the original tree where each branch is annotated with the bootstrap value. These values can be interpreted as a support for the branches. Bootstrapping only checks on the tree topology and not the branch lengths.

1.5 Orthology inference

The knowledge of orthologous relationships is important for analysing proteins of unknown function. Already annotated proteins of closely-related species can hint at starting points for further experimental examinations. Orthologs (which includes inparalogs as well) are more likely to have preserved a similar function than outparalogs. It is therefore important to separate in- from outparalogs during orthology inference.

The definition of orthology is usually illustrated on a phylogenetic tree. Hence, the usage of phylogenetic trees for orthology assignment is an obvious option, yet not the most commonly used one. Sequence similarity has instead been a much employed approach. This choice has been influenced by computing power restrictions. Other methods use the conservation of gene loci structure (synteny) for orthology inference (Wheeler *et al.*, 2006; Blake *et al.*, 2006).

1.5.1 Orthologs from phylogenetic trees

Orthostrapper (Storm and Sonnhammer, 2002) takes alignments from two groups of species plus alignment data from an outgroup of more distantly related sequences.

By using the outgroup, Orthostrapper is able to exclude outparalogs from the analysis. The Orthostrapper algorithm can identify orthologs and compute a confidence value for each orthology assignment. The confidence value is calculated by bootstrapping (Efron *et al.*, 1996). The program can be accessed and results can be displayed using the OrthoGUI program (**Paper I**).

The RIO algorithm uses a similar approach to determine orthology confidence values (Zmasek and Eddy, 2002), but is aimed at finding all orthologs to one sequence rather than all orthologs between two species. A large data analysis has been carried out using the Orthostrapper algorithm, resulting in the HOPS database (Storm and Sonnhammer, 2003).

The TreeFam database stores manually curated phylogenetic trees with orthology assignments (Li *et al.*, 2006). Data from the PhIGs database were used to get seed clusters (Dehal and Boore, 2006) which were expanded and subjected to algorithms for inferring duplication and speciation nodes.

1.5.2 Orthologs from pairwise comparisons

Similarity-based orthology inference for two species starts with a pairwise comparison of all proteins from one species and all proteins of some other species. The BLAST program has become the standard tool for sequence similarity (Altschul *et al.*, 1997). It assembles for each protein a list of similarity scores for all proteins from the other species. The protein that receives the highest score in terms of Reciprocal BLAST Hits (RBS) is considered an orthology candidate. The hypothesis is that orthologs should score higher than outparalogs as these were separated from each other earlier. This assumes a fixed rate of evolution for all proteins, which is not necessarily true. It has been shown that the highest scoring protein is sometimes not the nearest phylogenetic neighbour of the query sequence (Koski and Golding, 2001).

The COG (Clusters of Orthologous Groups) database was the first large-scale attempt to contruct a databse of orthologs (Tatusov *et al.*, 1997, 2000, 2001). The COG database currently contains data from 66 species. A eukaryotic addition to COG named KOG that holds data from seven eukaryotic species has been published as well Tatusov *et al.* (2003). A drawback of the COG/KOG clusters is that they often list outparalogs by mistake. The algorithm behind the COG/KOG database has also been used to generate the EGO database (Lee *et al.*, 2002b). Here, DNA sequences were used as input data to build the database.

The InParanoid database (Remm *et al.*, 2001; O'Brien *et al.*, 2005) is a collection of orthologs and inparalogs. If two genes in one species show a higher similarity than the ortholog in another species, the two genes are considered inparalogs. If they are less similar than the ortholog they are considered outparalogs. All RBS orthologs and inparalogs are assigned confidence values. In a recent study, InParanoid was found to be the best orthology inference program for the identification

of functionally equivalent proteins (Hulsen *et al.*, 2006). The OrthoMCL database uses a similar approach as the InParanoid database but builds orthologous groups of multiple species by using a Markov clustering algorithm (Li *et al.*, 2003; Chen *et al.*, 2006).

PhIGs is a graph-based method for orthology inference (Dehal and Boore, 2006). The algorithm follows known phylogenetic relationships to aggregate orthologs. At each internal node of the phylogenetic tree, ortholog clusters were calculated. The database currently holds data from 23 opisthokonts and 11 chordates.

Chapter 2

Results

Die Wissenschaft, sie ist und bleibt,
was einer ab vom andern schreibt.
Eugen Roth (1895-1976)

Paper I – OrthoGUI: graphical presentation of Orthostrapper results

The Orthostrapper algorithm (Storm and Sonnhammer, 2002; Storm, 2004) is a tree-based approach for orthology inference. The sequences within a multiple alignment are labeled as two (groups of) species and input into Orthostrapper. The algorithm builds a phylogenetic tree from the alignment and traverses the tree seeking for monophyletic subtrees. The output is a matrix of orthology reliability based on bootstrapping (Felsenstein, 1985; Efron *et al.*, 1996).

The Orthostrapper program is a Java command line tool but requires the presence of other executables as Belvu (Sonnhammer, 2006). To relieve users from these technical details, make Orthostrapper ubiquitously accessible and present the results in a human-readable fashion, we developed a graphical user web interface. OrthoGUI is a user-friendly interface which combines web access to Orthostrapper with a graphical presentation of the results.

