The relationship between orthology, protein domain architecture, and protein function

Kristoffer Forslund

Stockholm University

The relationship between orthology, protein domain architecture and protein function

Kristoffer Forslund

ISBN 978-91-7447-350-6

Printed in Sweden by US-AB, Stockholm 2011
Distributor: Department of Biochemistry and Biophysics

Dedicated to Dr Knut Åhs,
adored grandfather,
eternal role model.

List of publications

Publications included in this thesis

Paper I: Forslund K, Henricson A, Hollich V, Sonnhammer EL. Domain tree-based analysis of protein architecture evolution. Molecular Biology and Evolution 2008;25:254-64.

Paper II: Forslund K, Sonnhammer EL. Predicting protein function from domain content. Bioinformatics 2009;24:1681-7.

Paper III: Forslund K, Sonnhammer EL. Benchmarking homology detection procedures with low complexity filters. Bioinformatics 2009;25:2500-5.

Paper IV: Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer EL. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Research 2009;38:196-203.

Paper V: Henricson A, Forslund K, Sonnhammer ELL. Orthology confers intron position conservation. BMC Genomics 2010;11:412-25.

Paper VI: Forslund K, Pekkari I, Sonnhammer ELL. Domain architecture conservation in orthologs. BMC Bioinformatics 2011;12:326.

Other publications

Forslund K, Sonnhammer ELL. Evolution of protein domain architectures. In Anisimova M (Ed.), Evolutionary Genomics: statistical and computational methods. New York: Springer-Humana 2011, in press.

Forslund K, Schreiber F, Thanintorn N, Sonnhammer ELL. OrthoDisease: tracking disease gene orthologs across 100 species. Briefings in Bioinformatics 2011.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A. The Pfam protein families database. Nucleic Acids Research 2010;38:211-22.

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. The Pfam protein families database. Nucleic Acids Research 2008;36:281-8.

Grünewald S, Forslund K, Dress A, Moulton V. QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets. Molecular Biology and Evolution 2007;24:532-8.

Forslund K, Huson DH, Moulton V. VisRD - visual recombination detection. Bioinformatics. 2004 20:3654-3655.

Strimmer K, Forslund K, Holland B, Moulton V. A novel exploratory method for visual recombination detection. Genome Biol. 2003;4:33.

Abstract

Lacking experimental data, protein function is often predicted from evolutionary and protein structure theory. Under the 'domain grammar' hypothesis the function of a protein follows from the domains it encodes. Under the 'orthology conjecture', orthologs, related through species formation, are expected to be more functionally similar than paralogs, which are homologs in the same or different species descended from a gene duplication event. However, these assumptions have not thus far been systematically evaluated.

To test the 'domain grammar' hypothesis, we built models for predicting function from the domain combinations present in a protein, and demonstrated that multi-domain combinations imply functions that the individual domains do not. We also developed a novel gene-tree based method for reconstructing the evolutionary histories of domain architectures, to search for cases of architectures that have arisen multiple times in parallel, and found this to be more common than previously reported.

To test the 'orthology conjecture', we first benchmarked methods for homology inference under the obfuscating influence of low-complexity regions, in order to improve the InParanoid orthology inference algorithm. InParanoid was then used to test the relative conservation of functionally relevant properties between orthologs and paralogs at various evolutionary distances, including intron positions, domain architectures, and Gene Ontology functional annotations.

We found an increased conservation of domain architectures in orthologs relative to paralogs, in support of the 'orthology conjecture' and the 'domain grammar' hypotheses acting in tandem. However, equivalent analysis of Gene Ontology functional conservation yielded spurious results, which may be an artifact of species-specific annotation biases in functional annotation databases. I discuss possible ways of circumventing this bias so the 'orthology conjecture' can be tested more conclusively.

Contents

Abbreviations

BLAST	Basic Local Alignment Search Tool
DAS	Distributed Annotation Service
DDC	Duplication-Degeneration-Complementation
EC	Enzyme Classification
GBA	Guilt By Association
GO	Gene Ontology
GPD	Generalized Pareto Distribution
HMM	Hidden Markov Model
JTO	Jaccard-normalized Term Overlap
LUCA	Last Universal Common Ancestor
MCMC	Markov Chain Monte Carlo
OTU	Operational Taxonomic Unit
SVM	Support Vector Machine
TO	Term Overlap

1 Introduction

1.1 Purpose

Modern biological science aims at improving conditions for humanity, making it possible for us to live longer, healthier lives. To accomplish this, we need to be able to measure, understand, and predict how life functions, ranging from human life to that of all other organisms that affect it, such as pathogens or beneficial symbionts. The high-throughput biology revolution, with technologies that include expression and genomic microarrays, as well as large-scale genomic and transcriptomic sequencing, has granted us an understanding of the building blocks of organisms, and the next step then becomes understanding how they relate and interact, paving the path to full systems biology. While wet lab experiments are the Alpha and Omega of biology, practical, technological and ethical constraints necessitate a computational effort, the field of bioinformatics or computational biology, to help us go from disjointed facts to an understanding of the functional whole.

The work described here aims to promote this goal. I have intended to explore, evaluate and improve methods for computationally characterizing how the individual components of life act and interrelate. In the course of this effort, I have studied the underlying evolution of living organisms, because it is from evolutionary relationships that many biological hypotheses and conclusions are arrived at. More specifically, the underlying question that I have been trying to address is this: to what degree exists there easily measured, general properties of genes and proteins that can help us understand what they do and what processes they are part of?

By easily measured, general properties, I mean attributes that can be tested in high-throughput studies without any hypothesis on function *a priori*. There are many such properties that might be useful in this manner. I have focused on two: the presence of recurring protein sequence or structure elements in the form of *protein domains*, and the specific phylogenetic relationship of *orthology*, which holds for a pair of genes in two different species if they stem from a single gene in the last common ancestor (the

cenancestor) of those species. Orthology is complemented by the relationship of *paralogy*, which holds between any genes that descended by duplication with subsequent descent from a common ancestor.

In terms of what proteins do and what processes their products are part of, I have focused on whether or not those proteins can be sorted into broad or narrow functional categories, and on their involvement in human genetic diseases [1, 2]. While constantly expanding, these classifications are still crude, but it is my hope that any property that can be used to predict involvement in such crude categories might also be useful in predicting involvement in finer-grained categories, or at the very least narrow down the search space.

Within this problem area, I have investigated what functional information is available from the properties of protein domains and orthology relationships, and on how these two properties relate to each other, wherein selective pressure towards retaining an ancestral protein function might form the elusive hidden link. We have good reason to believe that orthology often confers functional similarity between proteins, and I have found that it also confers relatively higher conservation of domain architecture.

One interpretation of this is that orthologous evolution of proteins is associated with selective pressure towards retaining some ancestral function, which is achieved by making changes in the domain content of proteins less likely. This, then, provides support for the idea that many of the functions of proteins take place because of the domains they contain, either as direct or indirect consequences. This idea is very widely accepted, but not conclusively proven. An alternate hypothesis could be presented, under which analogous functions could be implemented equally well by proteins with very different structures, along very different pathways, in which case the connection between specific domains and specific functions would follow not from structural necessity but merely from shared ancestry and historical happenstance. If this would be the case, it would limit the conclusions that we could draw from domain content alone, and so the question has bearing on the larger problem of determining which functional conclusions we can draw reliably from which properties.

Foremost, I have thus wanted to test two commonly endorsed hypotheses. The first has been termed the *orthology conjecture* [3] or *standard model* [4], and states that proteins without paralogs tend to change in function more slowly as they evolve, unlike the case for proteins with paralogs, where functional redundancy following from the presence of gene duplicates would reduce selective pressure. The second could be called the *domain grammar hypothesis* [Paper II, 5], and I will consider a strong and weak form of it.

The weak form simply states that domain architectures contain information that can be used to predict the functions of proteins. The strong form states that the functions of proteins causally follow from their domain architectures, so that a given function is guaranteed to be achieved by a protein combining the proper domains, and, moreover, that the function in question cannot in fact be implemented without these domains being present. The weak form follows from the strong form, and is relevant only for practical purposes of designing computational function prediction pipelines. The strong form, as indicated above, may or may not hold true, and it is likely that it does so only in part: there may be some protein functions that follow as a necessary consequence of particular domain combinations, and of these, some may be impossible to achieve in any protein lacking that domain combination. If this is the case, determining how often a function is tied by necessity to a particularly domain combination becomes relevant.

In summary, in order to improve our knowledge of the functional roles of genes and proteins, I have studied the evolution of proteins to test which impact various factors have on these functional roles. Along the way, this work has serendipitously resulted in improvement to certain bioinformatics tools and resources, and alerted me to a number of open questions and potential sources of bias that may form obstacles we should strive to overcome in order to gain a clearer picture.

