From the Center for Genomics and Bioinformatics
Karolinska Institute, Stockholm

Assignment and Assessment of Orthology and Gene Function

Christian Storm

Stockholm 2004

Published and printed by Karolinska University Press
Box 200, SE-171 77 Stockholm, Sweden

ISBN 91-7349-810-6

"Nothing in biology makes sense except in the light of evolution."
- Theodosius Dobzhansky

Abstract

Several genomes from different species have been sequenced over the last years, most notably the human genome. An important task of computational biology is to classify and functionally annotate the large amount of sequence data created by the genome sequencing projects. The concept of orthology and paralogy, developed over 30 years ago by Fitch, plays an important role in this task: Orthologous genes are genes in different species that evolved from a single gene in the last common ancestor of these species. Paralogous genes are genes that evolved due to a duplication event. Orthologs can be seen as different versions of the same gene in different species. Therefore they are likely to have the same functional properties and play a similar biochemical role in the cell. Once an orthologous gene for a newly sequenced gene is known, the annotation of the ortholog can give reliable information about the function and the role of the new gene.

The main focus of the work was to improve existing and develop new approaches for the inference of orthology. We developed a novel method, called ortholog bootstrapping, to analyze a gene tree for orthologs. Instead of only assigning orthology from a single gene tree, ortholog bootstrapping analyses multiple trees calculated for the same gene family. The trees are reconstructed using the bootstrap technique, enabling us to calculate bootstrap support values for orthologous sequence pairs. Ortholog bootstrapping was then used to find orthologs between species with completely sequenced genomes. Here we employed a scheme for the hierarchical clustering of species based on their evolutionary history. The orthology inference was performed on the domain level, using the Pfam domain definitions. The results of the analysis were compared to a tree reconciliation method using a complete species tree for orthology inference. The comparison was based on a testset of putative orthologous proteins with experimentally characterized functional properties. The outcome of the comparison showed that our approach increases the sensitivity for assigning orthologs from a gene tree. Orthologous relations found using our approach were stored in a database. The database is available over the Internet, accessible by a previously developed Java applet for visualizing phylogenetic relations between domains.
In addition to inferring orthology by phylogenetic means we developed a pairwise sequence similarity based method for assigning orthology. It focuses on the correct separation of paralogs and the calculation of an orthology confidence value.

Original Publications

This thesis is based on the following articles, which will be referred to by their roman numerals in the text:

I **Storm C. E. V.,** Sonnhammer E. L. L. (2001)
NIFAS: visual analysis of domain evolution in proteins. *Bioinformatics* **17**: 343-348.

II Remm M., **Storm C. E. V.,** Sonnhammer E. L. L. (2001)
Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**: 1041-1052.

III **Storm C. E. V.,** Sonnhammer E. L. L. (2002)
Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* **18**: 92-99.

IV Hollich V., **Storm C. E. V.,** Sonnhammer E. L. L. (2002)
OrthoGUI: graphical presentation of Orthostrapper results. *Bioinformatics* **18**:1272-1273.

V **Storm C. E. V.,** Sonnhammer E. L. L. (2003)
HOPS: Hierarchical grouping of orthologous and paralogous sequences. *Genome Res* **13**:2353-2362.

Contents

Contents

List of Abbreviations

COG	Cluster of orthologous genes
DNA	Deoxyribonucleic acid
HOPS	Hierarchical grouping of orthologous and paralogous sequences
HMM	Hidden Markov model
KOG	Eukaryotic orthologous groups
MCL	Markov cluster algorithm
MCMC	Markov chain Monte Carlo
NIFAS	No idea for a shortcut
NJ	Neighbor joining
RIO	Resampled inference of orthologs
RNA	Ribonucleic acid

1. Introduction

"Everything should be made as simple as possible, but not simpler."
- Albert Einstein

The numerous completed and ongoing genome-sequencing projects have produced and continue to produce an unprecedented amount of sequence data. Genomes of several organisms are completely sequenced and publicly available, most notably the human genome (Lander *et al.* 2001; Venter *et al.* 2001) and genomes from model organism such as *M. musculus* (Waterston *et al.* 2002), *C. elegans* (*C. elegans* Sequencing Consortium 1998) and *D. melanogaster* (Adams *et al.* 2000). By the end of 2002 the public databases contained more than 22 million sequences from nearly 120.000 species (Stoesser *et al.* 2003; Benson *et al.* 2003).

However, only a tiny fraction of the gene-products from these sequences is experimentally characterized (Karp *et al.* 2001). Given the exponential growth of sequencing capacity it is very likely that the gap between sequenced and experimentally characterized genes will continue to increase in the foreseeable future. This makes functional annotation of sequences one of the most important fields in Bioinformatics. The task of *in-silico* annotation has led to a renewed interest in orthology and paralogy. Protein features such as catalytic activity, posttranslational modification, subcellular localization and physical / chemical properties are known to be more conserved between orthologous proteins than between non-orthologous proteins (Jensen *et al.* 2003). Therefore it is possible to infer the most likely function of a sequence if experimentally characterized orthologous sequences can be found.

Orthology and paralogy is a relatively old concept on a Bioinformatics time scale. Although it has already been introduced more than 30 years ago (Fitch 1970), the concept only became widely used in the late nineties (fig. 1) with the availability of completely sequenced genomes. These allowed for the first time to reliably assign orthology (Tatusov *et al.* 1997).

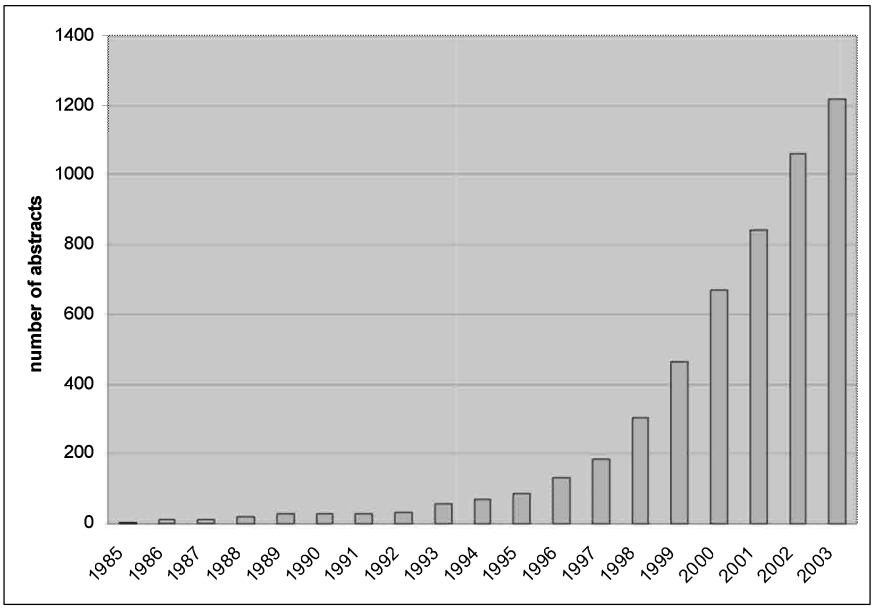

Figure 1 Number of abstracts in the PubMed literature database using the term 'ortholog'.

1.1 Homology, orthology and paralogy

Two sequences that have evolved from a common ancestral sequence are homologous[1]. Orthology and paralogy are a further sub-classification of homology. Dependent on the temporal order and type of events that separated the homologous sequences they are either orthologs or paralogs.

The definition for orthology and paralogy was originally given by Walter Fitch (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).*"

In other words: Orthologous genes are separated by speciation, paralogous genes by gene duplication. The difference to the definition of homology is small, but crucial. Orthology and paralogy refers to the last common ancestor

[1] Homology is not limited to sequences, any structural or behavioral features of two organisms can be homologous, too

of the species hosting the genes. Homology refers to the last common ancestor of the genes itself.

The definition of a "gene" is not as straightforward anymore as it seemed to be in 1970 (Epp 1997). However, the concept of orthology can be extended to any kind of biological sequence data: non-coding DNA, RNA, amino acid sequences and even sequence segments. This work focuses on the assignment of orthology between proteins. Of course strictly speaking it is not proteins that undergo duplication and speciation, but the genomic sequences encoding the information needed to assemble the amino acid sequences.

The definition of orthology has several implications (Fitch 2000):

- The true phylogenies of orthologous sequences and the corresponding species are the same.
  By definition orthologous sequences are separated by a speciation. Therefore two orthologous sequences go back to a single sequence in the last common ancestor of the hosting species. Differences in orthologous sequences can only accumulate after the speciation, thus reflecting the evolutionary history of the species.

- Orthology is not necessarily a one-to-one relation.
  A sequence can be orthologous to more than one sequence in a single species. One-to-many and many-to-many relations are possible. If one or more duplication events follow the speciation, all duplicated sequences are orthologous to the corresponding sequences in the other species (fig. 2).

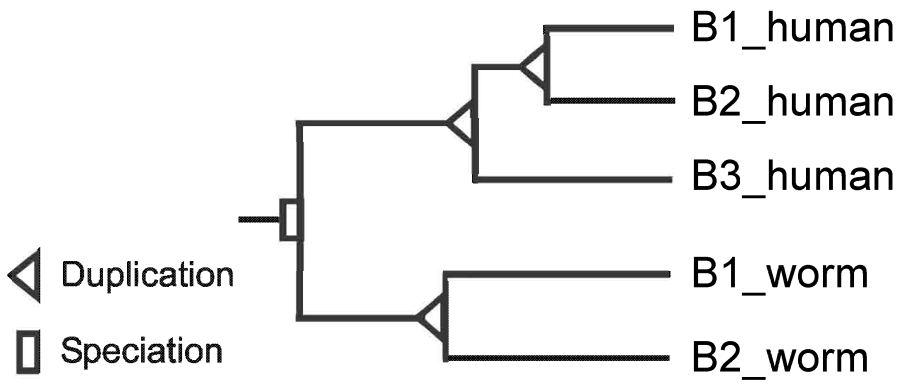

Figure 2 Many-to-many orthologous relations

Multiple duplications followed the speciation at the root of the tree. Nonetheless, the node connection the worm and human sequences in this hypothetical example is the speciation node. Therefore B1_human, B2_human and B3_human are orthologous to B1_worm and B2_worm.