The OrthoGUI homepage automates the Orthostrapper pipeline and guides the user smoothly through the analysis. The Orthostrapper result matrix is processed and orthologs are clustered based on the average linkage method. OrthoGUI displays the coloured clusters within a matrix as well as on a phylogenetic tree. The resulting matrix can also be exported to a local file via the browser.

Paper II – The Pfam protein families database

The Pfam database comprises a large number of protein families and domains. Seed alignments for each Pfam-A family are manually curated and used for generation of profile-HMMs (Eddy, 1998, 1996; Durbin *et al.*, 1998). Subsequently, the UniProt database (Bairoch *et al.*, 2005) is exhaustively sought for the Pfam-A domains, resulting in the full protein domain alignments. The domain families are annotated with literature references and links to popular databases (Murzin *et al.*, 1995; Mulder *et al.*, 2005; Hulo *et al.*, 2006). The manually curated Pfam-A domains are supplemented by the automatically generated Pfam-B domains.

Based on the Pfam domain predictions, the domain architecture of the proteins can be studied. The domain query function enables searching for a given domain architecture. Domains can be freely arranged and gaps between domains may be specified. The result is a list of all proteins that share the query architecture.

Paper III – Pfam: Clans, Web Tools and Services

The Pfam database has been enhanced with clans, a hierarchy level to group similar domain families. Some previous families have been split into several families that are subsumed into clans. The reason for this measure is that it proved infeasible for certain domain families to build a profile-HMM that catches all members of the domain family yet does not overlap with other entries. In addition, several new domain families have been incorporated in this release of the Pfam database. All tools have been updated to account for the changes. Additional web services have been integrated into the database to offer automated searches to external researchers.

Paper IV – PfamAlyzer: Domain-Centric Homology Search

Homology search aims to identify proteins with a common evolutionary descent. Traditional methods have treated all sequence sites equally and are thus sequence-centric. This paper introduces domain-centric homology search as a complement to sequence-centric homology search. It seems particularily suited for searching distant homologs. The idea behind domain-centric homology search is to take advantage of the fact that domain sites are much more important to a protein's function than non-domain sites. Usually, domain-sites are more stable during evoluton than non-domain sites and should be given priority when identifying distant homologs. It has been verified that convergent evolution of domain architectures is rare (**Paper VII**; Gough, 2005). The potential risk of type I errors is therefore low. We have studied the limitations of the traditional sequence-centric homology search. The results show that up to 16% of proteins with the same domain architecture are missed by BLAST homology search.

Traditonal homology search is typically carried out in one step with the BLAST tool (Altschul *et al.*, 1997). Domain-centric homolgy search uses two phases. At first, the given sequence is analysed for protein domains. We use the hmmer software package (Eddy, 2006) to examine the given sequence for Pfam domains. The outcome of this step is a domain architecture, the sequence of protein domains that have been found on the protein sequence. The second phase searches the Pfam-annotated UniProt (Bairoch *et al.*, 2005) database for proteins with the same domain architecture.

The PfamAlyzer application has been created to enable easy-to-use domain-centric homology searches. The user can analyse a protein sequence for Pfam domains and use the found domain architecture to further search UniProt. PfamAlyzer provides means for seeking specific domain architectures within Pfam. Arbitrary domains can be combined freely, optionally with gaps. The query may be limited to taxonomic groups. Results are displayed either in a list-fashion or as species distribution where the sequences are shown as leaves on a phylogenetic tree according to their origin. This allows exploring domain recombination and studying the spread of domain architectures within different species.

Paper V – *Scoredist*: A robust protein sequence distance estimator based on the BLOSUM scoring matrices

Pairwise evolutionary distances of amino acid sequences are frequently applied in many areas of Bioinformatics. Virtually all current high-troughput phylogenetic tree reconstruction algorithms use pairwise distances as input (Saitou and Nei, 1987; Gascuel, 1997; Desper and Gascuel, 2002; Bruno *et al.*, 2000). Multiple subsitutions occurring at one site are the major hinderance in determining the exact evolutionary distance. Eventually, a mutated site may change back to the original amino acid leaving no traces behind and this cannot be discovered later. Statistics has to be applied to estimate evolutionary time from the observed alignment.

The evolutionary distance between protein sequences is commonly measured in Percent Accepted (point) Mutation (PAM) (Dayhoff *et al.*, 1978). Either the original matrix series by Dayhoff *et al.* (1978) or subsequent series that follow the same principle are commonly used (Müller and Vingron, 2000; Jones *et al.*, 1992; Whelan and Goldman, 2001). The evolutionary distance is estimated by looking in the matrix series for the transition probability matrix that explains the observed differences most accurately. The optimal matrix can be found either by an iterative search for the Maximum Likelihood matrix, or by integration to find the Expected Distance (Agarwal and States, 1998). The drawback with both Maximum Likelihood and Expected Distance is the computational complexity. Other methods that only apply some correction to number of observed differences are known to be faster but less accurate.