1.2 Conventions

Throughout this work, in some contexts, references may be made interchangingly to genes and proteins, or to genomes and proteomes, notably with respect to orthology and other phylogenetic relationships, and to function. In these cases, this is to be understood as referring either to gene sequences or to their encoded and corresponding amino acid sequences, implying that the reasoning employed could be applied at either level. Any references to the function of a gene should be interpreted as referring to the function of its product. Likewise, any references to mutation, evolution or duplication of a protein refers to these events happening to the gene encoding it.

2 Background

2.1 Evolution and orthology

2.1.1 Homology

The term *homology* was first used by Owen [6] as referring to “the same organ in different animals under every variety of form and function”. That is to say: under this definition, a part of an organism is homologous to a part of another organism if they are in some sense “the same organ” and do the same thing. This is in fact a statement of the properties of the extant organism, rather than its origins, which is unsurprising given that this definition preceded Darwin's [7] concept of evolution by descent with modification under natural selection. The term, however, has come to be adapted [8] to an evolutionary framework and given a revised definition, which has nothing to do with what a biological trait does, at least directly, and everything to do with where it came from. This more recent definition of homology states that a part of some organism is homologous to another part of some organism if they both evolved through descent from a part found in some shared ancestor organism.

This work exclusively concerns homology as the term is used within molecular evolution. For a discussion on how this terminology translates to morphological evolution, see Patterson [9]. Within this framework, we can talk of homology at all levels of genetic materials, as well as indirectly at the level of the encoded proteins. Single nucleotides can be homologs, as can all or part of the genes they form, and the chromosomes where they are found. Perhaps ironically, given Owen’s definition, it is incorrect to refer to two parts (genes, organs etc.) as homologs if they did not evolve from the same ancestral part, even if they accomplish the same things for an organism, and they should then instead be considered *analogs* [9, 10].

2.1.2 Inferring homology

At the core of inferring homology between a pair of sequences lie implicit or explicit statistical models. These reveal when certain levels of observed similarity between sequences become unrealistic in the absence of a common origin [9, 10]. Generally, the models are defined relative to an *alignment* of the sequences, which is a set of hypotheses on which characters are descended from the same common ancestral characters. In the absence of a possible alignment (though structural features can often be aligned even when sequence features cannot) [11], homology is generally ruled out, whereas each (optimal) alignment corresponds to a potentially valid homology relationship. Given an alignment, it is thus possible for us to score how confident we are in the homology of a pair of sequences [10, 12, 13].

With the existence of large-scale sequenced genome databases, along with predicted and experimentally verified mRNA transcripts and translated proteins, methods have been developed for searching for and ranking potential homologs by evaluating potential alignments, in practice always in a heuristic fashion. This is an extremely common operation when trying to understand what role the protein expressed from a given sequence plays in the organism it is a part of. Existing tools for pairwise alignment of sequences build on *dynamic programming* techniques, like the Smith-Waterman local alignment algorithm [14] or the Needleman-Wunsch global alignment algorithm [15]. Subsequent developments such as FASTA [16, 17] or BLAST [18, 19] are attempts at heuristics to make similar approaches applicable to large-scale database searches. Alignment of multiple sequences could theoretically be performed using dynamic programming in a higher-dimensional space, but this is not feasible for practical applications due to time and memory constraints. As such, methods like Clustal [20, 21], Kalign [22, 23], Mafft [24-26] and Muscle [27, 28] all employ heuristics to merge multiple pair-wise alignments into a multiple sequence alignment.

More complex homology search and alignment reconstruction methods are available through the use of sequence profiles. These are based on the fact that most homologous sequence pairs are in fact part of larger homologous sequence families, and also on the fact that the nucleotide sequences are not random. Instead they are shaped by structural and chemical constraints to perform a biological function when translated into proteins. As such, a family will exhibit different degrees of variation at different sites in the sequence [29, 30]. Thus, considering known members of the family allows treating similarity or difference at a site as more or less important from the perspective of these constraints, as inferred from the relative conservation within the family at that site. Methods based on these facts – sequence

profiles, position-specific scoring matrices (PSSMs) [31], PSI-BLAST [19, 32], CS-BLAST [33] and the Hidden Markov Models [34-37] all allow detection of much more divergent homologs.

It should be noted here that the scores resulting from an alignment construction program applied to a pair do not necessarily correspond to the evolutionary distance between them in terms of time, although the two measures often seem to be correlated [38]. Many bioinformatics applications, including several described in this thesis [Paper IV, 39-43], do use measures of confidence in a homology inference, such as BLAST bit scores [12, 18, 44, 45], in order to rank homologs by order of distance. However, this should be considered a heuristic approach taken for ease of implementation, and it is an approach that the bioinformatics community might want to move away from [38, 46].

2.1.3 Low-complexity regions as an error source

While statistically significant sequence similarity between proteins generally is a consequence of their homology, there are factors that may cause non-homologous proteins to be unexpectedly similar in sequence. These include a shared but otherwise uncommon amino acid bias, but also the presence of internal repeats found in several unrelated genes. I refer to these as *low-complexity regions* [47, 48].

Approaches have been suggested that detect such sequence regions and mask them from subsequent analysis [47, 48]. Other suggested approaches make specialized sequence comparisons where the particular amino acid distributions are considered explicitly [32, 49-52], limiting inferences to using only the positional information found in a sequence.

2.1.4 Phylogenetics

For multiple homologous objects, such as a family of gene sequences (or, at a higher level, a group of species), their actual historical relationships will form a hierarchy: Species ‘A’ branched off from its ancestor, and its sibling branch subsequently split into ‘B’ and ‘C’. The latter two are more closely related to each other than to ‘A’, and this series of relationships matches a tree structure, with branches corresponding to periods of time separating events where new objects rise from the old. These trees are *phylogenies*, and the science or art of inferring them is called *phylogenetics*. Leaf nodes are called taxa or *OTUs* (*Operational Taxonomic Units*). Each phylogeny

corresponds to a hypothesis of (hierarchical) homology among a set of OTUs, which are typically but not always genes or species.

As historical events cannot be observed directly, we can only observe the effects they have on presently available observable objects, and infer the most credible history from there. We may reconstruct the phylogeny of genes or organisms with varying degrees of precision, from observations such as the molecular structures of genes, proteins and genomes, phenotypic characteristics of organisms, and the presence or absence of fossils or extant species. In this, we must also rely on inference criteria such as *maximum parsimony* [53-55] or *maximum likelihood* [56], both variations on Occam's idea of minimal assumptions.

Historically, animal and plant taxonomists have reconstructed the genealogies of entire species based on externally visible phenotypic traits as well as on fossilized organisms that could be dated using various methods [9]. These histories are very far-reaching but often not perfectly resolved, involving phylogenies that are multifurcating rather than bifurcating [57], and without well-defined branch lengths. Subsequently, the discovery of DNA and techniques for the detailed analysis of individual gene or protein sequences, or of protein structures, allowed molecular phylogenetics to complement these results, partly validating them and partly revising them. The last two decades, however, has seen sequence analysis techniques make a remarkable shift forward, allowing analysis of not only entire genomes of organisms, but entire genomes of multiple organisms simultaneously. As such, methods for analyzing the history of organisms based on their entire genetic content [58, Paper IV], or from multiple genes [59-65], in relation to a wide range of close and distant relatives, are becoming available.

2.1.5 Disagreements between trees

A central problem in bridging species and gene phylogeny lies in the possibility that they can disagree. A pair of genes may be *xenologs* [66], meaning that while they themselves are homologous, one or both has moved between species, through any of a variety of events [67], such as through transfer of bacterial plasmids [68, 69], through viral infection [68, 69], or from endosymbionts into nuclear genomes [70], so that their host organisms need not share the same historical relationships as these genes do. This type of *horizontal gene transfer* (*HGT*) is common in single-cell organisms of all stripes [67, 68, 71, 72], and have prompted many to ask whether it even makes sense to talk of a true, tree-like genome- or species-level phylogeny for prokaryotes [73], though others have argued differently [74, 75]. Similarly, *duplication* of genes, which is very common and a core concept in

this thesis, possibly followed by lineage-specific gene loss, also frequently gives rise to situations where gene and species phylogeny appear to disagree.

Analogous to the situation of gene versus species phylogeny, genes may experience recombination events of various types, causing subsequences to be gained, lost, duplicated or shuffled. Genes may be split or merged [76]. Recombinant subsequences may or may not correspond to relevant gene or protein features such as introns, exons or domains [77-80]. As a result, while each individual sequence region that has not been broken up by recombination in any of the organisms considered can be said to have a well-defined, tree-like phylogeny corresponding to its evolutionary history, the genes as a whole, like the mosaic genomes previously mentioned, sometimes cannot [81, 82].