- Orthology is a relative concept based on pairwise species comparison. The definition of orthology refers to the speciation event. This is different for each species pairing. Subsequent grouping of orthologous sequences is possible, but this should take into account the phylogeny of the species. This also means that orthology is not transitive: If a gene A in species 1 is orthologous to a gene B in species 2, and this gene B is orthologous to a gene C in species 3, that does not mean that C is necessarily orthologous to A (fig. 3).

- Within a single species only paralogous relations are present. Sequences within a species can only be separated due to a duplication event.

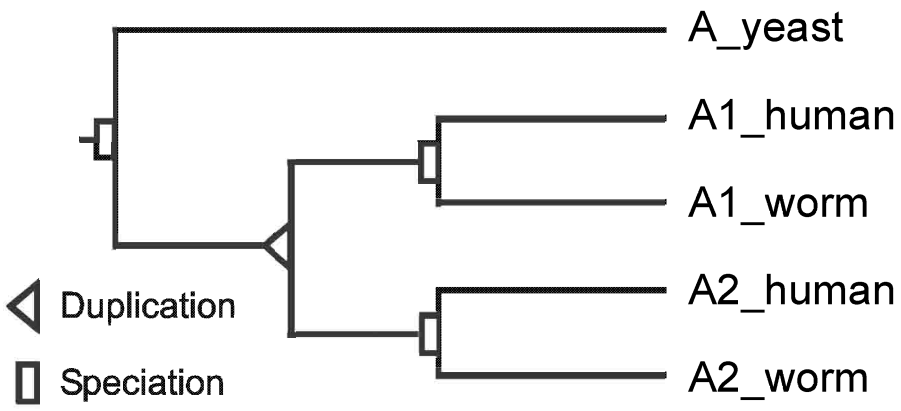

Figure 3 Orthology between sequences from multiple species

A_yeast is separated from all other sequences by a speciation event. Therefore A_yeast is orthologous to all other sequences in this tree. However, A1_human is not orthologous to A2_worm, the connecting node in the tree represents a gene duplication event. The duplication predates the speciation: A1_human and A2_worm, as well as A2_human and A1_worm, are out-paralogs (see text for further explanation).

The concept of paralogy has been refined in recent years, based on the temporal order of the speciation and duplication. We classify paralogs whose duplication occurred after a speciation as 'in-paralogs'. If the duplication occurred before the speciation we label them as 'out-paralogs'[2] (Sonnhammer and Koonin 2002). Unfortunately there is no agreement regarding the terminology. Most authors use their own definitions, for instance paralogs and metalogs (Koonin 2001) or ultra-paralogs and paralogs (Zmasek and Eddy 2002).

The discovery of horizontal gene transfer made it necessary to introduce an additional term – xenology (Gray and Fitch 1983). Sequences that are separated by a horizontal transfer event are called xenologs.

Definitions for orthology given in the literature sometimes include functional properties (Gerlt and Babbitt 2000). Here orthologs are defined as genes in different species with the same catalytic activity. Based on the original phylogenetic definition from Fitch, orthologs are likely to catalyze the same reaction. However, it is possible that orthologs develop different catalytic

[2] Interestingly a comment stating that orthology and paralogy is an unnecessary concept (Petsko, G. A. 2001) resulted in the discussion about the further sub-classification of paralogs into in-paralogs and out-paralogs.

properties. Orthologs with the same function are normally referred to as isofunctional orthologs, or short isorthologs (Fitch 2000).

From a more theoretical point of view a definition for orthology that includes functional properties would reduce its potential for most studies. Often an investigator is trying to infer functional properties of experimentally uncharacterized genes with the help of orthologs. A definition for orthology based on function would not allow this, if used strictly: To assign orthology it would be necessary to determine the functional properties of a gene first.

1.2 Applications

Orthologous genes evolved from a single gene in the last common ancestor. They can be seen as variations of the same gene in different species. This makes orthologous genes of fundamental interest in comparative genomics and neighboring fields. A few important applications for the concept of orthology are described below.

1.2.1 Comparative and evolutionary studies

When comparing genes between species, it is important to compare orthologous and not paralogous genes (Doolittle *et al.* 1996; Feng *et al.* 1997). Differences between orthologous genes go back to evolutionary influences since the speciation. They reflect the evolutionary history of species since the diversion of these species. Orthologous genes start with their evolutionary clock at zero; they only begin diverging with the split of the last common ancestor in two independent species. Paralogous genes were present in the last common ancestor in two or more copies; they already had a different evolutionary history before the two species under comparison evolved.

Therefore orthology is a very basic concept in comparative genomics. Thornton even claims (Thornton and DeSalle 2000): "[…] the fundamental activity of comparative genomics is to track the presence, structural characteristics, function, and map position of orthologs in multiple genomes."

A few examples of recent comparative and evolutionary studies using orthologs are the intron position correlation investigation of genes in different species (Muller *et al.* 2002), domain rearrangements (Yanai, I. *et al.* 2002), the evolution of biochemical pathways (Ranson *et al.* 2002), comparative genome analysis's of *Anopheles gambiae* and *Drosophila melanogaster* (Zdobnov *et al.*

2002), between cyanobacteria and plants (Sato 2002) and of the rice and *Arabidopsis* Dof gene families (Lijavetzky *et al.* 2003).

1.2.2 Transfer of functional annotation

Standard methods for functional annotation are based on sequence similarity. Different tools and databases are available for this task. A widely used approach for predicting functional features of a sequence is to refer to the annotation of homologous sequences (Bork *et al.* 1998). By assigning a predicted or uncharacterized protein to a family of homologous proteins it is possible to transfer functional annotation from experimentally characterized proteins within the same family. However, a higher resolution in the clustering of proteins can be necessary: Proteins with diverse function can still belong to the same protein family (Gogarten and Olendzenski 1999). Therefore in addition to having a complete list of all members of a protein family, researchers are often interested in the relations of the proteins within a family.

Ad hoc solutions to this problem are mostly based on highest-ranking BLAST (Altschul *et al.* 1997) hits. But this approach can give ambiguous results. A database similarity search will sometimes find multiple significant matching genes with different functions. Here it becomes necessary to choose from which genes to transfer the functional annotation. Often a cutoff is applied to chose the sequences, but the cutoffs used are arbitrary and do not solve the problems if multiple significant hits with different function are present. Simply picking the best hit is also problematic; sequence similarity does not ensure identical function (Bhatia *et al.* 1997; Karp 1998; Andrade *et al.* 1999; Eisen and Wu 2002)

However, a 'natural cutoff' is provided by evolution: the distinction between orthologs and non-orthologs. Information about the potential function and role of a gene can be gained by analyzing a set of homologous sequences for orthologs. The chance that orthologs retained the same function as their ancestral gene is high. In this approach evolutionary similarity is used for the transfer of functional annotation instead of sequence similarity. This use of phylogenetic information for the prediction of gene functions has been labeled phylogenomics (Eisen 1998).

In practice many-to-many orthologous relations complicate this procedure. While orthologs with a one-to-one relation are very likely to have the same function and similar role in different species, this is not the case for one-to-

many or many-to-many orthologous relations (Copley *et al.* 2002). After gene duplication, one copy of the gene has a lower evolutionary pressure to keep its function and is free to develop a new function (Kondrashov *et al.* 2002). Therefore for the transfer of functional annotation one-to-one orthologs are of main interest.

Another approach for functional annotation using orthologs is based on phylogenetic profiles (Pellegrini *et al.* 1999). Phylogenetic profiles represent the information about the presence or absence of orthologous sequences in all known genomes. Pellegrini showed that functionally linked proteins show the same or similar phylogenetic profiles. Various tools and databases use phylogenetic profiles for functional annotation (Wong *et al.* 2003; Mering *et al.* 2003).

1.2.3 Phylogenetic footprinting

Phylogenetic footprinting is a principle for identifying conserved regulatory elements in genomic DNA (Blanchette and Tompa 2002). It was first used in 1988 by Tagle (Tagle *et al.* 1988) for the identification of conserved regulatory elements in embryonic epsilon and gamma globin genes.

Standard models used for predicting transcription factor binding sites only include information stored in the naked DNA. The conserved region of a binding site might only consist of five or six nucleotides (Wingender *et al.* 2001), therefore this approach results in an abundant number of hits (Krivan and Wasserman 2001). While transcription factors bind to most of these predicted sides in vitro, often they are not active in vivo. Whether a potential transcription factor-binding site is active in vivo or not is determined by its accessibility (Wu and Grunstein 2000) and the combination of sites within the gene.

Phylogenetic footprinting is based on the observation that selective pressure causes functional elements in genomic sequence to diverge slower than non-functional elements. If in a set of orthologous genes conserved non-coding regions can be found, it is likely that these regions have a regulatory function (Lenhard *et al.* 2003). Predicted transcription factor binding sites that are conserved over multiple genes are therefore more likely to be active in vivo.

1.3 Methods and Databases for assigning orthology

Although the definition for orthology is simple – sequences separated by a speciation event – finding orthologous sequences is far from trivial. Gene loss, incompletely sequenced genomes, the modular structure of proteins, horizontal gene transfer and different rates of evolution between lineages are some of the issues complicating the extraction of orthologous relations from the sequence databases.
Different methods have been developed that attempt to solve these problems. The algorithms used for finding orthologs can be divided into two groups: Sequence similarity based approaches and methods employing phylogenetic trees for ortholog inference.

1.3.1 Sequence similarity based methods

All sequence similarity-based methods share the use of pairwise alignments and the subsequent scoring of the alignments for the assignment of orthology. In the simplest approaches sequences from two or more species showing mutual highest scoring alignments to each other are considered to be orthologous. To achieve this all known and predicted proteins from one species are aligned to all known and predicted proteins from another species. The scores of these pairwise alignments are then used to assign orthology: Only if the highest scoring protein for a query also has the query protein as the highest scoring alignment the proteins are considered to be orthologous. This is based on the assumption that orthologs should have a higher similarity to each other than to any other protein in the other species.
The aligning and scoring is often done with BLAST, the mutual best scoring pairs are referred to as 'best-best-BLAST hits'. Recently the 'best-best BLAST hit' scheme has been refined to not only find the highest scoring blast hit, but the nearest phylogenetic neighbor (Wall *et al.* 2003). This algorithm, named reciprocal smallest distance algorithm, uses maximum likelihood estimation of evolutionary distances in a 'best-best hit" way to detect orthologous sequences.