We developed a correction-based protein sequence estimator called *Scoredist*. It uses a logarithmic correction of observed divergence based on the alignment score according to the BLOSUM62 score matrix (Henikoff and Henikoff, 1992). We evaluated *Scoredist* and a number of optimal matrix methods using three evolutionary models for both training and testing Dayhoff (1978), Jones-Tylor-Thornton (1992), and Müller-Vingron (2000), as well as Whelan and Goldman (2001) solely for testing. Test alignments with known distances between 0.01 and 2 substitutions per position (1-200 PAM) were simulated using ROSE (Stoye *et al.*, 1998). *Scoredist* proved as accurate as the optimal matrix methods, yet substantially more robust. When trained on one model but tested on another one, *Scoredist* was nearly always more accurate. The Jukes-Cantor (1969) and Kimura (1983) correction methods were also tested, but were substantially less accurate. The *Scoredist* distance estimator is fast to implement and run, and combines robustness with accuracy. *Scoredist* has been incorporated into the Belvu (Sonnhammer, 2006) alignment viewer.

Paper VI – Assessment of Protein Distance Measures and Tree-Building Methods for Phylogenetic Tree Reconstruction

The construction of phylogenetic trees finds many applications in current research. It is a means to address evolutionary questions as observed in taxonomy or protein function inference. In molecular epidemiology, e.g., phylogenetic trees have been used to study the evolution of HIV (Kalish *et al.*, 2004) and may support future vaccine design.

Maximum Likelihood phylogeny inference is generally believed to produce the most accurate trees. Each topology is assigned a likelihood, summing over all ancestral sequences possible (Felsenstein, 1981; Kishino *et al.*, 1990). This is repeatedly carried out for all topologies and the tree with the highest likelihood is finally chosen. The major drawback of Maximum Likelihood is its poor scalability. Already with a small number of sequences, it becomes infeasible to examine every possible tree topology (Williams and Moret, 2003).

Distance-based methods have gained major importance and are today clearly dominating other approaches. Their popularity is due to being statistically consistent in all settings (Desper and Gascuel, 2004) as well as outperforming other methods by far in terms of speed. Applying distance-based approaches, tree reconstruction is thus conducted in two separate steps. First, pairwise distances are estimated for all sequences. Tree building is then based on the obtained pairwise distances. Previous studies have either compared performance of tree construction or distance methods (Russo *et al.*, 1996; Desper and Gascuel, 2004). However, the best distance measure with one tree method does not necessarily have to turn out as the right choice for some other tree algorithm. Additionally, real data do always need to pass distance estimation and tree reconstruction steps. We evaluated combinations of distance measures and tree construction methods and tested their ability to reconstruct correct trees.

Neighbour-joining (Saitou and Nei, 1987) is the most popular distance-based method. A number of variants of the standard algorithm have been proposed. We used the programs BIONJ (Gascuel, 1997), FastME (Desper and Gascuel, 2002), Weighbor (Bruno *et al.*, 2000), and standard neighbour-joining in combination with *Scoredist* (**Paper V**) and 11 other distance estimators (Jukes and Cantor, 1969; Dayhoff *et al.*, 1978; Kimura, 1983; Müller and Vingron, 2000; Jones *et al.*, 1992; Whelan and Goldman, 2001). These were evaluated on a test set based on real trees taken from 100 Pfam families (**Paper II**). Each tree was used to generate multiple sequence alignments with the ROSE (Stoye *et al.*, 1998) program using three evolutionary models. The accuracy of the methods was analysed with a modified version of the topology measure given by Robinson and Foulds (1981).

We found that BIONJ produced the overall best results, although the average accuracy differed little between the tree building methods (normally less than a percent). A noticeable trend was that FastME performed poorer than the rest on

long branches, which has also been reported by others (Bruno, 2005). Weighbor was several orders of magnitude slower than the other programs. Larger differences were observed when using different distance estimators. Jukes-Cantor and Kimura distance correction produced clearly poorer results than the other methods, even worse than uncorrected distances. Despite its computational simplicity, the *Scoredist* distance estimator was one of the best distance methods.

Paper VII – Gene Tree based Analysis of Domain Architecture Evolution

Protein domains are recombined throughout evolution to acquire new function (Murzin *et al.*, 1995; Rossmann *et al.*, 1974; Jaenicke, 1987). This work addressed the evolutionary processes that govern domain recombination. One approach to understanding the complex processes is the study of convergent evolution. We identified and analysed cases of convergent evolution of protein domain architectures. Previous approaches focused on the domain architectures only (Kummerfeld and Teichmann, 2005). This study considered the protein domain alignments. The outcome is a finer grained analysis that even allows the discovery of convergent evolution of domain architectures within the same species.

The algorithm outlined in this paper infers ancestral domain architecture given the phylogenetic trees from the protein domain families. As it is akin to traditional Maximum Parsimony algorithms (Semple and Steel, 2003), it processes a tree in two passes. The first pass starts at the leaf level and infers the least expensive ancestral architectures. Occasionally, it is not possible to make a decision if several potential ancestral architectures are attributed the same costs. The second pass starts at the root and aims to solve these cases.

The algorithm was appplied on domain-assigned sequences from 50 fully-sequenced genomes. The completeness of the domain assignments is of high importance to the study. Therefore, two datasets were formed. The *max50* data set required all N-, C-terminal or inter-domain sequence lengths to be below 50 acmino acids. The *nolimit* dataset did not embody these limitations. The phylogenetic trees were generated with QuickTree (Howe *et al.*, 2002) using the *Scoredist* (**Paper V**) distance estimator.