On a higher level, population genetics may face analogous problems. Defining a species is in fact non-trivial. A population of individual organisms might be considered to have a treelike history, which reflects the history of the set of species that arose when the population was divided by migrations, or by mutations making interbreeding between subgroups no longer possible. But this overall history, which is the species phylogeny, may only exist as an abstraction of the statistical behaviour of the individual organism histories, which in turn contain potentially conflicting gene histories, and they contain potentially conflicting domain histories.

Realistically, it is often necessary to work at the level of these abstractions when attempting phylogenetic reconstruction, i.e., we must construct gene phylogenies that unite the information from individual domains, organism phylogenies based on subsets of all available genes, and species phylogenies from those few or singular individuals we actually have sequence data for. This adds uncertainty to all such analyses, but by bearing these limitations in mind, we often are still able to make good use of their results [9].

Reconciling component phylogenies into a single whole can be done in several ways. In a situation where we would have access to the genomes of several individuals from the same species, we could use their consensus or average representation, or alternatively a profile describing the ranges within which they vary, to integrate the information the individuals carry for the species as a whole. This can be seen as integration along the dimension of population, but to my knowledge it has not yet been attempted to any great degree, mainly due to the scarcity of multiple genome sequences for the same species.

The other dimension, integrating information from multiple genes within the same genome, is better studied [59-65]. The term *phylogenomics* has been

suggested for approaches that integrate phylogenetic signals from multiple gene sequences [60], and involve both methods for crafting virtual "metagenes" from concatenated gene sequences [64, 83], as well as "supertree" methods for finding species phylogenies that are optimally compatible with multiple gene phylogenies [59-63, 65, 75, 83]. Other methods involve building species phylogenies from information on the gene or domain content of organisms [57, Paper IV], or reconciling sets of gene phylogenies with a hypothetical species phylogeny [59, 84, 85].

2.1.6 Orthology and gene duplication

The current terms *orthology* and *paralogy* were minted by Walter Fitch in a seminal article [10], though similar concepts were described using different terms by Zuckerkandl & Pauling [86] a few years earlier. Basically, orthology versus paralogy are properties that a homology relationship can possess in relation to a species tree. Two homologous sequences found in different species are orthologous if they descend from the same sequence in the *cenancestor*, the last common ancestral species of the species where they are presently found. Alternately put, the evolutionary event that gave rise to the two species – a speciation event – was also the event that separated the two sequence lineages.

In contrast, if the homologs exist within the same species, or are descended from different gene duplicates that arose before the cenancestor, they are instead paralogs. In the context of a particular species comparison, we can separate these two cases: *in-paralogs* are same-species paralogs that diverged through duplication events after the divergence of the species lineages under consideration [87, 88]. *Out-paralogs* are same-species or cross-species paralogs that stem from different ancestral paralogs in the cenancestor.

Can we talk of orthology when considering more than one species? An ortholog group defined relative to a particular lineage of organisms can be thought of as all the parts in those organisms that descended from a single part in their last common ancestor. However, in this case, it is not guaranteed that every cross-species homology relationship in this set will be an orthologous relationship. This is because orthology depends on whether or not two genes descend from a single gene or not in the cenancestor to the two species where they are found. With multiple species included in an orthology group, different pairs of species will have different cenancestors, allowing pairs of genes from the same multispecies group to be either orthologous or paralogous. Ultimately, the definition can be safely applied only to pairs of genes [89].

There is considerable confusion in the literature concerning the terminology for orthologs, however. Due to observations that many orthologs retain the same functions – that is to say, that for pairs of genes that are orthologous, there are many cases when the products of both genes in the pair can be shown to perform the same function – some authors have used orthology as a synonym for functional equivalence (see [90]), which does not directly follow from the original definition. Furthermore, when a cenancestral gene has duplicated in one lineage but not the other, all the descendent genes in the first lineage will independently be orthologs to the gene in the second lineage, whereas they will be inparalogs of each other. In some cases, authors have considered only one of these genes an "ortholog" [91], which misses the fact that orthology is a property of pairs of organism parts rather than a property of the individual parts themselves. Some authors have claimed that the orthology relationship has the logical property of transitivity [92]. This, however, can be clearly shown not to be the case [66, 89].

2.1.7 Inferring orthology

Like homology, orthology is a property we can only infer, not observe. Ultimately, the "true" answer would be gained from comparing a part (gene) phylogeny with a species phylogeny. If the subtrees defined by the two genes and by the two species are rooted at the same point, they are orthologs. More generally, methods for assigning orthology relationships given trees, such as tree reconciliation, exist and are well described [59, 84, 85]. A well-defined species tree must be available, however, which may be problematic in several cases [93].

More notably, for a long time, large-scale phylogenetics-based orthology reconstruction has been impractical or intractable for computational reasons. Though some recent developments suggest this may be changing [94], this state of things has nevertheless prompted a focus on development of heuristic methods for inferring orthology, generally from sequence distance networks of one type or another. As such, these methods have been collectively referred to as *graph-based* rather than *tree-based* methods [2, 95, 96]. The edges of such networks should ideally correspond to evolutionary distances, but many methods have instead used homology confidence measures such as BLAST bit scores [12] as a proxy.

The simplest graph-based method is the RBH or Reciprocal Best Hit [46, 97]. Pairs of sequences are considered orthologous if they are both each other's closest neighbour. In practice, this means that they are each other's top hits in all versus all genome-wide sequence comparisons, generally using

tools such as BLAST [18], but sometimes using dynamic programming, i.e. Smith-Waterman [14] or Needleman-Wunsch [15] alignment. The RSD or Reciprocal Smallest Distance method replaces alignment scores by maximum-likelihood estimates of branch length between the sequences [46]. Ranking next in complexity, a series of methods follow that add in additional inparalogs to the resulting orthology cluster, or that use triangles of reciprocal best hits between three species at a time to build multi-species ortholog groups. More complex clustering methods can also be used. There are also an increasing number of phylogenetics-based orthology resources. Some of the most commonly used or otherwise notable tools are briefly reviewed below.

Aside from sequence similarity- or phylogeny-based orthology reconstruction methods, there is a growing repertoire of *context-based* methods for orthology inference, i.e., using the homology or orthology of neighbouring genes as evidence for orthology [98-100]. This conservation of neighbourhood, termed *synteny*, is justified in that the original gene involved in a gene duplication event will remain where it was, making for segments of orthologous genes conserved across species, at least over sufficiently short evolutionary distances. It will, however, be unable to detect all the orthologous relationships in the case of one-to-many or many-to-many orthology groups.

2.1.8 Overview of some orthology inference resources

A comprehensive online repository of ortholog databases was recently established by the Quest for Orthologs initiative, and is located at http://questfororthologs.org/orthology_databases. The following sections briefly describe some of these resources, as well as a few others not listed there.

2.1.8.1 COGs/KOGs

Possibly the best-known orthology inference resource, the Clusters of Orthologous Groups (COGs) [39-41] are constructed by linking together genes that are reciprocal best BLAST hits, following a step where obvious inparalogs are merged. Where such reciprocal best hits can be found between at least three species, an ortholog group is inferred, and successively added to by including sequences that are likewise reciprocal best hits to existing group members. Finally, the resulting groups are inspected manually. While the original version contained mainly

prokaryotes, with yeast as the sole eukaryote, the euKaryotic Ortholog Groups (KOGs) [41] version included additional eukaryote species.

2.1.8.2 eggNOG

The 'evolutionary genealogy of genes: Non-supervised Orthologous Groups' (eggNOG) [82, 101] is in many respects similar to COGs/KOGs, but with vastly higher coverage as well as extensive integration of functional information. Like COGs, it is built based on triangles of reciprocal best hits, in this case using Smith-Waterman rather than BLAST scores as a distance measure. Inparalogs are initially merged, as are very similar genes in closely related species. Several different clusterings are performed at different taxonomic levels, and an additional filtering step breaks up clusters artificially joined by genes that arose by two separate genes fusing into one.

2.1.8.3 TOGA/EGO

The TIGR Orthologous Gene Alignments (TOGA, at present called EGO) [102], works similar to COGs by clustering eukaryotic genes into multiple-species ortholog groups on the grounds of reciprocal best BLAST hits linking at least three species together.

2.1.8.4 EnsemblCompara GeneTrees

The Ensembl sequence database clusters genes into families, constructs trees from them and infers orthology and paralogy by reconciliation with a species tree [103, 104].