Orthologs assigned by mutual best BLAST hits have been used in a comparison of proteins from human, yeast, worm and fly (Mushegian *et al.* 1998). In this study the orthologous proteins are used to reconstruct the evolutionary history of the four species, concluding that the established species tree might not be correct. However, a drawback of a 'best-best BLAST hit' scheme is its limitation to one-to-one orthologous relations. Only the best scoring sequences are considered to be orthologous, thereby missing potential in-paralogs. By definition in-paralogs of the best hit sequence are also orthologous to the query

sequence. The data from Mushegian's study has been re-analyzed using phylogenetic trees (Xie and Ding 2000). This revealed that indeed most of the orthologous relations found in the original study are not one-to-one, but one-to-many or even many-to-many relations.

The limitation to one-to-one orthologs is not intrinsic to similarity-based methods. More sophisticated approaches to assign orthology by sequence similarity have been developed. These methods take into account possible one-to-many and many-to-many relations.

1.3.1.1 Clusters of Orthologous Groups (COG)

The COG database (Tatusov *et al.* 1997) was the first systematic large-scale approach to delineate orthologous relations between all completely sequenced genomes. The first release was based on the analysis of seven complete genomes, which was extended to 66 genomes from unicellular organisms in the latest release (Tatusov *et al.* 2003). In addition clusters of orthologs for seven multicellular eukaryotic genomes, named KOGs after eukaryotic orthologous groups, were added in the last release. The focus of the COG (respectively KOG) database is on delineating one-to-many and many-to-many orthologous relations and to cluster the sequences accordingly.
The assignment is based on an all-against-all protein comparison using BLAST. In a first step all proteins from the same genome that are more similar to each other than to any protein from any other species are considered to be (in-) paralogs. These groups of putative (in-) paralogs are then analyzed for triangle patterns of mutual best hits between proteins from at least three different species to form a COG. Triangles that have two proteins in common are merged into a single COG. COGs then undergo a manual case-by-case analysis in order to find COGs that are incorrectly cross-linked by multidomain proteins. Proteins that are composed of multiple domains might only have one or a few homologous domains in common (Hegyi and Bork 1997), while the other domains or the proteins are not related. If this is the case, the proteins in the COG are split into single domain sequence segments and reanalyzed. In a final step all proteins that have at least 2 best hits to any COG are added to it.

The COG database has been criticized (Eisen 1998; Eisen and Wu 2002; Zmasek and Eddy 2002), mainly on the basis that phylogenetic relations are inferred by sequence similarity without explicitly applying an evolutionary model.
In addition, Thornton criticized the general setup of the COGS approach to delineate orthologs. He provides a specific example why COGs might give

wrong results (Thornton and DeSalle 2000). The criticism is based on a one-to-many orthology example, claiming that the COG approach would not assign orthology correctly for this case. However, the authors fail to notice that orthology in their example would have been assigned correctly by COGs. The way COGs assigns orthology (in-) paralogs are collapsed in the first step or added to a cluster of orthologs by expanding it through single best hits in the final step

A more substantial point of critique arises from the non-transitivity of orthology. In the COG database current release orthology for proteins from up to 66 species is reported in a single COG. Any approach that summarizes orthology between three or more species in such a way will give a simplified view of the actual orthologous relations.
In the introduction (fig. 3) an example is shown where the genes A1_human and A2_worm are orthologous to the gene A_yeast. But A1_human and A2_worm itself are out-paralogs, separated by a duplication predating the speciation. Summarizing orthologous relations in a way done by COGs is bound to miss the orthology between A1_human and A2_worm to A_yeast or will assign orthology incorrectly between A1_human and A2_worm. In paper II a non-hypothetical example is given where COGS fails to delineate orthology correctly.

1.3.1.2 OrthoMCL

OrthoMCL (Li *et al.* 2003) is a BLAST based approach that works similar to Inparanoid (II): OrthoMCL groups all sequences together as in-paralogs that are (reciprocally) more similar to each other than to any sequence from another genome. But in a next step putative orthologous and paralogous relationships are converted into a graph. In this graph protein sequences are represented as nodes and the weighted edges of the graph model the relationships between the proteins. The resulting graph is represented by a symmetric similarity matrix. The MCL algorithm (van Dongen 2000) is then used to extract orthologous and (in-) paralogous relations from the similarity matrix.
The results for a pairwise species comparison are very similar to the ones obtained with Inparanoid. In contrast to Inparanoid, OrthoMCL can also extract and cluster orthologous sequences between more than two species.

1.3.2 Phylogenetic approaches

The definition for orthology is based on phylogenetic relations. Therefore it is not surprising that the first approaches to assign orthology were based on

phylogenetic trees. Goodman *et al.* (Goodman *et al.* 1979) were the first to use a mapping function between a tree calculated from sequence data, the gene tree, and a tree representing the evolution of the species, the species tree, to find duplication and speciation events.

However, finding orthologs in a gene tree is only the last step in any automated approach. The following preprocessing of the sequences is necessary:

- Clustering of homologs
- Calculation of evolutionary distances
- Reconstruction of phylogenetic trees.

These are well-defined problems in bioinformatics. Different solutions applicable for each step are briefly outlined below:

1.3.2.1 Clustering of homologs

Standard programs like BLAST or Fasta (Pearson and Lipman 1988) search a database for sequences that are similar to a query sequence. Although there are a few reported cases of convergent evolution (Gandbhir *et al.* 1995; Oren 1995; Haney *et al.* 1999), the number of all theoretical possible protein sequences is so large that any significant similarity found between two proteins is very likely the result of homology (Lipman and Pearson 1985).
The results of a database search can be used to create groups of homologs. This is for instance done by the SYSTERS (Krause *et al.* 2000) and ProtoMap (Yona *et al.* 2000) databases. Tools for the delineation of protein families using pairwise hits are GeneRage (Enright and Ouzounis 2000) and Tribe-MCL (Enright *et al.* 2002).

Multidomain proteins complicate the delineation of protein families from BLAST or Fasta searches. Some proteins are only homologous over a short stretch of their sequences. Therefore two unrelated protein families might be incorrectly merged because they are cross-linked via a multidomain protein. Several databases to delineate protein families based on domains exist, including Pfam (Bateman *et al.* 2002) and TIGRFAM (Haft *et al.* 2001). These databases use Hidden Markov Models, short HMMs, (Durbin *et al.* 1998) to group domains into families. A related approach, but employing PSI-BLAST (Altschul *et al.* 1997) instead of HMMs, is used by ProDom (Servant *et al.* 2002).

In addition, databases like Blocks+ (Henikoff *et al.* 1999), MetaFam (Silverstein *et al.* 2001) and Interpro (Mulder *et al.* 2003) combine the information from several protein family sources.

1.3.2.2 Calculation of evolutionary distances

Placing a sequence in the context of other members of a protein family requires an alignment (Lecompte *et al.* 2001). The goal of an alignment is to make sure that only homologous sites are aligned and subsequently compared.

Some protein family databases, such as Pfam, provide multiple alignments. If this is not the case, or the families are created *ab initio*, a multiple alignment has to be calculated. See Notredame (Notredame 2002) for a review of programs and algorithm for calculating multiple alignments.

A number of methods have been developed to improve the quality of automatically created alignments. These include the MaxHom (Sander and Schneider 1991) and the Probe program (Neuwald *et al.* 1997). Based on a BLAST database search the programs include the most conserved segments in the multiple alignments, but exclude more variable regions between these segments. DbClustal (Thompson *et al.* 2000) also constructs multiple alignments from database searches, but in addition includes the less conserved regions. DbClustal uses conserved sequence segments identified by BLAST as anchor points and then calculates a multiple alignment weighted towards these anchor points. Another program for automated multiple alignment computation is the implementation of the Shuffle-LAGAN algorithm (Brudno *et al.* 2003). Shuffle-LAGAN improves the automated alignment calculation by incorporating a combination of global and local alignment methods.

Once a multiple alignment is calculated it is possible to estimate the evolutionary distance from pairwise sequence similarity (Phillips *et al.* 2000).

However, programs such as DbClustal use a guidance tree derived from pairwise distances to construct a multiple alignment. The topology of the guidance tree will influence the bootstrap support for a tree topology reconstructed from the multiple alignment (Thorne and Kishino 1992). An alternative method to infer evolutionary distances is based on obtaining a maximum-likelihood estimate of evolutionary distances between each pair of sequences in a protein family (Thorne *et al.* 1991). Here all possible pairwise alignments between two sequences contribute to the evolutionary distance estimates.

1.3.2.3 Reconstruction of phylogenetic trees

A field that has received a lot attention over the last three decades is the reconstruction of phylogenetic trees from sequence data. A wide variety of different methods exist (Nei 1996; Whelan *et al.* 2001).

In addition to pairwise distance based methods for tree reconstruction such as neighbor joining (Saitou and Nei 1987), more advanced approaches have been developed. Frequently used are maximum likelihood approaches. Maximum likelihood tree reconstruction will give the correct topology, if enough data and a correct model of evolution is available (Rogers 1997). The likelihood framework makes it also possible to use techniques like Markov Chain Monte Carlo (MCMC) simulation and Bayesian inference (Huelsenbeck and Ronquist 2001).

However, likelihood methods are time-intensive to compute and because of this not yet applicable to large sets of data. Comparisons suggest that simple methods like neighbor joining give a good estimate of the true phylogenetic tree (Nei *et al.* 1998; Kumar and Gadagkar 2000). In addition Morrison and Ellis examined the influence of different sequence alignment methods on tree topologies (Morrison and Ellis 1997). They conclude that the resulting tree is more dependent on the quality of the underlying multiple alignment than on the method used for phylogenetic tree reconstruction.