Previous studies could identify 27 cases of convergent domain architecture evolution (0.7% of the examined domain architectures) (Gough, 2005). We could confirm that convergent evolution is rare, yet more frequent than originally thought. About 1.9% of the architectures in the *max50* and 4.2% of the architectures in the *nolimit* were found to be examples of convergent evolution. Convergent evolution seems to be a random process. There was no bias towards certain domains or functions in the datasets.

Chapter 3

Discussion and Further Work

Indes sie forschten, röntgten, filmten, funkten,
entstand von selbst die köstlichste Erfindung:
der Umweg als die kürzeste Verbindung
zwischen zwei Punkten.
Erich Kästner (1899-1974)

It is in the nature of science that there always remain questions to be answered. New insights may cast doubt on what was previously believed conclusively solved. New genome sequencing projects enlarge our knowledge and supply many researchers with data for their investigations. This work has sought to further the understanding of protein domain architecture combinations. The task has been approached by examining existing methods and databases.

The implications the reliability of phylogenetic trees has on the conclusions in current research can hardly be overestimated. We have carried out a comprehensive study of existing phylogenetic tree reconstruction algorithms using real phylogenetic trees instead of artificially created ones. The results can hopefully advise scientists on chosing appropriate algorithms for their research projects. Currently, computational limitations in many cases only permit distance-based tree reconstruction. In fact, the protein distance estimation is the main bottle-neck and not the tree reconstruction algorithms itself. The *Scoredist* protein distance estimator has been developed with the aim for fast alternative to Maximum Likelihood and Expected Distance estimators. In our findings, *Scoredist* performed unmatched in speed and robustness. However, as technology advances and the need for low complexity loses importance, it may be possible to deploy more complex phylogenetic tree reconstruction algorithms, e.g. based on Maximum Likelihood, for large data sets as well.

Using data from completely sequenced genomes, we were able to show that convergent evolution on domain level is in fact more common than previously thought.

This was achieved using a Maximum Parsimony inspired ancestral architecture inference algorithm. The algorithm uses a tree topology of protein domains to predict ancestral architectures. By combining the results for all domains of a particular domain architecture introduces a confidence estimation into the predictions. The branch lengths of the phylogenetic trees are currently not considered. An enhanced algorithm could make use of this information as well and assign better confidence values. The functional aspects of domain recombination have not yet been fully examined. Recent pulications indicate that the sequence identity plays an important role in protecting proteins against misfolding (Wright *et al.*, 2005). This may lead to formulating a language of domain recombination which could be incorporated into the algorithm for an improved ancestral domain architecture inference.

Chapter 4

Acknowledgements

Nullum enim officium referenda gratia magis necessarium est.
Marcus Tullius Cicero (106 - 43 a.C.n.)

I am indebted to the numerous people who helped and supported me while carrying out the work on which this thesis is founded.

Erik Sonnhammer for welcoming me in your group, introducing me to science and your guidance and support over the years. Thank you for giving me the opportunity for an early start into the industry.

Anna Henricson for a splendid collaboration, nice discussions and your valuable comments on this thesis.

Lena Milchert and **Lars Arvestad** for a fruitful collaboration.

Christian Storm for helping me at the start, for taking so much time to explain things to me.

Carsten Daub for scientific discussions and other chats and for always being helpful.

Lukas Käll for sharing a room in Glasgow, discussions about so many topics and your help while preparing the thesis.

Abhiman Saraswathi for teaching me much about India and its delicious food, and for proofreading.

Markus Wistrand for discussions about various aspects of science, life and current politics.

Andrey Alexeyenko and **Timo Lassmann** for sharing an office with me.

everybody else at CGB who made my time there a pleasant experience.

Bibliography

Ehab Abouheif, Michael Akam, William J Dickinson, Peter W Holland, Axel Meyer, Nipam H Patel, Rudolf A Raff, V Louise Roth, and Gregory A Wray. 1997. Homology and developmental genes. *Trends Genet*, 13(11):432–433.

Pankaj Agarwal and David J States. 1998. Comparative accuracy of methods for protein sequence similarity search. *Bioinformatics*, 14(1):40–47.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. 1990. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Z Zhang, Webb Miller, and David J Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402.

Antonina Andreeva, Dave Howorth, Steven E Brenner, Tim J P Hubbard, Cyrus Chothia, and Alexey G Murzin. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*, 32(Database issue):226–229.

Gordana Apic, Julian Gough, and Sarah A Teichmann. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol*, 310(2):311–325.

Gordana Apic, Wolfgang Huber, and Sarah A Teichmann. 2003. Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *J Struct Funct Genomics*, 4(2-3):67–78.

K Atteson. 1997. *The performance of the NJ method of phylogeny reconstruction*, volume 37 of *Mathematical Hierarchies and Biology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 133–147. American Mathematical Society, Providence.

Amos Bairoch, Rolf Apweiler, Cathy H Wu, Winona C Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J Martin, Darren A Natale, Claire O'Donovan, Nicole Redaschi,

and Lai-Su L Yeh. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 33(Database issue):154–159.

Matthew Bashton and Cyrus Chothia. 2002. The geometry of domain combination in proteins. *J Mol Biol*, 315(4):927–939.