2.1.8.5 InParanoid/MultiParanoid

InParanoid [105-107, Paper IV], is a graph-based method for inferring orthology and paralogy relationships between all members of two complete proteomes. As such, it particularly focuses on correctly including species-specific inparalogs while excluding outparalogs predating the speciation that a particular comparison of two proteomes define. This is done at a price, as the framework is then limited to comparing two species at a time. An attempt to extend InParanoid was made to allow inference of hierarchical orthology groups from multiple closely related species in the form of MultiParanoid [43], but it has not been updated since publication.

2.1.8.6 Homologene

The NCBI makes available the Homologene database [108] which consists of gene families for which phylogenetic trees have been reconstructed. While not explicitly presented as an orthology resource, it has been treated as such in some contexts [109].

2.1.8.7 KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) [110-115] assigns genes in a genome to orthology groups through automatic and manual inspection of sequence similarity, presence or absence of genes found together in other organisms, and by chromosomal proximity. Its main feature is its strong focus on functional annotation, assigning as many genes as possible to specific roles in biochemical pathways.

2.1.8.8 MetaPhOrs

The recently introduced MetaPhOrs server [116] is effectively a metaserver for tree reconciliation-derived orthology, drawing on as many gene trees as possible and deriving consistency scores from the degree to which an orthology or paralogy inference is supported by multiple trees.

2.1.8.9 OMA

The Orthologous MAtrix (OMA) [117, 118] resource is a graph-based tool for inference of groups of 1-1 orthologs (sometimes termed *super-orthologs*) [119]. It uses reciprocal best hits using Smith-Waterman alignment, which are then filtered by searching for relationships to genes in other genomes that would contradict the inferred orthology relationships [120].

2.1.8.10 OrthoMCL

OrthoMCL [42] is a graph-based method similar to InParanoid, in that it is based on sequence similarity between genes. Same-species genes more similar to each other than any gene in another species are clustered together as inparalogs. Following this step, the Markov Cluster algorithm (MCL) [121] is used to link inparalog groups together into ortholog groups. This heuristic approach can be applied both to the pairwise species comparison

case or to multiple species at once. OrthoMCL-DB [122], a database of orthology inferences using this method, is also available.

2.1.8.11 PhIGs

The PhIGs [123] system is fundamentally a tree reconciliation-based orthology inference tool. However, it clusters gene sequences into families by help of a species tree before family tree reconstruction and reconciliation, assigning each resulting ortholog group to the taxonomic level where it is first seen. It also contains Hidden Markov Models used to rapidly assign query sequences to a PhIG.

2.1.8.12 PHOG

The PHOG orthology resource [119] is in some sense a hybrid of a tree reconciliation method and a graph-based method. To avoid computational costs associated with tree reconciliation, pairs of genes within the same family trees are classified as orthologs or paralogs based on the distance between them along connecting tree branches.

2.1.8.13 PhyOp

Goodstadt & Ponting [124] presented a high-precision reconstruction of orthologs and paralogs between dog and human based on sequence clustering followed by phylogenetic tree reconstruction and gene-species tree reconciliation. It is noteworthy in that the analysis incorporates multiple splice forms for each gene, thus potentially avoiding artefacts resulting from unfortunate choices of splice form representatives for each gene.

2.1.8.14 Roundup

Roundup [125] is similar to InParanoid in that it infers orthology and paralogy relationships between the proteins of two genomes at a time, but differs from it by being based on reciprocal smallest distances [46] rather than using BLAST bit scores as a distance measure. It is available in a number of builds using different sequence inclusion thresholds.

2.1.8.15 TreeFam

The TreeFam database contains family trees for genes mainly from animals [126] but later also updated to include some fungi and plants [127]. Sequences are based on families from PhIGs, but extended through BLAST searches and Hidden Markov Model searches within the included genomes. Based on the tree reconciliation algorithm of Zmasek & Eddy [128], orthology and paralogy relationships are then inferred.

2.1.9 Accuracy of ortholog inference

How well do the various orthology reconstruction methods work? Given the definition, a true benchmark of orthology can only be done where gene and species phylogenies are both known, so that the true and inferred relationships can be compared. This can be done directly, by considering agreement between the sets of orthologous and non-orthologous pairs stemming from an inference and from the known relationships; such analyses is performed for a small number of manually curated orthology relationships in the benchmark studies by Hulsen and co-workers [90] and by Altenhoff & Dessimoz [109], and was also done during the initial testing of InParanoid [105]. However, the small datasets limit the applicability of these results. An indirect approach instead samples groups of genes wherein each is predicted to be orthologous to all other genes in the group, with one gene taken from each species. A phylogenetic tree is built from these genes, and compared to the phylogeny of the species from which they are sampled [109]. In this manner, large-scale phylogenetic evaluation of orthology inferences becomes possible. Very recently, a phylogenetic benchmark consisting of 70 manually curated protein families in animals was applied to evaluate a selection of orthology resources, as well as to try to determine the influence of various error sources. Errors in genome annotations stood out as the major factor limiting the resources under comparison, as well as problems where domain shuffling had obfuscated orthology relationships [96].

Since it is known that relative chromosomal position, or gene order, is often conserved between sets of orthologous genes (the phenomenon of *synteny*), the degree of synteny exhibited under different methods for orthology inference may provide some guidance. While not every true orthology relationship will be reflected in synteny, it can nevertheless be useful for the relative comparison of methods, though with the risk of bias. This was done as part of the evaluations of Altenhoff & Dessimoz [109] and Hulsen and co-workers [90], as well as when optimizing algorithms for the

EnsemblCompara GeneTrees database [104], and for evaluation of the PhyOp resource [124].

Most comparative evaluation of orthology inference methods, however, have not actually focused on testing for actual orthology as defined by Fitch, but have instead measured conservation of various functional properties among orthologous pairs. Hulsen and co-workers [90] measured similarity of tissue expression/co-expression profiles and interactions for orthologs inferred through various methods. Altenhoff & Dessimoz [109] similarly compared conservation of Gene Ontology terms and Enzyme Classification (EC) numbers, as well as expression profile. While these tests may be of value for researchers in order to determine how similar in these respects they should expect orthologous pairs inferred by different methods to be, it reveals nothing regarding the reliability of the orthology inferences themselves. Moreover, neither study contrasts the average conservation of orthologs to that of paralogs, which, if performed, would have allowed evaluation of the relationship of these properties to the orthology phenomenon in itself, i.e. the orthology conjecture [3]. Zmasek & Eddy [129] perform bootstrap tests on orthology inferences, an approach also used in InParanoid [105]. However, this merely tests the extent to which the results are robust to input data noise, not the extent to which they actually capture the true evolutionary relationships.

On the whole, different orthology resources have different species coverage and may provide different utility depending on the sensitivity and precision needs of a particular application [90, 130]. The only large-scale evaluation of agreement with phylogeny, by Altenhoff & Dessimoz [109], evaluates only one tree reconciliation-based orthology inference method (EnsemblCompara), which surprisingly enough did not strongly outperform competing heuristic methods in this benchmark, which might otherwise have been expected. One remaining obstacle to comparative evaluation on a large scale of orthology inference methods is the technical difficulty of matching up the different representations used. Common standards for sequence data (SeqXML) and orthology inferences (OrthoXML) have been suggested for circumventing this problem [131], and may soon enable more comprehensive evaluations of competing methods.

2.2 Protein domains

2.2.1 Protein modularity and domain architecture

As techniques for analysis of protein structure developed, it became apparent that some structural forms appeared in multiple proteins, which were otherwise structurally different [132, 133]. Such recurring elements, termed *domains*, came to be seen as building blocks for protein structure on a level higher than secondary structure elements [134, 135], and were shown to often be independently folding (with the term '*fold*' often used in the same sense as 'domain') [133, 136]. Moreover, the sequences corresponding to these protein domains can be aligned, and from the resulting alignments, powerful methods for sequence profile searches can be used to find additional sequences belonging to such *domain families*. Likewise, novel gene sequences can be assigned to protein domain families from the library of already-known domains, and unassigned regions from many proteins can be subjected to sequence clustering methods that aim to discover novel domain families [132, 134, 137, 138].

From a theoretical perspective, the existence of structurally and potentially functionally well-defined subsequences may provide a vital piece of the puzzle regarding how protein complexity can evolve [133, 139]. Recombination of domain sequences – either through exon shuffling [139-142] or through mechanisms such as gene fusion or fission [76, 143, 144] – might allow a relatively small number of mutational steps to result in protein variants with novel functional specificities, resulting from the combination of the properties of their constituent domains. It also allows refinement of a domain in one context to be reused in another protein context, meaning that not every protein family must evolve from scratch.