1.3.2.4 Analysis of a gene tree for orthologs – tree reconciliation

Inferring the evolutionary position of species (the species tree) is a application of gene phylogeny (Page 2000). A finding from these analyses is that gene trees do not necessarily reflect the species tree (Martin and Burg 2002). This has several reasons, including gene duplication, gene loss, horizontal transfer of genes and convergent evolution.

Based on this observation is the idea of tree reconciliation. The topology of the true phylogenetic tree for orthologous sequences is identical with the species tree. Given a trusted species tree, differences in the topology of a gene tree and the species tree can therefore be used to infer orthologous and paralogous relations from the gene tree.
This is done by postulating the minimum number of gene duplications and gene losses that are necessary to reconcile a gene tree with a given species tree (Goodman *et al.* 1979; Page 1994). Once potential gene duplications and losses are assigned to the gene tree, orthologous and paralogous relations can be read from the reconciled tree (fig. 4a).

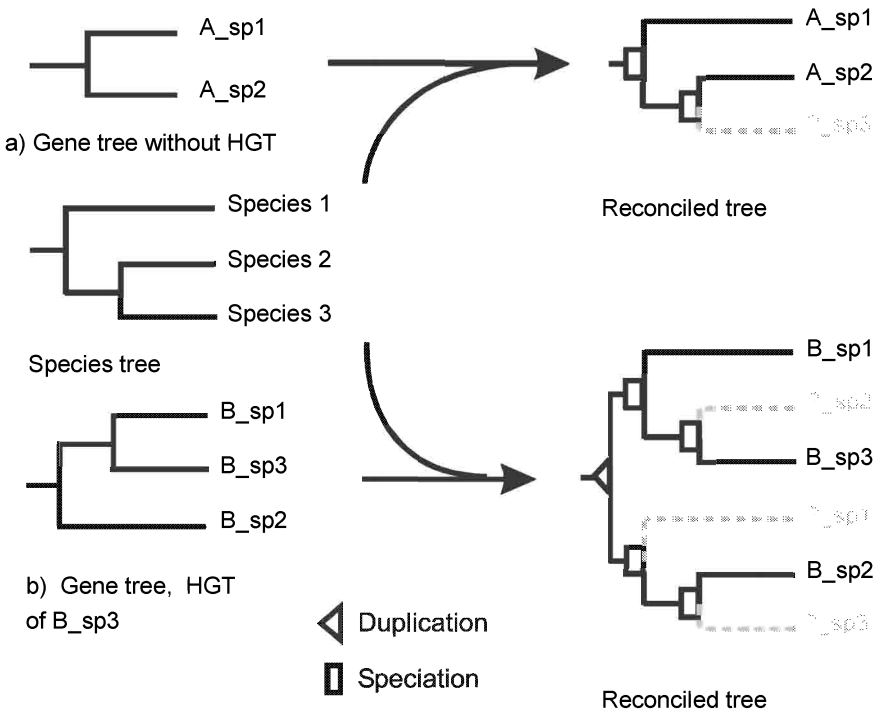

Figure 4 Tree reconciliation and horizontal gene transfer (HGT)

Predicted gene losses are shown in gray. a) Reconciling a gene tree with no HGT and a species tree results in the correct assignment of speciation events. b) Gene tree for a gene family with a recent HGT from sp1 to sp3. B_sp1 and B_sp3 are xenologs. Current algorithms for tree reconciliation cannot account for xenology, and try to reconstruct the evolutionary history of the genes by incorrectly postulating duplications. Not only is B_sp1 incorrectly assigned orthologous to B_sp3 in this example, but the orthology between B_sp1 and B_sp2 is missed.

An underlying assumption is that only gene duplications and subsequent losses took place. Horizontal transfer events and adaptive evolution are not taken in to account in this approach. If they have occurred then current implementations for tree reconciliation algorithm will give erroneous results (fig. 4b).

1.3.2.5 Algorithms for tree reconciliation

Two implementations of a tree reconciliation algorithm are available. Roderick Page first used a brute force algorithm running in $O(n^3)$ time in his program GeneTree (Page 1998). This was later replaced with an algorithm developed by Eulenstein (Eulenstein 1998) that runs in nearly linear time.

Zmasek and Eddy (Zmasek and Eddy 2001) implemented a tree reconciliation algorithm based on Eulenstein's work. Although the theoretical runtime performance of their implementation is worse than the original algorithm developed by Eulenstein, they claim that it is superior in time for trees with up to 550 genes and species, due to a lower time constant in the initialization process.

The only orthology database currently available using tree reconciliation is based on the implementation from Zmasek and Eddy. The tree reconciliation algorithm is integrated in an automated pipeline, named RIO (Resampled Inference of Orthologs) (Zmasek and Eddy 2002).

The web-based interface[3] for RIO takes a query protein with annotated Pfam domain information as the input. The output is a list of proteins orthologous to the query protein, sorted by a bootstrap confidence value. One-to-one orthologs are named "super-orthologs" and in-paralogs are labeled "ultra-paralogs".

The procedure for assigning orthology relies on the use of the Pfam alignments. First the query protein is aligned to the corresponding protein family in Pfam. The alignment is bootstrapped 100 times and a pairwise distance matrix is calculated for each alignment. Unrooted trees reconstructed from the distance matrices are then rooted by minimizing the number of postulated duplication events. These rooted bootstrap trees are reconciled with a complete species tree derived from the "Tree of Life" project[4]. For each sequence in the protein family it is counted in how many bootstrap trees it is orthologous to the query sequence. This number is reported as the ortholog bootstrap support for two sequences.

1.3.2.6 Bayesian tree reconciliation using MCMC

Tree reconciliation is based on a parsimony model. Lars Arvestad and co-workers (Arvestad *et al.* 2003) developed a tree reconciliation algorithm in a Bayesian framework using Monte Carlo Markov Chains (MCMC).
In their model a gene tree evolves inside a species tree; gene duplication and loss events are simulated by a birth-death process. Based on this model the *posteriori* distribution for the reconciliation of the species tree with the gene tree is approximated using MCMC. This allows to find the most probable reconciliation of a gene tree with a species tree and to estimate the probability of any reconciliation.

[3] http://www.rio.wustl.edu/
[4] http://tolweb.org/tree/phylogeny.html

1.4 Comparison of sequence similarity based and phylogenetic methods for inferring orthology

Similarity based methods are fast and easy to automate. The COG database is probably the orthology resource with the highest impact. Many subsequent studies use the orthology relations from COGS or developed a similar scheme for inferring orthology.(Braun *et al.* 2000; Gonczy *et al.* 2000; Walhout *et al.* 2000; Koonin *et al.* 2002; Zdobnov *et al.* 2002).

However, orthology is a phylogenetic concept. Trying to assign orthology without an explicit model of evolution can be problematic. Orthologous sequences in different species are known to evolve at different rates (Kondrashov *et al.* 2002). This means that two orthologs might not necessarily have the highest similarity to each other (Koski and Golding 2001). In this case a pairwise similarity method will assign orthology incorrectly. Dependent on the tree reconstruction algorithm used, a phylogenetic approach still assigns orthology correctly. The neighbor joining method, for instance, reconstructs the evolutionary history correctly even for different rates of evolution (Durbin *et al.* 1998).

Similarity based approaches assign orthology from pairwise sequence similarity, eliminating the error-prone step of constructing a multiple alignment or the time intensive estimation of maximum-likelihood distances. On the other hand, the goal of using a multiple alignment is to make sure to infer the correct evolutionary distances between the sequences (Phillips *et al.* 2000). This is not the case for similarity-based scores derived from pairwise alignments. These scores are only rough estimates of the true evolutionary distances.

In response to these critiques, Tatusov wrote: “The approach used for the construction of COGS does not supplant a comprehensive phylogenetic analysis. Nevertheless, it provides a fast and convenient short-cut to delineate a large number of families that most likely consist of orthologs.” (Tatusov *et al.* 2000)

While in theory a phylogenetic approach is better suited to assign orthology than a similarity based one, there are some pitfalls.
A crucial point is the size of the protein families. To allow computation of multiple alignments and phylogenetic trees, large protein families might have to be split up, using a program like Secator (Wicker *et al.* 2001). On the other side the protein families must be large enough to assign orthologous sequences to the same protein (sub-)family.

Introduction

In an automated approach the quality of the multiple alignments calculated for the protein families can be poor. Sequences with repeats and large deletions and insertions can be difficult to align (Thompson *et al.* 1999).

Sophisticated methods for reconstructing phylogenetic trees, like Bayesian inference, are time-intensive and not applicable to large data sets, even with today's computational resources. With the current computer power available, fast methods such as neighbor joining (NJ) are still the only feasible algorithms for large-scale calculation of trees. However, NJ is known to give wrong trees (Nei 1996) for a number of cases, especially for short sequences or heterogeneous rate of evolution among sites.

The assumption of a given, correct species tree is also problematic. Most of modern phylogeny is based on molecular phylogenetics, often derived from ribosomal RNA (Woese and Fox 1977; Woese 1987). Artifacts related to differences in evolutionary rate and mutational saturation can lead to ambiguous species trees (Doolittle 1999). An example is the phylogeny of *H. sapiens*, *D. melanogaster* and *C. elegans*. In an analysis done by Mushegian (Mushegian *et al.* 1998) the majority of trees reconstructed for 42 candidate orthologs from these species do not reflect the established species tree, wherein fly and nematode are sister taxa. Instead 66% of the trees in the study show a topology with a fly sequence as a sister taxon to a human sequence.

2 Methods and Databases

This section is intended as a brief overview of the main methods and database used. For details refer to the corresponding sections of the papers I-V and references given in the text.

2.1 Pfam (Paper I and V)

Pfam (Bateman *et al.* 2002) is a protein family database based on Hidden Markov Models (HMM). It consists of two parts, Pfam-A and Pfam-B. Pfam-A families are derived from manually annotated high quality multiple alignments. From these 'seed' alignments HMMs are calculated and used to search public databases for all members of the corresponding families. The HMMs are also available as a library to search a query sequence against. In the current release 11.0, Pfam contains 7255 Pfam-A families. Pfam-B domains are automatically generated employing the Domainer algorithm (Sonnhammer, E. L. and Kahn, D. 1994).