Judith A Blake, Janan T Eppig, Carol J Bult, James A Kadin, and Joel E Richardson. 2006. The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res*, 34(Database issue):562–567.

Erich Bornberg-Bauer, Francois Beaussart, Sarah K Kummerfeld, Sarah A Teichmann, and January 3rd Weiner. 2005. The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci*, 62(4):435–445.

William J Bruno. 2005. URL `http://www.t10.lanl.gov/billb/weighbor/fastme/`.

William J Bruno, Nicholas D Socci, and Aaron L Halpern. 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol Biol Evol*, 17(1):189–197.

Peter Buneman. 1971. *The recovery of trees from measures of dissimiarity*, pages 387–395. Mathematics in the Archaeological and Historical Sciences. Edinburgh University Press, Edinburgh.

Feng Chen, Aaron J Mackey, Christian J Jr Stoeckert, and David S Roos. 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*, 34(Database issue):363–368.

Cyrus Chothia. 1992. Proteins. One thousand families for the molecular biologist. *Nature*, 357(6379):543–544.

Charles Darwin. 1859. *On the origin of species my Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.

Charles Darwin. 1872. *On the origin of species*, 6th edition. John Murray, London.

Margaret O Dayhoff. 1972. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Springs.

MO Dayhoff, RM Schwartz, and BC Orcutt. 1978. *A model of Evolutionary Change in Proteins*, volume 5 supplement 3 of *Atlas of Protein Sequence and Structure*, pages 353–352. National Biomedical Research Foundation, Silver Springs.

Fernando de la Cruz and Julian Davies. 2000. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol*, 8(3):128–133.

Paramvir S Dehal and Jeffrey L Boore. 2006. A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics*, 7 (1):201.

Richard Desper and Olivier Gascuel. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J Comput Biol*, 9(5): 687–705.

Richard Desper and Olivier Gascuel. 2004. Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Mol Biol Evol*, 21(3):587–598.

Russell F Doolittle. 1995. The multiplicity of domains in proteins. *Annu Rev Biochem*, 64:287–314.

Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological sequence analysis*. Cambridge University Press, Cambridge.

Sean Eddy. 2006. Hmmer: profile hmms for protein sequence analysis. URL `http://hmmer.wustl.edu/`.

Sean R Eddy. 1996. Hidden Markov models. *Curr Opin Struct Biol*, 6(3):361–365.

Sean R Eddy. 1998. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763.

Robert C Edgar. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113.

B Efron and R Tibshirani. 1993. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York.

Bradley Efron, Elizabeth Halloran, and Susan Holmes. 1996. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci U S A*, 93(23):13429–13434.

Warren J Ewens and Gregory R Grant. 2001. *Statistical Methods in Bioinformatics*. Springer-Verlag, New York.

Joseph Felsenstein. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol*, 17:368–376.

Joseph Felsenstein. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(5):783–791.

Walter M Fitch. Towards defining the course of evolution: minimum change for a specified tree topology. *Systematic Zoology*, 20:406–416.

Walter M Fitch. 1970. Distinguishing homologous from analogous proteins. *Syst Zool*, 19(2):99–113.

Walter M Fitch. 2000. Homology a personal view on some of the problems. *Trends Genet*, 16(5):227–231.

LR Foulds and RL Graham. 1982. The steiner problem in phylogeny is np-complete. *Adv in Appl Math*, 3:43–49.

Olivier Gascuel. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*, 14(7):685–695.

Gaston H Gonnet, Mark A Cohen, and Steven A Benner. 1992. Exhaustive matching of the entire protein sequence database. *Science*, 256(5062):1443–1445.

Julian Gough. 2002. The SUPERFAMILY database in structural genomics. *Acta Crystallogr D Biol Crystallogr*, 58(Pt 11):1897–1900.

Julian Gough. 2005. Convergent evolution of domain architectures (is rare). *Bioinformatics*, 21(8):1464–1471.

Julian Gough and Cyrus Chothia. 2002. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res*, 30(1):268–272.

Gary S Gray and Walter M Fitch. 1983. Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus. *Mol Biol Evol*, 1(1):57–66.

Caroline Hadley and David T Jones. 1999. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure Fold Des*, 7(9): 1099–1112.

Ernst Haeckel. 1866. *Generelle Morphologie der Organismen*. Georg Riemer, Berlin.

Masato Hasegawa, Hirohisa Kishino, and Takahisa Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22 (2):160–174.

Steven Henikoff and Jorja G Henikoff. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Res*, 19(23):6565–6572.

Steven Henikoff and Jorja G Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919.

Aaron E Hirsh and Hunter B Fraser. 2001. Protein dispensability and rate of evolution. *Nature*, 411(6841):1046–1049.

Kevin Howe, Alex Bateman, and Richard Durbin. 2002. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics*, 18(11):1546–1547.

John P Huelsenbeck and Fredrik Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Richard Hughey and Anders Krogh. 1996. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput Appl Biosci*, 12 (2):95–107.

Nicolas Hulo, Amos Bairoch, Virginie Bulliard, Lorenzo Cerutti, Edouard De Castro, Petra S Langendijk-Genevaux, Marco Pagni, and Christian J A Sigrist. 2006. The PROSITE database. *Nucleic Acids Res*, 34(Database issue):D227–30.