Categorizing proteins into families is a vast project to which much effort has been dedicated. Families at the protein level may be defined by the presence of a domain, or there may be distinct combinations of domains that characterize a higher-level family [145, 146]. As sequence databases grow more complete, better and better surveys can be made of the diversity of proteins, both with regards to domain families and multidomain combinations. A picture also gradually emerges of the distribution of these families across different lineages, enabling analysis of when particular domains or combinations first emerged [142, 147] and subsequently shedding light on which genetic innovations played a part in the rise of particular classes of organisms [139, 148]. A specific subproblem is

estimation of the domain content of the *LUCA* or *Last Universal Common Ancestor*, which can be addressed by methods such as maximally parsimonious reconstruction of its domain repertoire [149, 150].

2.2.2 Overview of some protein domain databases

As more and more protein sequences and structures were identified through structural genomics projects and genome projects, domain families were identified either through manual curation or through clustering approaches. This process has been carried out independently by several different groups using different datasets and methods, and as a result, multiple redundant domain classification systems exist. While conclusions drawn from one are often valid relative to another, the systems are not directly compatible. This section lists some of the most widely used systems.

2.2.2.1 CATH/Gene3D

The CATH database [151-156] is a hierarchical classification system of structural domains. Each letter ('C', 'A', 'T', 'H') corresponds to a hierarchical level, with *Homologous superfamilies* belonging to *Topologies*, which belong to *Architectures*, which belong in turn to one of the four *Classes*. It is constructed from the protein structures available in the Protein Data Bank (PDB) [157]. Domains sharing the H or T level can be assumed to be homologous, whereas this cannot be guaranteed for the higher levels (Orengo, personal communication). It is built using both computational structure comparisons and manual curation. While only proteins with experimentally determined structures are thus part of CATH, it has been used to build sequence family Hidden Markov Models, which are used to search sequence databases and assign CATH classifications to proteins without experimentally determined structures. These models and assignments make up the affiliated Gene3D database [154, 158-161].

2.2.2.2 CDD

The Conserved Domain Database (CDD) [162, 163] at NCBI is a domain metadatabase in that it imports domain models from many other databases, as well as unique models based on 3D structure data. These models are used to assign domains to sequences in the NCBI databases. Unlike the other databases listed here which typically use Hidden Markov Models, CDD uses the RPS-Blast algorithm [164] for this purpose.

2.2.2.3 Interpro

Interpro [165, 166] is a domain metadatabase, which integrates domain assignments from many different schemas (including those listed here) for a large set of protein sequences. It also contains functional predictions made using this domain information.

2.2.2.4 Pfam

The Pfam database [167-174] is analogous to the Gene3D and SUPERFAMILY databases in that it is built by training Hidden Markov Models (HMMs) for known domain families. These are then used to search sequence archives in order to assign protein sequence regions to these families. However, instead of structure-defined families, Pfam is built from manually curated *seed alignments* either based in literature or from automated clusterings of sequences not currently assigned to any domain family, the Pfam-B database. Various methods have been used to perform this clustering [132, 134, 137, 138]. The Pfam database is not hierarchical as such, but later versions include a higher level of organization in that homologous domain families are gradually grouped together into *Clans* [172].

2.2.2.5 SCOP/SUPERFAMILY

SCOP [175, 176] is highly similar to CATH in that it defines protein domains from structure, though it relies more on manual assignment and curation than on automated structure comparisons. It, too, has four hierarchical levels, though these do not correspond exactly to the CATH levels [177]. Similarly, SUPERFAMILY [178-182] is analogous to Gene3D, or to Pfam, with the SCOP families serving as seed alignments.

2.2.2.6 SMART

The SMART database [183-189] is similar to Pfam in that it is populated using Hidden Markov Models from seed alignments. Unlike other databases in this section, it does not aim to be exhaustive but rather specializes in signalling and regulatory domains, as well as in integrating associated functional information.

2.2.3 Mechanisms of domain architecture evolution

What are the mechanisms that allow the domain architectures of proteins to change between successive generations, and what kind of changes do these mechanisms actually cause? Notably, there are two aspects to the evolution of domain architectures. The first is which mutations actually take place. The second is whether a given mutation is retained in the population and perhaps brought to fixation, or purged from it through reduced fitness or random chance [190].

As for the mutations, there are a variety of ways in which the sequences encoding proteins in the genomes can change, either in place or through introduction elsewhere into the genome of a modified duplicate. Homologous recombination [191, 192] is a DNA repair mechanism that replaces material in one region with that from a homologous region. Non-homologous, or illegitimate, recombination [193] may exchange material between entirely different genes. Mobile elements such as retrotransposons [194, 195] or DNA transposons [195, 196] provide further mechanisms for larger-scale genetic changes. Point mutations may add or remove start or stop codons, splice sites, or other sequence markers that affect which DNA regions will end up in the translated proteins [143]. This may shorten, lengthen, split or fuse genes. Processed transcripts may be inserted in the genome through retrotranscription, generally as inactive pseudogenes but not always [195]. In combination with atypical splicing, retrointegration can even involve chimeric transcripts spliced from exons of several genes [197]. Segmental duplication of genome regions may duplicate all or part of genes, creating architecture copies or novel architectures [197]. A novel coding sequence may arise through the process of exonization [141, 142, 197]. An uncommon but notable phenomenon is that of circular permutation of a domain architecture (e.g. ABCD -> DABC), the mechanisms of which have been explored by Weiner & Bornberg-Bauer [198] and by Vogel & Morea [199].

There is not room within the scope of this thesis to give a detailed rundown of the population genetics behind duplicate or mutant retention and/or fixation in a population, but the core premise is this: a modified trait may persist either through *genetic drift* (i.e. by chance), or through positive selection. The former is more likely the smaller the effective population size is, allowing for population bottleneck phenomena.

2.2.4 Reconstructing domain architecture evolution

How can we chart the processes that change domain architectures over evolutionary time? Fundamentally, this is done by inferring changes from present-day architectures, usually through maximum parsimony assumptions where the scenario involving the smallest number of changes is concluded. For each multidomain architecture, Ekman and co-workers [200] identified the most similar architecture found elsewhere, and treated the differences between them as corresponding to an observed change.

Several studies [150, 201] have considered the repertoire of domains, architectures or domain combinations as a set of binary characters defined for each species, and reconstructed the most parsimonious ancestral assignments of these characters. The more restrictive *Dollo parsimony criterion* [202] has also been used [201, 203], which requires that any traits, such as the presence of a domain, can only be gained once but lost multiple times in parallel, based on the assumption that specific gain should be much less likely than loss of a domain already present. Other studies [Paper I, 142] have used explicit gene trees and assigned ancestral domain architectures to internal nodes so as to minimize the number of domain gain and loss events along each tree.

All of these approaches suffer from the presence of many largely *ad hoc* assumptions that risk biasing the results depending on the particular framework. While parsimony has sometimes been described as being assumption-neutral, it corresponds in effect to a scoring scheme where all changes are assigned the same score. However, other studies have shown that gene fission and fusion events [76, 201] and domain gain and loss events [150] occur with different frequencies. In response to this, Itoh and co-workers [150] scored gain events as three times more costly than loss events, but found that their results were robust to changes in this parameter. Architecture change events also occur with different frequencies at the termini of proteins than in central positions in an architecture. This is likely because certain architecture changing events – fusions, fissions, and insertion or deletion of start or stop codons – always affects the terminal positions [142, 143, 204, 205].

However, parsimonious reconstruction of ancestral properties along a tree also does not consider the time that passes along a tree branch, whereas in reality we would expect that changes should be much more likely along very long branches than very short branches. The relative relationships between branch lengths can in fact be thought of as a property of the taxon sampling more than anything else. Including many closely related taxa will make for

either very short branches or multifurcating rather than bifurcating trees, if unreliably short branches are collapsed. Revisions to the original maximum parsimony algorithm [150] that do not assume strict bifurcations have been suggested, and may alleviate this problem somewhat, though issues resulting from unequal branch lenghts will still remain.

2.2.5 Evolution of domain family sizes

The most accessible way of understanding the evolution of domain architectures has been charting the distributions of domains, domain combinations and domain architectures across individual genomes, taxonomic groups, and kingdoms of life, as this does not require direct reconstruction of phylogenies. Early on, as bacterial genomes began to be sequenced, the distribution of sets of paralogous genes – i.e. gene families, overlapping with single domains or multi-domain architectures – within genomes was studied. It was found that a particular family size distribution occurred again and again; the *power law* [206-208].