2.2 Neighbor-Joining (Paper III)

The neighbor-joining (NJ) algorithm reconstructs a tree from a matrix of pairwise distances. It was developed by Saitou & Nei (Saitou and Nei 1987) and modified by Studier & Keppler (Studier and Keppler 1988). NJ is guarantied to reconstruct the correct tree, if the distance between any pair of sequences in the tree is equal to the sum of the lengths of the branches connecting them. This property of a tree is called additivity. A detailed description of the neighbor-joining algorithm can be found in (Durbin *et al.* 1998).

2.3 The Bootstrap Technique (Paper II and III)

The bootstrap method is a resampling technique developed by Efron (Efron and Tibshirani 1993). It can be used to obtain the distribution of a data sample when the underlying statistical population is unknown. The distribution is estimated by resampling with substitution from the data sample. Felsenstein was the first to apply the bootstrap method to estimate the reliability of a given tree topology (Felsenstein 1985): From a multiple alignment new alignments of the same length are created by repeatedly sampling columns with replacement. For these pseudo-samples trees are reconstructed and compared to a tree calculated for the original alignment.

3 Aims of this Investigation

The aim of this study has been to improve and automate methods for inferring orthology. At the commencement of this work no method for finding orthologs assigned a confidence value to orthologous pairings. The only large-scale analysis of orthology – the COG database - was done by pairwise similarity.

Based on this initial situation, the goals in detail were:

- To develop methods for calculating confidence values for putative pairs of orthologs
- To set up a database of orthologous sequences
- To develop a user interface for easy and effective access to this database

4 Results and Discussion

"42"
- Douglas Adams, The Hitch-Hiker's Guide to the Galaxy.

4.1 Paper I - NIFAS: visual analysis of domain evolution in proteins

In this work we wanted to investigate the evolutionary history of protein domains rather than whole length sequences. In an attempt to visualize and allow effective analysis of domain evolution in proteins, we developed NIFAS.

By comparing the phylogenetic trees of the domains in one protein it is possible to infer the most likely evolutionary scenario to explain differences in domain architecture between proteins. If proteins are composed of the same domains in the same order and the phylogenetic trees show the same topology, the analysis is trivial: Domain recombination did not take place. However, if proteins are composed of different domains, or the same domains in different order, comparing the phylogenetic trees for each domain can reveal the possible evolutionary events that shaped the current proteins.

The two examples for domain recombinations analyzed in this work show that orthology analysis should take the domain structure of proteins into account. Proteins can be composed of domains with different evolutionary history. This means that two proteins might not only be separated by a single speciation or duplication event, but parts of the protein by a speciation and other parts by a duplication event. In other words, while some domains of two proteins might be orthologous, other domains within the same proteins might be paralogous.

4.2 Paper II - Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons

Parallel to our work on phylogenetic inference of orthology, M. Remm developed a similarity-based approach for assigning orthology. This work relies, like COG, on best-best BLAST hits. It focuses on distinguishing between in- and out-paralogs. Due to the non-transitivity of orthology the analyses are performed on a pairwise basis between all proteins from species with completely sequenced genomes.

Results and Discussion

A question in this work was how to assign a confidence value to a given orthologous pair. An analytic approach would have to take into account the size of the protein family and the amino acid distribution of its members (Hertz and Stormo 1999). This would mean losing the key advantages of similarity-based approaches: speed and ease of automation. Instead we chose to use the bootstrap method. Here it is not necessary to make any assumptions about the amino acid distribution of the sequences compared.
The way the bootstrapping is done is a straightforward application of the non-parametric bootstrap. Two alignments, the main ortholog pair A1B1 and a lower scoring pair A1B2, are compared. The columns in each alignment are sampled with replacement. For each pseudo-sample the BLOSUM62 score (Henikoff and Henikoff 1992) is calculated. From the distribution of scores a confidence value is estimated (fig 5). This estimate reflects the confidence in the hypothesis that A1B1 is the higher scoring alignment, rather then A1B2.

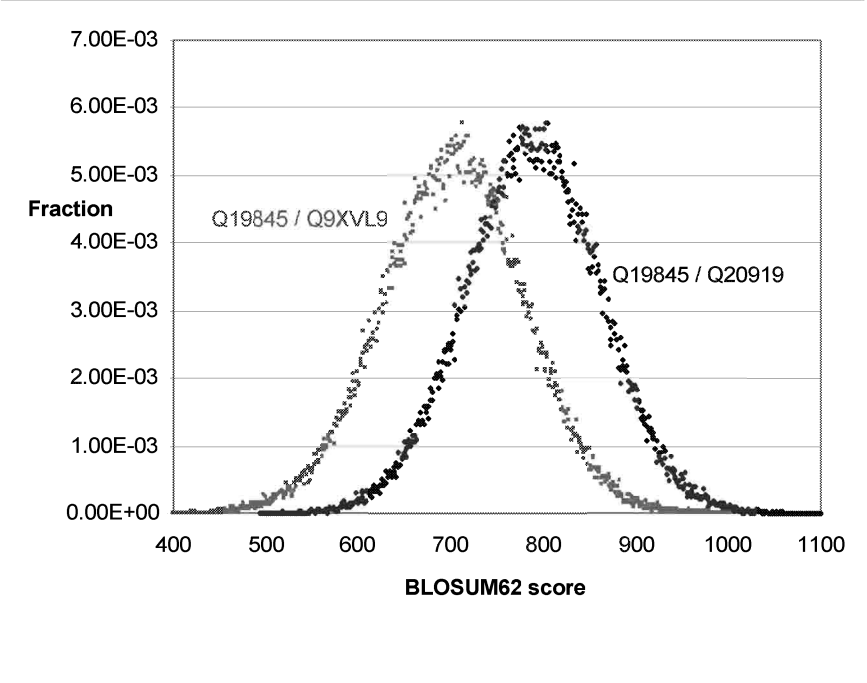

Figure 5 Distribution of bootstrap scores

Shown are the distributions of BLOSUM62 bootstrap scores for two pairwise alignments: Q19845/Q20919 and Q19845/Q9XVL9. The scores of the original alignments are 788 (Q19845/Q20919) and 703 (Q19845/Q9XVL9). The confidence calculated from this data that Q19845/Q20919 is a higher scoring alignment than Q19845/Q9XVL9 is 0.77.

The bootstrap method only works on independent data points. For protein sequence this is not strictly true, but a reasonable simplification (Efron *et al.* 1996). However, this simplification cannot be made for insertions and deletions in the alignment. Rather then sampling the amino acids of an insertion or deletion independently, an insertion/deletion is considered to be a single character. Therefore insertions/deletions are sampled as a single unit, with the same probability of being sampled as any amino acid character, independent of the length of the insertion/deletion.

4.3 Paper III - Automated ortholog inference from phylogenetic trees and calculation of orthology reliability

The last step in an automated pipeline to assign orthology by phylogenetic means is the analysis of a phylogenetic tree. The central point of this publication is the calculation of confidence values respectively orthology support values.

Often the phylogenetic bootstrap value for the connecting node of the orthologous sequences in the tree is taken to assign a confidence to a phylogenetic estimate (Remm and Sonnhammer 2000). However, this value does not include any information of the branching pattern of the sequences below this node. The ortholog bootstrap method developed during this work reflects all possible tree topologies supporting orthology that are found in the pseudo-trees. This approach to calculate an orthology bootstrap value was also used later in the program RIO.

The algorithm developed for this publication analyses a phylogenetic tree to find orthology between two (groups of) species, rather than trying to infer orthology between all species as this is done by tree reconciliation.
This design opened a very flexible way of including or excluding any phylogenetic information in the orthology assignment process. Closely related species can be clustered in one group, without having to make an assumption of the taxonomic relations within this group. Species whose evolutionary position is unsure or still debated can be excluded from the analysis. This enabled us later on (see V) to avoid some problems normally connected to orthology assignment from a gene tree.

The way ortholog bootstrapping assigns orthology could have been approximated with a tree reconciliation algorithm. This would have been

possible by using a species tree with three artificial species, corresponding to the three species groups used in ortholog bootstrapping. However, at the time of this work the only available implementation of a tree reconciliation algorithm was GeneTree. GeneTree is only available for Windows and MacOS, therefore lacking the flexibility and power of UNIX machines.

4.4 Paper IV - OrthoGUI: graphical representation of Orthostrapper results

'Orthostrapper' is only available as a command line based Java program. To simplify its use and add features for an effective analysis a Java applet, 'OrthoGUI', was developed.

OrthoGUI takes a multiple alignment in 'Aligned Fasta' format as input. Species information must be given in the alignment by appending '&1', '&2' and '&O' to the sequence names, marking them as either belonging to species group 1, species group 2 or the outgroup species.

An average-linkage clustering algorithm was implemented to calculate groups of orthologs and paralogs for the input data. This average linkage-clustering algorithm is now also included in the 'Orthostrapper' command line program. This allows clustering orthologs from more than two species. For instance if one is interested in groups of orthologs between worm, fly and human, this algorithm can be used to find consistent groups of in-paralogs derived from a pairwise analysis.

4.5 Paper V - HOPS: Hierarchical grouping of orthologous and paralogous sequences

After the development of the ortholog bootstrapping algorithm the question was how to apply it. Early experiments with protein families delineated by BLAST based approaches failed because of the poor quality of automatically calculated multiple alignments and problems due to the cross-linking of unrelated sequences by multiple domain proteins.

The phylogenetic investigation done with NIFAS already indicated that orthology inference should be done between domains rather than whole length proteins. We chose the Pfam database for the analysis, because it provides multiple alignments for the protein families. In addition a tool for visualizing phylogenetic relations in Pfam was already present (NIFAS).