Tim Hulsen, Martijn A Huynen, Jacob de Vlieg, and Peter MA Groenen. 2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol*, 7(4):R31.

Laurence D Hurst and Nick G Smith. 1999. Do essential genes evolve slowly? *Curr Biol*, 9(14):747–750.

Rainer Jaenicke. 1987. Folding and association of proteins. *Prog Biophys Mol Biol*, 49(2-3):117–237.

Ravi Jain, Maria C Rivera, and James A Lake. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A*, 96(7): 3801–3806.

Joel Janin and Cyrus Chothia. 1985. Domains in proteins: definitions, location, and structural principles. *Methods Enzymol*, 115:420–430.

Albert Jeltsch. 1999. Circular permutations in the molecular evolution of DNA methyltransferases. *J Mol Evol*, 49(1):161–4.

David T Jones, William R Taylor, and Janet M Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, 8 (3):275–282.

David T Jones, William R Taylor, and Janet M Thornton. 1994. A mutation data matrix for transmembrane proteins. *FEBS Lett*, 339(3):269–275.

I King Jordan, Igor B Rogozin, Yuri I Wolf, and Eugene V Koonin. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res*, 12(6):962–968.

TH Jukes and CR Cantor. 1969. *Evolution of protein molecules*, pages 21–132. Mammalian Protein Metabolism. Academic Press, New York, London.

Marcia L Kalish, Kenneth E Robbins, Danuta Pieniazek, Amanda Schaefer, Nzila Nzilambi, Thomas C Quinn, Michael E St Louis, Ae S Youngpairoj, Jonathan Phillips, Harold W Jaffe, and Thomas M Folks. 2004. Recombinant viruses and early global HIV-1 epidemic. *Emerg Infect Dis*, 10(7):1227–1234.

Georgy P Karev, Yuri I Wolf, and Eugene V Koonin. 2003. Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics*, 19(15):1889–1900.

Georgy P Karev, Yuri I Wolf, Andrey Y Rzhetsky, Faina S Berezovskaya, and Eugene V Koonin. 2002. Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC Evol Biol*, 2(1):18.

Samuel Karlin and Stephen F Altschul. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*, 87(6):2264–2268.

Kazutaka Katoh, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, 33(2):511–518.

Motoo Kimura. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120.

Motoo Kimura. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

H Kishino, T Miyata, and M Hasegawa. 1990. Maximum likelihood inference of protein in of chloroplasts. *J Mol Evol*, 31:151–160.

Eugene V Koonin, Yuri I Wolf, and Georgy P Karev. 2002. The structure of the protein universe and genome evolution. *Nature*, 420(6912):218–223.

Liisa B Koski and G Brian Golding. 2001. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol*, 52(6):540–542.

Antje Krause and Martin Vingron. 1998. A set-theoretic approach to database searching and clustering. *Bioinformatics*, 14(5):430–438.

Sarah K Kummerfeld and Sarah A Teichmann. 2005. Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet*, 21(1):25–30.

E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, N Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin,

A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blocker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowski, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, and Y J Chen. 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Timo Lassmann and Erik LL Sonnhammer. 2002. Quality assessment of multiple alignment programs. *FEBS Lett*, 529(1):126–130.

Christopher Lee, Catherine Grasso, and Mark F Sharlow. 2002a. Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464.

Yuandan Lee, Razvan Sultana, Geo Pertea, Jennifer Cho, Svetlana Karamycheva, Jennifer Tsai, Babak Parvizi, Foo Cheung, Valentin Antonescu, Joseph White, Ingeborg Holt, Feng Liang, and John Quackenbush. 2002b. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res*, 12(3):493–502.

Heng Li, Avril Coghlan, Jue Ruan, Lachlan James Coin, Jean-Karim Heriche, Lara Osmotherly, Ruiqiang Li, Tao Liu, Zhang Zhang, Lars Bolund, Gane Ka-Shu

Wong, Weimou Zheng, Paramvir Dehal, Jun Wang, and Richard Durbin. 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*, 34(Database issue):572–580.

Li Li, Christian J Jr Stoeckert, and David S Roos. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9):2178–2189.

Jinfeng Liu and Burkhard Rost. 2004. CHOP: parsing proteins into structural domains. *Nucleic Acids Res*, 32(Web Server issue):569–571.

Michael Lynch and John S Conery. 2000. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155.

Zhi M Ma and Michael Röckner. 1992. *Introduction to the theory of (non-symmetric) Dirichlet Forms*. Springer-Verlag, Berlin.

Martin Madera and Julian Gough. 2002. A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res*, 30(19): 4321–4328.

Martin Madera, Christine Vogel, Sarah K Kummerfeld, Cyrus Chothia, and Julian Gough. 2004. The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res*, 32(Database issue):235–239.

Thomas Mailund and Christian N S Pedersen. 2004. QuickJoin–fast neighbour-joining tree reconstruction. *Bioinformatics*, 20(17):3261–3262.