The power law distribution, which is a specific case of the Generalized Pareto Distribution (GPD) [209], corresponds to "the dominance of the population by a selected few" [208], i.e., that a majority of families are sparsely populated or singletons. On the other end of the scale, this distribution has a "heavy tail" [210], such that a few very large families exist and a majority of instances in fact come from a minority of families. There is a vast body of literature on power law phenomena in a variety of contexts, including computer network architecture [211], word usage in languages [212], wealth distributions [213] and scientific citations [213]. In biology, it has also been demonstrated for some genomic features, such as distribution of pseudogenes or short DNA sequences [208]. Furthermore, power law distributions are seen in the node degrees of protein-protein interaction networks [214] and metabolic networks [215, 216].

Power law distributions have been linked to concepts such as self-similarity [217], a property shared with fractals, as well as scale-freeness and small-world network properties [218], but the use of these terms is not always stringent [217]. It has been further noted that family size distributions may be even better modelled using the general GPD [209, 219, 220], with additional parameters varying between kingdoms of life, but as those distributions nevertheless yield asymptotic power law behaviour, this point is mostly academical. It is also the case that several different network architectures can display the same power law degree distributions, even while differing with respect to such properties as the propensity for "network

hubs" to be connected [217], impacting the extent to which the network exhibits modular behaviour.

In general, situations where power laws are observed are such that the most likely expected alternative would be an exponential decay or binomial distribution, which are different from the power law mainly by the absence of those few very large families in the heavy tail. Classic random graphs such as Erdös-Renyi networks [221] will have a binomial degree distribution, prompting development of alternative random graph models for biological networks, such as *preferential attachment* [222], where already well-connected nodes are proportionally more likely to acquire additional connections. The prevalence of power law behaviour in domain family distributions [206-208], domain combinations (i.e. supra-domains) and domain architectures [223-225], thus provide a basis for conclusions on how the domain architecture repertoire of proteomes evolve. In a recent review work [205], we further validated these power law distributions based on the state of the Pfam database in 2010, and found that the same trends remained clear.

Huynen & van Nimwegen [206] interpreted the power law distribution of family sizes as consistent with a model of random gene duplication, but with family- and organism lineage-specific probabilities of duplication (or, more likely, of duplicate retention). This would follow if duplicates within certain functional categories were more useful than others in a particular organism, depending on its functional requirements. Yanai and co-workers [226] disputed this, claiming that uniform duplication probabilities across families provided a sufficiently good fit to the observed data.

Later work suggested more complex evolutionary models such as *birth-death* [219] or *birth-death-innovation* [207, 209], and modelled domain family size distributions accordingly. Karev and co-workers [209], like Huynen & van Nimwegen [206], concluded that different domain families gained or lost members at different rates, based on the selective advantage of having more or fewer members of those families available. They also concluded that domain gain and loss rates must be asymptotically equal for simulations to match observed distributions, which would follow from a punctuated equilibrium type model where family size evolution may shift, though rarely, between different evolutionary submodels but otherwise be relatively stable. These shifts would then correspond to shifts in organismal complexity. The number of genes involved in particular functional categories within genomes also follow power law distributions [227, 228], with power law coefficients differing between functional categories, which is consistent with the selection-based models discussed above, particularly as

these functional categories are likely to match up with domain family categories to some extent.

As such, there seems to be some support for a model in which the sizes of domain families are controlled mainly by duplication and loss of whole genes (rather than by domain-level mutations), processes which in turn may vary in frequency depending on the utility of particular families for particular niches. The fact that the same distributions are seen both for single domains, combinations of domains and entire architectures further imply that evolutionary events affecting whole genes play a larger part in shaping protein repertoires than domain architecture changing mutations do, though the latter also has an impact.

2.2.6 Evolution of domain combinations

Similar mathematics as apply to the size distribution of domain families also applies to the number of different combinations that domains are found in. If two domains are present together in a protein, the corresponding nodes are linked in a *domain co-occurrence network*. Similarly, *domain neighbour networks* link domains for which there is at least one protein where those domains are found next to each other. Mutations causing domains to recombine, then, leads to the creation of new edges in such networks. Przytycka and co-workers [203], Itoh and co-workers [150] as well as Kummerfeld & Teichmann [229] have all analyzed such networks. Apic and co-workers [223] found that the degree distribution matches a power law, and Wuchty [218] demonstrated that it can also be fitted to a general GPD, results that remained consistent in our recent validation [205].

The existence of a power law for domain combinations makes for the existence of certain '*promiscuous*' [230], '*mobile*' [184, 231] or '*versatile*' [231-234] domains, which have very many different combination partners, most of which in turn are found only in that combination or in a small number of combinations. Several different metrics have been suggested for measuring the "intrinsic" versatility of a domain family, in the sense of how likely members of the domain family are to enter into novel combinations as they grow more abundant through duplications [233, 234]. Attempts have been made at investigating if particular functional categories are overrepresented among promiscuous domains, but no statistically significant trends have been demonstrated [231-234].

The most important question for the purpose of this work, however, is whether there is selection for or against particular domain combinations or if their evolution is primary characterized by random drift. Apic and co-

workers [224] suggested a random model for domain architectures sampled from a domain repertoire (the 'bag of domains' model). They concluded that far fewer combinations are actually observed than would be expected under this model. This would follow from gene duplication being the dominant mechanism for formation of new proteins, regardless of whether selection is exerted on particular combinations or not. This is also consistent with the relatively small numbers of convergently evolved architectures [203, 235, Paper I], and with most domain combinations only seen in one of the two possible N- to C-terminal orientations, despite there being few structural constraints against this [236].

Kummerfeld & Teichmann [229] extended the co-occurrence network representation to have directed edges (i.e. 'occurs to the left of' and 'occurs to the right of' represented separately). By comparing the observed network to a random model, they could conclude that while supra-domains (i.e. domain combinations) were found in more than one N- to C-terminal orientation only very rarely, it was still more common than expected from this network-informed random model. Similarly, the network exhibited more prominent clustering behaviour than expected from the random model. These findings may reflect selective pressure in the form of positive selection for certain domain combinations, but more work will be required to fully determine the truth.

2.2.7 Monophyly versus polyphyly of domain architectures

Yet another approach to address the question of the direct functional importance of particular domain architectures (the strong form of the domain grammar hypothesis) is to determine how frequently multi-domain architectures evolve convergently. If it can be concluded that selective pressure operates on the formation of novel domain architectures, it would imply that particular combinations were better than others at implementing particular functions. However, concluding that selection occurs may also be difficult in this case, due to the absence of an obvious neutral null model to compare against. Regardless, it is clear that most multi-domain architectures are *monophyletic* rather than *polyphyletic* traits within gene trees; that is, they have arisen only once [Paper I, 203, 232, 235].

The assumptions underlying the use of Dollo parsimony [142, 202, 203] in attempts at reconstructing ancestral domain architectures also reflect this insight. However, while uncommon, convergent evolution of domain architectures is not unheard of [Paper I, 203, 235]. An interesting result was obtained by Przytycka and co-workers [203] in that, based on graph theoretical constraints and Dollo parsimony, it is possible to prove

stringently, given a set of extant protein architectures, whether a given multi-domain architecture can have evolved monophyletically or not. As such, the results of such an analysis forms the most conservative possible estimate for the prevalence of convergent evolution of domain architectures, while those reported in Paper I may form a most permissive upper bound.

2.2.8 Comparing domain architectures

It is not trivial to compare two proteins with respect to their domain architectures in a systematic manner. More specifically, there are several ways to consider the difference, and depending on the purpose of the comparison, different methods may be most useful.

Some approaches simply consider the set of domains shared between both architectures, focusing on the presence or absence of particular domains without considering their exact arrangement [237]. It can be argued that this is, in fact, the chemically most relevant aspect. However, from a perspective of the evolution of the architectures, it is likewise important to note how they have changed. By aligning domain architectures, the differences between them can be explicitly considered, and it is possible to determine an edit distance [204], which is the number of domain-level changes of any type required to turn one architecture into another, as well as to classify the changes qualitatively.

The edit distance is a simplified model of reality, because different changes are more or less likely to occur, and have different impacts on overall protein structure and function. However, weighting particular changes has so far not been attempted in a comprehensive manner, likely because reliably determining their relative costs would require a thorough analysis of a very large benchmark dataset. The closest that studies have come this far, is by ignoring cases where only the number of consecutive small repeats have changed, such as by collapsing consecutive repeats into single “pseudo-domains” [Paper I, Paper V, Paper VI, 200, 201, 237]. The rationale for this is that such regions are known to be highly volatile in evolution. Another example is found in the ancestral architecture reconstruction analysis of Itoh and co-workers [150], where gain events are considered three times as costly as loss events.

Next, there is the fact that if two proteins differ by one domain, this is a much larger part of their wholes if they are two-domain proteins than if they are twenty-domain proteins. When comparing architectures from a structural

perspective, then, some normalization for number of domains may also be called for [81, 238].