Results and Discussion

The sub-classification of Pfam in orthologs resulted in the HOPS database. By clustering evolutionary related species into distinct groups many of the problems connected to tree reconciliation can be avoided. Only reliable and robust phylogenetic information is used to infer orthology:

If the ratio (time between speciation) / (time passed since the speciation) is too small, the speciation events might not be reflected correctly in a gene tree of orthologous sequences. As shown in the analysis, this can be the case for orthologs from worm, human and fly or for metazoan, fungi and green plant orthologous sequences. The HOPS approach accounts for this by the way how orthology is analyzed. Speciation events from the same level as the groups being analyzed are not used to assign orthology. Only speciation events for sequences from higher levels, this means more phylogenetic distant speciation events, are included in the analysis.

However, not only ambiguousness' in the species tree can result in wrong assignments. Six percent of the cases RIO failed to assign correctly were due to possible co-evolution of proteins from intracellular parasites. Plasmodium falciparum is an example for this. Established taxonomy for Plasmodium falciparum places it basal to a metazoan / yeast clade (Hashimoto, T. *et al.* 1994). But due to its parasitic life form, some proteins from Plasmodium falciparum tend to co-evolve with the proteins of the host (Jiggins *et al.* 2002). Gene trees reconstructed from such sequences will place the Plasmodium falciparum proteins incorrectly on a branch with proteins from the host (Hafner and Nadler 1988). If these gene trees are reconciled with the species tree, the misplaced P. falciparum sequence prevents assignment of orthology between proteins from the host and other species (fig. 6).

The assignment of orthology was further improved by minimizing the size of the gene trees. In a first approach the bootstrap trees were only calculated once for each Pfam family, using all sequences of that family. These trees were then analyzed for orthologs multiple times. Only the group assignments were changed, based on the different species pairings. This resulted in low ortholog bootstrap support values.
In general large trees are known to have a lower bootstrap support than small trees (Zharkikh and Li 1995). Excluding sequences from the alignment that added no information to the assignment of orthology for a given species pairing resulted in more reliable trees and therefore higher ortholog bootstrap values. In this approach new bootstrap trees were calculated for every species pairing.

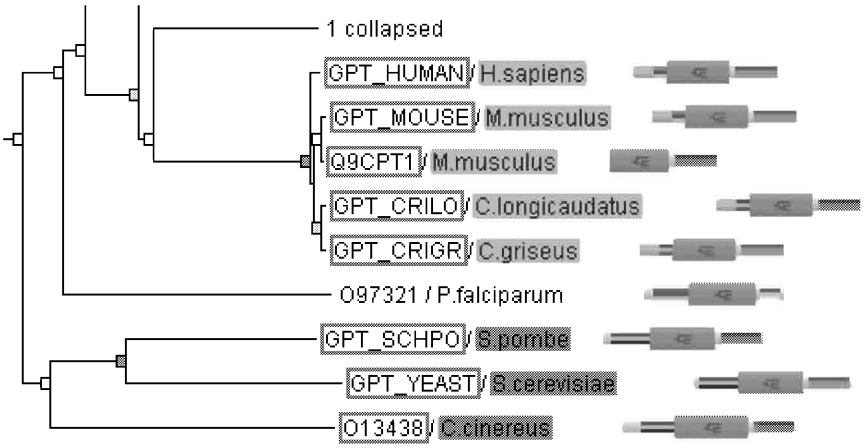

Figure 6 Example for host-parasite co-evolution

The Plasmodium falciparum protein O97321 is on a branch with metazoan proteins. If included in the analysis this would prevent orthology assignment between the metazoan and fungi sequences. The domain shown in this example is glycosyl transferase (Glycos_transf_4).

5 Conclusions & Perspectives

"Ach, Luise, laß ... das ist ein zu weites Feld."
- Theodor Fontane, Effi Briest

Similarity-based approaches are today the main methods used for orthology inference. They have, like the UPGMA algorithm for tree reconstruction, an underlying molecular clock assumption. Tree based approaches that allow incorporating more advanced evolutionary models, like tree reconciliation or our 'Orthostrapper', are available, but not widely used yet. With the application of the bootstrap method to orthology inference a tool for calculating orthology support values is available. In addition a new generation for orthology inference in a Bayesian framework has just been developed.

The work done for the HOPS database shows that there are still problems to overcome. Bacterial sequences are excluded in HOPS because of horizontal gene transfer. A recent study suggests that up to 20% of bacterial sequences underwent horizontal gene transfer (Snel *et al.* 2002). While for functional inferences ortholog miss-assignments due to horizontal gene transfer might not be problematic, it would mislead any comparative or evolutionary study. Methods for reliably distinguishing xenologs from orthologs would improve the potential of the orthology / paralogy concept for bacterial species.

For HOPS we chose to use the alignments provided by Pfam for orthology inference. These alignments can be rather short, resulting in ambiguous trees. A more preferable way would be to use alignments from full-length sequences, but automated multiple alignment creation still gives inferior results. An important step to improve ortholog detection from phylogenetic trees would therefore be the development of more sophisticated multiple alignment methods. Ideally these methods should be able to correctly align sequences with repeats, large extensions and insertions, multiple domains and circular permutations.

Any method for inferring orthologs tries to find sequences that are separated by speciation. The evolutionary model used to infer orthology limits the conclusions that can be derived from studying orthologs.
Therefore it is beneficial to use more advanced evolutionary models in the reconstruction of the trees. Programs to do this exist, but are limited by the

current computer power available. Super-computers consisting of large numbers of cheap PC's are revolutionizing the price of computational power. Re-implementation of existing programs to run on parallel system would improve their usability for a large-scale analysis.
In combination with the availability of automated phylogenetic approaches for assigning orthology and the ongoing developments it will be interesting to see if phylogenetic methods for inferring orthology replace similarity based approaches in the near future.

Another challenge will be the clustering of orthologs from multiple species. Some methods, for instance COGs and OrthoMCL, already group orthologs from multiple species. However, they do not account for the intransitivity of orthology, resulting in a sometimes simplified view of the actual orthologous and paralogous relations. While this simplified view is acceptable for applications such as phylogenetic profiles, it does not always suit the demands for detailed comparative studies. Applications making use of orthologous relations could benefit from a new way of representing orthology. This representation should on the one hand capture the complexity of all possible orthologous relations. On the other hand it should be capable of efficiently summarizing orthologous relations for applications such as phylogenetic profiles or whole genome comparisons between multiple species.

6 Acknowledgements

"Doch hängt mein ganzes Herz an dir,
Du graue Stadt am Meer;
Der Jugend Zauber für und für
Ruht lächelnd doch auf dir, auf dir,
Du graue Stadt am Meer."

- Theodor Storm, Die Stadt

I would like to thank Erik Sonnhammer for supervising the Ph. D. project and for encouraging me to develop and realize my ideas.

Boris Lenhard for providing office space and advice in the final phase of the project.

Anthony Brooks, without whose help the thesis might never have been finished.

Lars Arvestad for his patience in answering my mathematical questions, his valuable comments and interesting scientific discussions.

My time at the CGB would have been a lot less interesting without Alistair Chalk. Thanks for the help on organizational and computer related issues, bearing my cynical comments over the years, and last but not least the numerous parties.

Special thanks also to Jennifer Lee for sharing her vast knowledge of green plant evolution, the delicious cookies and making the coffee breaks so enjoyable.

Volker Hollich for his introduction to SQL.

All the other members of the exploratory bioinformatics groups, present and past: Timo Lassman, Markus Wistrand, Lukas Käll, Saraswathi Abhiman, Gang Liu, Timothy Bailey, Raf Podowski and Michael Åsman. Thanks for the

Acknowledgements

help, stimulating scientific discussions and all the nice memories I will take home!

Anna Polgren for the wonderful kayak trips.

Peer Angström and Albin Sandlin for sharing their office with me during the last six month.

Bill Wilson for being nice about my buggy perl scripts and his help in my struggle with the various KI registration forms.

Bengt Sennblad for interesting discussions on evolution.

Mark Reimers for his support.

The people from the general science course - especially Tatjana Steiler, Eszter Somogyi and James O' Brian - who made the course so much more interesting.

Patricia Dwight, Emily Hodges and the "Irish twins" Shan McCarthy & Allan McKenna for all the nice chats, Cecilia Bosdotter for her superb cocktails, David Fredman for helping me with Access, Yvonne Kallberg and Erik Nordling for the nice time we spend during ISMB at the beach, err, talks.

An unknown gypsy woman for a prophecy she made to my grandmother nearly 80 years ago.

All my friends from Germany for keeping in touch with me over the years.

My family (all of the Storm and Maybauer clans) for allowing me to relax and refuel whenever I went home.

And finally to my wife Berit for her unwavering support throughout our stay in Sweden.

7 References

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y. H. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor, J. F. Abril, A. Agbayani, H. J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M. H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G. H. Karpen, Z. Ke, J. A. Kennison, K. A. Ketchum, B. E. Kimmel, C. D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. C. McIntosh, M. P. McLeod, D. McPherson, G. Merkulov, N. V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S. M. Mount, M. Moy, B. Murphy, L. Murphy, D. M. Muzny, D. L. Nelson, D. R. Nelson, K. A. Nelson, K. Nixon, D. R. Nusskern, J. M. Pacleb, M. Palazzolo, G. S. Pittman, S. Pan, J. Pollard, V. Puri, M. G. Reese, K. Reinert, K. Remington, R. D. Saunders, F. Scheeler, H. Shen, B. C. Shue, I. Siden-Kiamos, M. Simpson, M. P. Skupski, T. Smith, E. Spier, A. C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A. H. Wang, X. Wang, Z. Y. Wang, D. A. Wassarman, G. M. Weinstock, J. Weissenbach, S. M. Williams, WoodageT, K. C. Worley, D. Wu, S. Yang, Q. A. Yao, J. Ye, R. F. Yeh, J. S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. H. Zheng, F. N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. O. Smith, R. A. Gibbs, E. W. Myers, G. M. Rubin and J. C. Venter (2000). "The genome sequence of Drosophila melanogaster." *Science* **287**: 2185-2195.

References

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res* **25**: 3389-3402.

Andrade, M. A., N. P. Brown, C. Leroy, S. Hoersch, A. de Daruvar, C. Reich, A. Franchini, J. Tamames, A. Valencia, C. Ouzounis and C. Sander (1999). "Automated genome sequence analysis and annotation." *Bioinformatics* **15**: 391-412.