Nicola J Mulder, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Paul Bradley, Peer Bork, Phillip Bucher, Lorenzo Cerutti, Richard Copley, Emmanuel Courcelle, Ujjwal Das, Richard Durbin, Wolfgang Fleischmann, Julian Gough, Daniel Haft, Nicola Harte, Nicolas Hulo, Daniel Kahn, Alexander Kanapin, Maria Krestyaninova, David Lonsdale, Rodrigo Lopez, Ivica Letunic, Martin Madera, John Maslen, Jennifer McDowall, Alex Mitchell, Anastasia N Nikolskaya, Sandra Orchard, Marco Pagni, Chris P Ponting, Emmanuel Quevillon, Jeremy Selengut, Christian J A Sigrist, Ville Silventoinen, David J Studholme, Robert Vaughan, and Cathy H Wu. 2005. InterPro, progress and status in 2005. *Nucleic Acids Res*, 33(Database issue):D201–5.

Tobias Müller. 2001. *Modellierung von Proteinevolution*. PhD thesis, Interdisziplinäres Institut für wissenschaftliches Rechnen, Heidelberg.

Tobias Müller and Martin Vingron. 2000. Modeling amino acid replacement. *J Comput Biol*, 7(6):761–776.

Alexey G Murzin, Steven E Brenner, Tim Hubbard, and Cyrus Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540.

SB Needleman and CD Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3): 443–453.

Masatoshi Nei and Sudhir Kumar. 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

Pauline C Ng, Jorja G Henikoff, and Steven Henikoff. 2000. PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics*, 16(9):760–766.

Cédric Notredame, Desmond G Higgins, and Jaap Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1): 205–217.

Artem S Novozhilov, Georgy P Karev, and Eugene V Koonin. 2005. Mathematical modeling of evolution of horizontally transferred genes. *Mol Biol Evol*, 22(8): 1721–1732.

Kevin P O'Brien, Maido Remm, and Erik L L Sonnhammer. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, 33(Database issue):476–480.

C Orengo, AD Michie, S Jones, DT Jones, MB Swindells, and JM Thornton. 1997. CATH–a hierarchic classification of protein domain structures. *Structure*, 5(8): 1093–1108.

Christina A Orengo, Frances M G Pearl, and Janet M Thornton. 2003. The CATH domain structure database. *Methods Biochem Anal*, 44:249–271.

Richard Owen. 1843. *Lectures on the comparative anatomy and physiology of the invertebrate animals*. Longman, Brown, Green & Longmans, London.

Colin Patterson. 1988. Homology in classical and molecular biology. *Mol Biol Evol*, 5(6):603–625.

Frances M G Pearl, C F Bennett, James E Bray, Andrew P Harrison, Nigel Martin, Arian Shepherd, Ian Sillitoe, Janet Thornton, and Christine A Orengo. 2003. The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res*, 31(1):452–455.

William R Pearson. 1996. Effective protein sequence comparison. *Methods Enzymol*, 266:227–258.

Christopher P Ponting and Robert B Russell. 1995. Swaposins: circular permutations within genes encoding saposin homologues. *Trends Biochem Sci*, 20(5): 179–80.

GR Reeck, C de Haen, DC Teller, RF Doolittle, WM Fitch, RE Dickerson, P Chambon, AD McLachlan, E Margoliash, and TH Jukes. 1987. "Homology"in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell*, 50(5):667.

Maido Remm, Christian EV Storm, and Erik LL Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314(5):1041–1052.

Monica Riley and Bernhard Labedan. 1997. Protein evolution viewed through Escherichia coli protein sequences: introducing the notion of a structural segment of homology, the module. *J Mol Biol*, 268(5):857–868.

DR Robinson and LR Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147.

Fredrik Ronquist and John P Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Michael G Rossmann, Dino Moras, and Kenneth W Olsen. 1974. Chemical and biological evolution of nucleotide-binding protein. *Nature*, 250(463):194–199.

Claudia AM Russo, Naoko Takezaki, and Masatoshi Nei. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol Biol Evol*, 13(3):525–536.

Naruya Saitou and Tadashi Imanishi. 1989. Relative efficiencies of the fitch-margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. *Mol Biol Evol*, 6:514–525.

Naruya Saitou and Masatoshi Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425.

David Sankoff and Robert J Cedergren. 1983. *Simultaneous comparison of three or more sequences related by a tree*, pages 253–264. Time Warps, String Edits, and Macromolecules: the Theory and Practivce of Sequence Comparisons. Addison-Wesley, Reading.

Heiko A Schmidt, Korbinian Strimmer, Martin Vingron, and Arndt von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18(3):502–504.

Charles Semple and Mike Steel. 2003. *Phylogenetics*, volume 24 of *Oxford lecture series in mathematics and its applications*. Oxford University Press, Oxford.

Boris E Shakhnovich and Max J Harvey. 2004. Quantifying structure-function uncertainty: a graph theoretical exploration into the origins and limitations of protein annotation. *J Mol Biol*, 337(4):933–949.

Temple F Smith and Michael S Waterman. 1981. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197.

RR Sokal and CD Michener. 1958. A statistical method for evalutating systematic relationships. *Univ Kansas Sci Bull*, 28:1409–1438.

Erik LL Sonnhammer. 2006. Belvu – multiple sequence alignment viewer. URL `http://www.cgb.ki.se/cgb/groups/sonnhammer/Belvu.html`.

Erik LL Sonnhammer, Sean R Eddy, Ewan Birney, Alex Bateman, and Richard Durbin. 1998. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res*, 26(1):320–322.