It is possible to construct a comparison metric for a particular purpose. Notably, this was done by Song and co-workers [81] for the purpose of finding a weighted difference metric that could optimize identification of true homologs from a manually curated reference set. This imposes limits on interpretation of such differences, but is useful for applications that rely on domain architectures to find distant homologs [237, 239, 240]. Likewise, Lin and co-workers [238] optimized their similarity metric for recapturing shared KOG eukaryotic ortholog group membership.

2.3 Protein function

2.3.1 Classification of protein function

A great deal is known about what individual proteins do. Much of this knowledge, however, is available only as unstructured text in books or articles. This makes propagating the information to all those who could make use of it more difficult to do efficiently, and it also prevents stringent application of statistics, which is the formal tool the sciences must use for drawing any type of general conclusions from a collection of data. Moreover, it makes this knowledge inaccessible as a source of information for automated prediction of the function of uncharacterized proteins.

In its most general form, exactly defining the concept of protein function is non-trivial. Ultimately, it refers to the relevance of a protein for the biology of the organism. It has been argued [241] that the totality of a protein's interactions can describe this role. However, to fully capture function, these interactions must then be annotated both with respect to the conditions under which they occur and with respect to how they contribute to phenotype-level phenomena.

One trend has been to classify proteins into pathways, assigning them a role in a chain of catalysts or binders that carry out some metabolic or signalling function. There exists a variety of pathway resources, both for specific model organisms and more general. Another form of functional classification, particularly for enzymes, concerns describing a protein mechanistically, in the sense of which particular affinities it has and what reactions it might affect. However, the most complex schemes for

formalized descriptions of protein function are probably the *ontology* approaches [242]. These are arbitrarily extendable *controlled vocabularies* for wide-range functional annotation.

2.3.2 Overview of some protein function definition systems

2.3.2.1 COGs

While technically an orthology inference framework, some studies consider the COG ortholog group [39-41] to which a gene is assigned as a description of its function. This approach arguably is justified only if orthology can be assumed to imply functional conservation.

2.3.2.2 Enzyme Classification (EC)

The oldest system for functional classification of protein was designed by the Enzyme Commission to group enzymes according to activity. It is hierarchical, and represents each enzymatic function by four numbers, corresponding to increasingly specific description of mechanism and substrate [243]. Each enzyme usually is described only by a single EC number.

2.3.2.3 FunCat

The MIPS [244-250] FunCat [251] functional categories form a controlled vocabulary for protein function description. It is hierarchical, with a relatively large number of root terms divided into broad categories, each having successive, more specific, child terms. A protein is frequently annotated with multiple terms, and presence of child terms should be assumed to imply presence of their parent terms.

2.3.2.4 Gene Ontology (GO)

This system was designed to aid curators of model organism genome databases in annotating genes and gene products in a consistent manner, allowing for easier comparative analysis. Arguably the most ambitious of the systems described, the Gene Ontology [252-254] is similar to the FunCat in that it is a hierarchy of description terms connected through parent-child relationships. However, rather than a tree it forms a directed acyclic graph,

as each term may be the child of multiple parents. Each term is also assigned an *evidence code* representing its origin, which may in some sense be considered a kind of qualitative confidence measure. There are three main root terms, representing biological process participation, cellular localization and molecular function respectively. Most child terms imply all their ancestral terms, though there are some exceptions because of subtleties in the exact relationships represented by each link. The system can also contain qualifiers, enabling for example a protein to be explicitly flagged as not possessing a certain term, but this is still rarely used by annotators [255].

2.3.2.5 KEGG pathways

The Kyoto Encyclopedia of Genes and Genomes [110-115] has long compiled chemical pathways in living systems, listing both chemical species and the proteins controlling the reactions. Through orthology inferences (mainly based on reciprocal best hits from sequence similarity searches), proteins may be assigned to these pathways, producing species-specific pathway maps. As such, membership in a KEGG pathway is often used as a functional descriptor for proteins.

2.3.2.6 OMIM

The Online Mendelian Inheritance in Man (OMIM) [108] database lists heritable genetic conditions, mainly diseases, as well as human genes to which they have been mapped. As such, it can be used to describe disease involvement of genes, which can be considered an indirect descriptor of function – each such disease is fundamentally the result of the absence of a particular gene function.

2.3.2.7 UniProt keywords

The UniProt [256] database curators assign functional descriptions to proteins using a system of keywords from several controlled vocabularies.

2.3.3 Introns and alternative splicing

As more and more gene sequences became available, the odd phenomenon was observed that the gene sequences encoding eukaryote proteins did not make up unbroken contiguous stretches of nucleotides, but rather chains of several such contiguous stretches called *exons*, flanked by non-coding

regions called *introns* [257]. Not all eukaryotic genes are split up into exons, but most are [258, 259]. During transcription, the entire gene is transcribed, after which a complex of proteins and catalytic RNAs, termed the *spliceosome*, removes all introns and *splices* most or all of the exons together into a processed transcript, which can then be translated [79]. In many cases, there is a combinatorial variety with respect to which exons are retained and which are skipped, which is thought to depend on the particular cellular state [260, 261]. That is, expression of genes can be dynamically controlled not only with respect to how much is produced, but also with respect to the particular sequence which is expressed. As such, intron structure can in some sense be considered a functional property of genes. It also causes technical difficulties in orthology analysis and gene annotation, since comparison of different splice forms is likely to exaggerate sequence divergence [124, 259, 262].

This splicing-based extended regulatory framework is thought to be an important part of the solution to the so-called *N-value paradox* [263], which is that seemingly very complex organisms, such as ourselves, do not have much larger gene repertoires than organisms like yeast. By having much larger alternate splice form repertoires, multicellular eukaryotes may thus have much larger protein repertoires than their gene repertoires allow.

The sequence signals controlling the splicing process are not entirely understood, and reliable analysis of splicing therefore requires transcriptome sequence data for classifying genome regions as either introns or exons. Concerns have been raised in that differences in the degree of transcriptome sequence coverage between species may bias the results studies comparing exon-intron structures of proteins [259]. Despite this, some notable results have emerged.

It appears clear that there is a strong signal of conserved introns between species [264-266], but also a considerable degree of species-unique splice forms [267, 268] and exons [269]. The original discovery of introns sparked a debate on whether these sequences represents an initial state – an appealing hypothesis since it matches a model of a pre-protein ‘RNA world’ – which was subsequently lost in prokaryote lineages, or if they instead more recently have invaded eukaryotic genes [79], perhaps in the process enabling increased genomic complexity and subsequently more complex multicellular life. While there may still not be a definitive answer, probabilistic analysis has shown that a model with significant conservation of some very old introns explains the observed genomic structures well [266, 270].

2.3.4 Comparing protein function

Just as with domain architectures, comparisons of protein function are non-trivial. While free-text descriptions can be compared using text mining tools, such comparisons are best made on controlled vocabulary annotations. Schemes placing any protein into a single category allows for only similar/dissimilar comparisons, whereas hierarchies of annotations necessitate more complex metrics.

Much research has gone into finding similarity metrics for Gene Ontology annotation sets, ranging from considering the number of shared terms [271] to various normalized methods [272] as well as methods weighted by term rarity [272-275], many of which employ semantic similarity [276]. There are also methods explicitly taking the GO structure into account [277]. Ultimately, there is no objective way to benchmark the utility of such metrics except for a particular purpose, such as predicting shared disease involvement.

2.3.5 Predicting protein function

Experimental determination of what a protein does in an organism takes many forms. Chemical assays may demonstrate particular activity or affinity, whereas identification and isolation of mutants or their artificial construction in knock-out studies may demonstrate the phenotypic impact of the absence of a functional form of the protein [278]. Fluorescence studies may demonstrate where and when it is expressed [279], and methods like the yeast two hybrid technique can evaluate whether or not a pair of proteins interact physically [280]. However, there are strong limitations in practice to our capacity of characterizing the protein universe experimentally. The first is a factor of simple numbers – while experimental characterization of all genes may be carried out for a small number of model organisms, the number of sequenced genomes grows very rapidly at this stage in the history of science. The second is that not every interesting wet lab experiment is possible, either for technical reasons or for ethical reasons, in every relevant organism. Last, the sheer number of possibilities to test is problematic, and reducing the hypothesis space, or at the very least ranking possible functions by how probable they are given available data, would speed up the process.

As a result, the possibility of predicting function is an interesting one. Through access to the formalized results of experimental validation of a part of the protein universe, computational techniques drawn from various subfields such as sequence and language analysis [34, 277, 281] graph

theory [282, 283] and machine learning techniques like Bayesian networks [284, 285], Support Vector Machines (SVMs) [286, 287], or decision trees [288-291], we may be able to predict what function a protein is likely to have. Statistical techniques may allow assigning confidence measures to these predictions.