Arvestad, L., A. Berglund, J. Lagergren and B. Sennblad (2003). "Bayesian gene/species tree reconciliation and orthology analysis using MCMC." *Bioinformatics Suppl.* **19**: i7-i15

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall and E. L. Sonnhammer (2002). "The Pfam protein families database." *Nucleic Acids Res* **30**: 276-280.

Benson D. A. , I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, D. L. Wheeler (2003). "GenBank. " *Nucleic Acids Res* **31**:23-27.

Bhatia, U., K. Robison and W. Gilbert (1997). "Dealing with database explosion: a cautionary note." *Science* **276**(5319): 1724-1725.

Blanchette, M. and M. Tompa (2002). "Discovery of regulatory elements by a computational method for phylogenetic footprinting." *Genome Res* **12**: 739-748.

Bork, P., T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen and Y. Yuan (1998). "Predicting function: from genes to genomes and back." *J Mol Biol* **283**: 707-725.

Braun, E. L., A. L. Halpern, M. A. Nelson and D. O. Natvig (2000). "Large-scale comparison of fungal sequence information: mechanisms of innovation in Neurospora crassa and gene loss in Saccharomyces cerevisiae." *Genome Res* **10**: 416-430.

Brudno M., S. Malde, A. Poliakov, C. B. Do, O. Couronne, I. Dubchak, S. Batzoglou (2003) "Glocal alignment: finding rearrangements during alignment. "*Bioinformatics Suppl.* **19**:I54-I62

C. elegans Sequencing Consortium (1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology." *Science* **282**: 2012-2018.

Copley, R. R., I. Letunic and P. Bork (2002). "Genome and protein evolution in eukaryotes." *Curr Opin Chem Biol* **6**: 39-45.

Doolittle, R. F., D. F. Feng, S. Tsang, G. Cho and E. Little (1996). "Determining divergence times of the major kingdoms of living organisms with a protein clock." *Science* **271**: 470-477.

References

Doolittle, W. F. (1999). "Phylogenetic classification and the universal tree." *Science* **284**: 2124-2129.

van Dongen, S. (2000). "Graph clustering by flow simulation." Ph.D thesis, University of Utrecht, The Netherlands.

Durbin, R., S. Eddy, A. Krogh and G. Mitchison (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press.

Efron, B., E. Halloran and S. Holmes (1996). "Bootstrap confidence levels for phylogenetic trees." *Proc Natl Acad Sci U S A* **93**: 13429-13434.

Efron, B. and R. J. Tibshirani (1993). Monographs on Statistics and Applied Probability: An Introduction to the Bootstrap, Chapman & Hall.

Eisen, J. A. (1998). "Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis." *Genome Res* **8**: 163-167.

Eisen, J. A. and M. Wu (2002). "Phylogenetic analysis and gene functional predictions: phylogenomics in action." *Theor Popul Biol* **61**: 481-487.

Enright, A. J. and C. A. Ouzounis (2000). "GeneRAGE: a robust algorithm for sequence clustering and domain detection." *Bioinformatics* **16**: 451-457.

Enright, A. J., S. Van Dongen and C. A. Ouzounis (2002). "An efficient algorithm for large-scale detection of protein families." *Nucleic Acids Res* **30**: 1575-1584.

Epp, C. D. (1997). "Definition of a gene." *Nature* **389**: 537.

Eulenstein, O. (1998). Vorhersage von Genduplikationen und deren Entwicklung in der Evolution. GMD - Forschungszentrum Informationstechnik GMBH.

Felsenstein, J. (1985). "Confidence limits on phylogenies: an approach using the bootstrap." *Evolution* **39**: 783-791.

Feng, D. F., G. Cho and R. F. Doolittle (1997). "Determining divergence times with a protein clock: update and reevaluation." *Proc Natl Acad Sci U S A* **94**: 13028-13033.

Fitch, W. M. (1970). "Distinguishing homologous from analogous proteins." *Syst Zool* **19**: 99-113.

Fitch, W. M. (2000). "Homology a personal view on some of the problems." *Trends Genet* **16**: 227-231.

Gandbhir, M., I. Rasched, P. Marliere and R. Mutzel (1995). "Convergent evolution of amino acid usage in archaebacterial and eubacterial lineages adapted to high salt." *Res Microbiol* **146**: 113-120.

Gerlt, J. A. and P. C. Babbitt (2000). "Can sequence determine function?" *Genome Biol* **1**: Reviews 0005.

References

Gogarten, J. P. and L. Olendzenski (1999). "Orthologs, paralogs and genome comparisons." *Curr Opin Genet Dev* **9**: 630-636.

Gonczy, P., C. Echeverri, K. Oegema, A. Coulson, S. J. Jones, R. R. Copley, J. Duperon, J. Oegema, M. Brehm, E. Cassin, E. Hannak, M. Kirkham, S. Pichler, K. Flohrs, A. Goessen, S. Leidel, A. M. Alleaume, C. Martin, N. Ozlu, P. Bork and A. A. Hyman (2000). "Functional genomic analysis of cell division in C. elegans using RNAi of genes on chromosome III." *Nature* **408**: 331-336.

Goodman, M., J. Czelusniak, G. W. Moore, A. E. Romero-Herrera and G. Matsuda (1979). "Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences." *Systematic Zoology* **28**: 132-168.

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

Hafner, M. S. and S. A. Nadler (1988). "Phylogenetic trees support the coevolution of parasites and their hosts." *Nature* **332**: 258-259.

Haft, D. H., B. J. Loftus, D. L. Richardson, F. Yang, J. A. Eisen, I. T. Paulsen and O. White (2001). "TIGRFAMs: a protein family resource for the functional identification of proteins." *Nucleic Acids Res* **29**: 41-43.

Haney, P. J., J. H. Badger, G. L. Buldak, C. I. Reich, C. R. Woese and G. J. Olsen (1999). "Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species." *Proc Natl Acad Sci U S A* **96**: 3578-3583.

Hashimoto, T., Y. Nakamura, F. Nakamura, T. Shirakura, J. Adachi, N. Goto, K. Okamoto and M. Hasegawa (1994). "Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, Giardia lamblia." *Mol Biol Evol* **11**: 65-71.

Hegyi, H. and P. Bork (1997). "On the classification and evolution of protein modules." *J Protein Chem* **16**: 545-551.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." *Proc Natl Acad Sci U S A* **89**: 10915-10919.

Henikoff, S., J. G. Henikoff and S. Pietrokovski (1999). "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations." *Bioinformatics* **15**: 471-479.

Hertz, G. Z. and G. D. Stormo (1999). "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences." *Bioinformatics* **15**: 563-577.

References

Huelsenbeck, J. P. and F. Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees." *Bioinformatics* **17**: 754-755.

Jensen L. J., D. W. Ussery, S. Brunak (2003) "Functionality of system components: conservation of protein function in protein feature space." *Genome Res.* **13**:2444-2449.

Jiggins, F. M., G. D. Hurst and Z. Yang (2002). "Host-symbiont conflicts: positive selection on an outer membrane protein of parasitic but not mutualistic Rickettsiaceae." *Mol Biol Evol* **19**: 1341-1349.

Karp, P. D. (1998). "What we do not know about sequence analysis and sequence databases." *Bioinformatics* **14**: 753-754.

Karp P. D., S. Paley, J. Zhu (2001) "Database verification studies of SWISS-PROT and GenBank." *Bioinformatics* **17**:526-532;

Kondrashov, F. A., I. B. Rogozin, Y. I. Wolf and E. V. Koonin (2002). "Selection in the evolution of gene duplications." *Genome Biol* **3**: Research 0008.

Koonin, E. V. (2001). "An apology for orthologs - or brave new memes." *Genome Biol* **2**: 1005.

Koonin, E. V., Y. I. Wolf and G. P. Karev (2002). "The structure of the protein universe and genome evolution." *Nature* **420**: 218-223.

Koski, L. B. and G. B. Golding 2001. "The closest BLAST hit is often not the nearest neighbor. " *J Mol Evol.* **52**:540-542.

Krause, A., J. Stoye and M. Vingron (2000). "The SYSTERS protein sequence cluster set." *Nucleic Acids Res* **28**: 270-272.

Krivan, W. and W. W. Wasserman (2001). "A predictive model for regulatory sequences directing liver-specific transcription." *Genome Res* **11**: 1559-1566.

Kumar, S. and S. R. Gadagkar (2000). "Efficiency of the neighbor-joining method in reconstructing deep and shallow evolutionary relationships in large phylogenies." *J Mol Evol* **51**: 544-553.

Lander, E. S., 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,

References

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, 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, J. Szustakowki, 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**: 860-921.

References

Lecompte, O., J. D. Thompson, F. Plewniak, J. Thierry and O. Poch (2001). "Multiple alignment of complete sequences (MACS) in the post-genomic era." *Gene* **270**: 17-30.

Lenhard, B., A. Sandelin, L. Mendoza, P. Engström, N. Jareborg and W. W. Wasserman (2003) "Identification of conserved regulatory elements by comparative genome analysis" *J Bio.* **2**: 13

Lijavetzky D., P. Carbonero, J, Vicente-Carbajosa(2003). "Genome-wide comparative phylogenetic analysis of the rice and Arabidopsis Dof gene families. " *BMC Evol Biol.* **3**:17-28.

Li L., C. J. Stoeckert Jr, D. S. Roos (2003) "OrthoMCL: identification of ortholog groups for eukaryotic genomes. " *Genome Res* **13**:2178-2189.

Lipman, D. J. and W. R. Pearson (1985). "Rapid and sensitive protein similarity searches." *Science* **227**: 1435-1441.

Martin, A. P. and T. M. Burg (2002). "Perils of paralogy: using HSP70 genes for inferring organismal phylogenies." *Syst Biol* **51**: 570-587.

Mering C., M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, B. Snel (2003). "STRING: a database of predicted functional associations between proteins. " *Nucleic Acids Res* **31**:258-261.

Morrison, D. A. and J. T. Ellis (1997). "Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa." *Mol Biol Evol* **14**: 428-441.