Erik LL Sonnhammer, Sean R Eddy, and Richard Durbin. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–420.

Erik LL Sonnhammer and Daniel Kahn. 1994. Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci*, 3(3):482–492.

Erik LL Sonnhammer and Eugene V Koonin. 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet*, 18(12):619–620. Letter.

Rainer Spang and Martin Vingron. 1998. Statistics of large-scale sequence searching. *Bioinformatics*, 14(3):279–284.

Christian EV Storm. 2004. *Assignment and Assessment of Orthology and Gene Function*. PhD thesis, Karolinska Institutet, Stockholm.

Christian EV Storm and Erik LL Sonnhammer. 2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 18(1):92–99.

Christian EV Storm and Erik LL Sonnhammer. 2003. Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res*, 13(10): 2353–2362.

Jens Stoye, Dirk Evers, and Folker Meyer. 1998. Rose: generating sequence families. *Bioinformatics*, 14(2):157–163.

Amarendran R Subramanian, Jan Weyer-Menkhoff, Michael Kaufmann, and Burkhard Morgenstern. 2005. DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, 6(1):66.

David L Swofford. 1996. *PAUP:Phylogenetic analysis using parsimony (and other methods)*. Sinauer Associates, Sunderland.

Naoko Takezaki, Andrey Rzhetsky, and Masatoshi Nei. 1995. Phylogenetic test of the molecular clock and linearized trees. *Mol Biol Evol*, 12(5):823–833.

Roman L Tatusov, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Boris Kiryutin, Eugene V Koonin, Dmitri M Krylov, Raja Mazumder, Sergei L Mekhedov, Anastasia N Nikolskaya, B Sridhar Rao, Sergei Smirnov, Alexander V Sverdlov, Sona Vasudevan, Yuri I Wolf, Jodie J Yin, and Darren A Natale. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41.

Roman L Tatusov, Michael Y Galperin, Darren A Natale, and Eugene V Koonin. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*, 28(1):33–36.

Roman L Tatusov, Eugene V Koonin, and David J Lipman. 1997. A genomic perspective on protein families. *Science*, 278(5338):631–637.

Roman L Tatusov, Darren A Natale, Igor V Garkavtsev, Tatiana A Tatusova, Uma T Shankavaram, Bachoti S Rao, Boris Kiryutin, Michael Y Galperin, Natalie D Fedorova, and Eugene V Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*, 29(1):22–28.

Shalini Veerassamy, Andrew Smith, and Elisabeth R M Tillier. 2003. A transition probability model for amino acid substitutions from blocks. *J Comput Biol*, 10 (6):997–1010.

J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy,

L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigo, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. 2001. The sequence of the human genome. *Science*, 291(5507):1304–1351.

Christine Vogel, Matthew Bashton, Nicola D Kerrison, Cyrus Chothia, and Sarah A Teichmann. 2004a. Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol*, 14(2):208–216.

Christine Vogel, Carlo Berzuini, Matthew Bashton, Julian Gough, and Sarah A Teichmann. 2004b. Supra-domains: evolutionary units larger than single protein domains. *J Mol Biol*, 336(3):809–823.

Christine Vogel, Sarah A Teichmann, and Jose Pereira-Leal. 2005. The relationship between domain duplication and recombination. *J Mol Biol*, 346(1):355–365.

January 3rd Weiner and Erich Bornberg-Bauer. 2006. Evolution of circular permutations in multidomain proteins. *Mol Biol Evol*, 23(4):734–43.

January 3rd Weiner, Geraint Thomas, and Erich Bornberg-Bauer. 2005. Rapid motif-based prediction of circular permutations in multi-domain proteins. *Bioinformatics*, 21(7):932–7.

David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Lewis Y Geer, Wolfgang Helmberg, Yuri Kapustin, David L Kenton, Oleg Khovayko, David J Lipman, Thomas L Madden, Donna R Maglott, James Ostell, Kim D Pruitt, Gregory D Schuler, Lynn M Schriml, Edwin Sequeira, Stephen T Sherry, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tugba O Suzek, Roman Tatusov, Tatiana A Tatusova, Lukas Wagner, and Eugene Yaschenko. 2006. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 34(Database issue):173–180.

Simon Whelan and Nick Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*, 18(5):691–699.

Tiffani L Williams and Bernard ME Moret. 2003. An investigation of phylogenetic likelihood methods. In *Proc. 3rd IEEE Symp. on Bioinformatics and Bioengineering*, pages 79–86.

Markus Wistrand and Erik LL Sonnhammer. 2005. Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. *BMC Bioinformatics*, 6(1):99.

Yuri I Wolf, Nick V Grishin, and Eugene V Koonin. 2000. Estimating the number of protein folds and families from complete genome data. *J Mol Biol*, 299(4): 897–905.

Caroline F Wright, Sarah A Teichmann, Jane Clarke, and Christopher M Dobson. 2005. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature*, 438(7069):878–881.

Stefan Wuchty. 2001. Scale-free behavior in protein domain networks. *Mol Biol Evol*, 18(9):1694–1702.

J Zhang and M Nei. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J Mol Evol*, 44 Suppl 1: 139–146.

Christian M Zmasek and Sean R Eddy. 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3(1):14.