Mulder N. J., R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R. R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S. E. Orchard, M. Pagni, D. Peyruc, C. P. Ponting, J. D. Selengut, F. Servant, C. J. Sigrist, R. Vaughan, E. M. Zdobnov (2003). "The InterPro Database, 2003 brings increased coverage and new features." *Nucleic Acids Res* **31**:315-318.

Muller, W. E., M. Bohm, V. A. Grebenjuk, A. Skorokhod, I. M. Muller and V. Gamulin (2002). "Conservation of the positions of metazoan introns from sponges to humans." *Gene* **295**: 299-309.

Mushegian, A. R., J. R. Garey, J. Martin and L. X. Liu (1998). "Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes." *Genome Res* **8**: 590-598.

Nei, M. (1996). "Phylogenetic analysis in molecular evolutionary genetics." *Annu Rev Genet* **30**: 371-403.

References

Nei, M., S. Kumar and K. Takahashi (1998). "The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small." *Proc Natl Acad Sci U S A* **95**: 12390-12397.

Neuwald, A. F., J. S. Liu, D. J. Lipman and C. E. Lawrence (1997). "Extracting protein alignment models from the sequence database." *Nucleic Acids Res* **25**: 1665-1677.

Notredame, C. (2002). "Recent progress in multiple sequence alignment: a survey." *Pharmacogenomics* **3**: 131-144.

Oren, A. (1995). "Comment on "Convergent evolution of amino acid usage in archaebacterial and eubacterial lineages adapted to high salt", by M. Gandbhir et al. (Res. Microbiol., 1995, 146, 113-120)." *Res Microbiol* **146**: 805-806.

Page, R. D. (1998). "GeneTree: comparing gene and species phylogenies using reconciled trees." *Bioinformatics* **14**: 819-820.

Page, R. D. (2000). "Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny." *Mol Phylogenet Evol* **14**: 89-106.

Page, R. D. M. (1994). "Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas." *Systematic Biology* **43**: 58-77.

Pearson, W. R. and D. J. Lipman (1988). "Improved tools for biological sequence comparison." *Proc Natl Acad Sci U S A* **85**: 2444-2448.

Pellegrini, M, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. " *Proc Natl Acad Sci U S A* **96**:4285-4288.

Petsko, G. A. (2001). "Homologuephobia." *Genome Biol* **2**.

Phillips, A., D. Janies and W. Wheeler (2000). "Multiple sequence alignment in phylogenetic analysis." *Mol Phylogenet Evol* **16**: 317-330.

Ranson, H., C. Claudianos, F. Ortelli, C. Abgrall, J. Hemingway, M. V. Sharakhova, M. F. Unger, F. H. Collins and R. Feyereisen (2002). "Evolution of supergene families associated with insecticide resistance." *Science* **298**: 179-181.

Remm, M. and E. Sonnhammer (2000). "Classification of transmembrane protein families in the Caenorhabditis elegans genome and identification of human orthologs." *Genome Res* **10**: 1679-1689.

Rogers, J. S. (1997). "On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences." *Syst Biol* **46**: 354-357.

References

Sato N. (2003). "Comparative analysis of the genomes of cyanobacteria and plants." *Genome Inform Ser Workshop Genome Inform.* **13**:173-182.

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

Sander, C. and R. Schneider (1991). "Database of homology-derived protein structures and the structural meaning of sequence alignment." *Proteins* **9**: 56-68.

Servant, F., C. Bru, S. Carrere, E. Courcelle, J. Gouzy, D. Peyruc and D. Kahn (2002). "ProDom: automated clustering of homologous domains." *Brief Bioinform* **3**: 246-251.

Silverstein, K. A., E. Shoop, J. E. Johnson, A. Kilian, J. L. Freeman, T. M. Kunau, I. A. Awad, M. Mayer and E. F. Retzel (2001). "The MetaFam Server: a comprehensive protein family resource." *Nucleic Acids Res* **29**: 49-51.

Snel, B., P. Bork and M. A. Huynen (2002). "Genomes in flux: the evolution of archaeal and proteobacterial gene content." *Genome Res* **12**: 17-25.

Sonnhammer, E. L. and D. Kahn (1994). "Modular arrangement of proteins as inferred from analysis of homology." *Protein Sci* **3**: 482-492.

Sonnhammer, E. L. and E. V. Koonin (2002). "Orthology, paralogy and proposed classification for paralog subtypes." *Trends Genet* **18**: 619-620.

Stoesser G., W. Baker, A. van den Broek, M. Garcia-Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, F. Nardone, P. Stoehr, M. A. Tuli, K. Tzouvara, R. Vaughan (2003). "The EMBL Nucleotide Sequence Database: major new developments." *Nucleic Acids Res.* **31**:17-22.

Studier, J. A. and K. J. Keppler (1988). "A note on the neighbor-joining algorithm of Saitou and Nei." *Mol Biol Evol* **5**: 729-731.

Tagle, D. A., B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess and R. T. Jones (1988). "Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints." *J Mol Biol* **203**: 439-455.

Tatusov R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, D. A. Natale (2003). "The COG database: an updated version includes eukaryotes." *BMC Bioinformatics* **4**:41-55.

References

Tatusov, R. L., M. Y. Galperin, D. A. Natale and E. V. Koonin (2000). "The COG database: a tool for genome-scale analysis of protein functions and evolution." *Nucleic Acids Res* **28**: 33-36.

Tatusov, R. L., E. V. Koonin and D. J. Lipman (1997). "A genomic perspective on protein families." *Science* **278**: 631-637.

Thompson, J. D., F. Plewniak and O. Poch (1999). "A comprehensive comparison of multiple sequence alignment programs." *Nucleic Acids Res* **27**: 2682-2690.

Thompson, J. D., F. Plewniak, J. Thierry and O. Poch (2000). "DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches." *Nucleic Acids Res* **28**: 2919-2926.

Thorne, J. L. and H. Kishino (1992). "Freeing phylogenies from artifacts of alignment." *Mol Biol Evol* **9**: 1148-1162.

Thorne, J. L., H. Kishino and J. Felsenstein (1991). "An evolutionary model for maximum likelihood alignment of DNA sequences." *J Mol Evol* **33**: 114-124.

Thornton, J. W. and R. DeSalle (2000). "Gene family evolution and homology: genomics meets phylogenetics." *Annu Rev Genomics Hum Genet* **1**: 41-73.

Venter, J. C., 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,

References

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**: 1304-1351.

Walhout, A. J., R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg and M. Vidal (2000). "Protein interaction mapping in C. elegans using proteins involved in vulval development." *Science* **287**: 116-122.

Wall D. P., H.B. Fraser, A. E. Hirsh (2003) "Detecting putative orthologs. " *Bioinformatics* **19**:1710-1711.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R.

References

David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigo, M. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P. Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. P. Vinson, A. C. Von Niederhausern, C. M. Wade, M. Wall, R. J. Weber, R. B. Weiss, M. C. Wendl, A. P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. K. Wilson, E. Winter, K. C. Worley, D. Wyman, S. Yang, S. P. Yang, E. M. Zdobnov, M. C. Zody and E. S. Lander (2002). "Initial sequencing and comparative analysis of the mouse genome." *Nature* **420**: 520-562.

Whelan, S., P. Lio and N. Goldman (2001). "Molecular phylogenetics: state-of-the-art methods for looking into the past." *Trends Genet* **17**: 262-272.

Wicker, N., G. R. Perrin, J. C. Thierry and O. Poch (2001). "Secator: a program for inferring protein subfamilies from phylogenetic trees." *Mol Biol Evol* **18**: 1435-1441.

References

Wingender, E., X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhauser, M. Pruss, F. Schacherer, S. Thiele and S. Urbach (2001). "The TRANSFAC system on gene expression regulation." *Nucleic Acids Res* **29**: 281-283.

Woese, C. R. (1987). "Bacterial evolution." *Microbiol Rev* **51**: 221-271.

Woese, C. R. and G. E. Fox (1977). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." *Proc Natl Acad Sci U S A* **74**: 5088-5090.

Wong, P., G. Kolesov, D. Frishman, W. A. Houry (2003). "Phylogenetic web profiler. " *Bioinformatics* **19**:782-783.

Wu, J. and M. Grunstein (2000). "25 years after the nucleosome model: chromatin modifications." *Trends Biochem Sci* **25**: 619-623.

Xie, T. and D. Ding (2000). "Investigating 42 candidate orthologous protein groups by molecular evolutionary analysis on genome scale." *Gene* **261**: 305-310.

Yanai, I., Y. I. Wolf and E. V. Koonin (2002). "Evolution of gene fusions: horizontal transfer versus independent events." *Genome Biol* **3**.

Yona, G., N. Linial and M. Linial (2000). "ProtoMap: automatic classification of protein sequences and hierarchy of protein families." *Nucleic Acids Res* **28**: 49-55.

Zdobnov, E. M., C. von Mering, I. Letunic, D. Torrents, M. Suyama, R. R. Copley, G. K. Christophides, D. Thomasova, R. A. Holt, G. M. Subramanian, H. M. Mueller, G. Dimopoulos, J. H. Law, M. A. Wells, E. Birney, R. Charlab, A. L. Halpern, E. Kokoza, C. L. Kraft, Z. Lai, S. Lewis, C. Louis, C. Barillas-Mury, D. Nusskern, G. M. Rubin, S. L. Salzberg, G. G. Sutton, P. Topalis, R. Wides, P. Wincker, M. Yandell, F. H. Collins, J. Ribeiro, W. M. Gelbart, F. C. Kafatos and P. Bork (2002). "Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster." *Science* **298**: 149-159.

Zharkikh, A. and W. H. Li (1995). "Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique." *Mol Phylogenet Evol* **4**: 44-63.

Zmasek, C. M. and S. R. Eddy (2001). "A simple algorithm to infer gene duplication and speciation events on a gene tree." *Bioinformatics* **17**: 821-828.

Zmasek, C. M. and S. R. Eddy (2002). "RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs." *BMC Bioinformatics* **3**: 14.

Web Site References

http://www.rio.wustl.edu/; RIO
http://tolweb.org/tree/phylogeny.html; Tree of Life Project