There are several main approaches that functional prediction has taken. The one most widely used is arguably to base predictions on the sequence of the query protein. The sequence can be used to identify high-confidence homologs or orthologs where function is known, and the existing annotations of those proteins can be transferred to the query [292]. Parts of proteins, such as domains, can be identified and functional annotations associated with them, if available, can be assigned. There are also non-homology based sequence-oriented function prediction methods, where the presence of small motifs [293, 294], secondary structure elements [293-296] or other low-level sequence features are used as a basis for functional classification through machine learning methods. It can be argued, however, that homology is difficult to rule out as the underlying reason behind the information content of these features.

Beyond sequence, query proteins with known structures, or with structures predicted through homology with subsequent molecular dynamics optimization, can be compared with respect to these structures to the structures of proteins with known function [295-297]. Furthermore, annotations can be transferred in a similar manner as with sequence. Alternatively, structure features such as active sites [295, 298], chemical surface properties and so forth [295], can be used to predict functional properties.

Using neither sequence nor structure information directly, Guilt By Association (GBA) methods [299] exploit the fact that interacting proteins are often part of the same processes [300], as are coexpressed proteins [300], particularly if their orthologs are also coexpressed [301], and proteins that lie in close chromosomal proximity [302]. An extreme example of the latter is when genes are found in the same bacterial operon, which strongly implies a functional relationship [303]. Annotations have also been transferred between proteins based on similarity of their *phylogenetic profiles*, i.e. the tendency of their orthologs or homologs to co-occur across a range of genomes [304-308].

Recent developments in constructing protein-protein interaction or association networks from a variety of experimental data, including coexpression data [309], shared genomic context [310], protein complex formation [280, 311], and transient interactions such as those detected

through yeast two-hybrid systems [280, 311] may also be used to define something similar to pathways. There are several composite network resources explicitly intended for functional representation, such as FunCoup [312] and STRING [313-316]. Using graph theoretic tools, networks may be subdivided into modules that often are functionally coherent; however, see Song & Singh [317] for a critical view. Interactions conserved across multiple species appear to be of particular predictive value [318, 319]. Multiple approaches have been suggested for propagating functional annotations across a network of this type [283, 320-327].

Another approach, that shares its names with a family of methods for multi-gene phylogenetic analysis, is *phylogenomics*. This involves mapping known function onto gene family trees and extrapolating functional subfamilies from them [285, 328-331].

All in all, there has been a veritable deluge of publications of algorithms for predicting protein function from one feature or other. Generally they are evaluated under some form of holdout analysis such as cross-validation, compared with some set of competing methods, and made publically available. It is unclear how many of these methods ever come to be applied in a practical context outside of their original or follow-up publications. One trend opposing this balkanization of the subfield lies in the construction of metaservers that apply a wide range of different algorithms to a protein query and integrates the results [295, 296, 332], and it could also be remediated by more widespread use of tools like the Distributed Annotation Service (DAS) framework [333].

2.3.6 Gene duplications and functional changes

What happens to a gene that is duplicated? Suzumu Ohno [334] explored this early by suggesting a model for the effects of a gene duplication on the selective pressures experienced by the resulting copies. There are several possible outcomes of this scenario.

One is the *nonfunctionalization* of one of the copies through pseudogenization and subsequent gene loss. While, conceivably, genetic drift may sometimes fix a duplicate in the population without its loss being associated with a fitness decrease, it appears unlikely that this would happen often, and as such, an entirely redundant duplicate should generally become non-functional [335].

How can the duplicate make itself useful? Ohno suggested a scenario, later termed *neofunctionalization* [336] under which a redundant copy would

become free to evolve (by redundancy relaxing negative selection), but soon experience positive selection that optimizes it to perform a novel function. Thus, the copy becomes necessary again in order for the organism to maintain its now-extended functional repertoire, and therefore it is fixed in the population.

Another option, termed *subfunctionalization* [336], could follow if the original gene function had multiple aspects or facets. Some of these could be lost due to redundancy in one copy, some in the other copy, which would lead to both duplicates complementing each other with respect to the original set of functions. This would remove constraints on one set of subfunctions resulting from a need to also maintain the other set of subfunctions within the same gene, and would thus allow subsequent parallel optimization of the subfunctions in different genes, leading to increased fitness relative to before the duplication. Thus, the duplicates would be more likely to be fixed in the population. Force and co-workers [336] refer to this model of subfunctionalization as the *Duplication-Degeneration-Complementation* (*DDC*) model. See also Stoltzfus [337] for a very similar model published at around the same time. Notably, both neofunctionalization and subfunctionalization could be achieved either through changes to protein coding sequence or to gene regulatory regions [338].

Fourth, if there would be some benefit of maintaining multiple copies of genes with the same function, this could also lead to fixation, an outcome termed *gene conservation* [339]. Ohno [334] originally suggested that this may be beneficial as a protection against gene loss, but later research has largely disproven this scenario (discussed in [338]). Another possibility lies in *dosage effects* [340, 341], where multiple genes could enable either faster or more flexibly controlled production of a protein.

While it appears likely that all the described scenarios occur throughout evolution, their relative importance remains under debate [342], and much ongoing research aims to clarify this matter. Regardless, it is clear that some degree of functional change, either regulatory or at the protein level, is expected to occur in most duplicates, while it should be much less likely in non-duplicated genes.

2.3.7 Protein function versus orthology

Experimental research demonstrated early on that gene knockouts in model organisms could be "rescued", restoring the lost function, by introducing an orthologous gene via gene technology [88, 343]. This matches well the predictions mentioned above, and the commonly held belief came to be that

orthologous genes generally perform the same function whereas paralogous genes generally do not.

There is ample evidence that genes that are orthologs generally perform the same function. Not only have many complementation experiments of the type described above been performed [343], but there is also a vast body of literature that supports functional and structural conservation in homologs [344-350]. Since many homologs are also orthologs, this tells us that orthologs generally are well conserved in these respects. What it does not tell us, is if the orthologs are better conserved than other homologs, i.e. paralogs [4].

This orthology conjecture is theoretically attractive, for reasons described in the previous section, but has not been explicitly tested to any great extent. In recent years, some scientists have questioned it and instead proposed a model under which the rate by which homologs diverge functionally does not depend on their orthology status [3, 4]. Explicit and comprehensive testing of the hypothesis of increased conservation of orthologs, then, becomes relevant.

The central problem in analyzing the relative conservation of some property of orthologs versus paralogs, comes from the issue of determining evolutionary time. Obviously, we can only infer the time that may have passed since the divergence of a pair of homologs, not observe it directly. Moreover, we can infer it only through degree of observed differences between the sequences. Evolutionary rates may vary both between genes and between lineages. Organisms with short generations and little proofreading will accumulate neutral mutations at a higher rate, whereas variations in selective pressures - including occurrence of evolutionary bottlenecks - will further affect the relative rate of non-neutral mutation accumulation [351].

Some studies have explicitly investigated differences between orthologs and paralogs in the conservation of other features relative to (some measure of) time. It has been shown that protein sequence divergence increases following gene duplications [352, 353]. Another analysis demonstrated that the structural divergence of domain sequences is lower in orthologs than in paralogs at the same level of sequence identity [354]. However, overall, the question has not been addressed to a great degree.

Another obstacle to analyses of this type, which may be harder to address, is that of biased retention of gene duplicates, or the ‘Davis and Petrov effect’ [353]. The likelihood of a duplicate gene coming to fixation appears to differ between different classes of genes, meaning that comparisons of orthologous gene pairs and paralogous gene pairs may in fact also be comparisons of

genes involved in different processes, and thus, subject to different evolutionary pressures [4]. Curiously, it appears that the set of genes most frequently experiencing fixation of duplications in worm and yeast is in fact enriched for slow-evolving genes [355]. If this effect holds in general, then, it might serve to hide rather than cause artificial semblance of an increased functional divergence following gene duplications.

2.3.8 Protein function versus protein sequence conservation

The more general question of the relationship between homology and functional similarity has, in contrast, been adressed to a relatively large extent. The basic idea of homologs being more functionally similar on average than unrelated proteins can be considered to have been effectively proven true, though exceptions have been documented where either protein function has evolved convergently or where homologs have diverged in function [356]. More interesting is perhaps the extent to which functional conservation varies with sequence divergence or with time that has passed since divergence.

The early studies in the field concerned themselves primarily with enzymes, as the EC system from the start provided datasets where the questions asked would be well-defined. A variety of sequence identity threshold studies were carried out, where the main objective was to identify above which level of sequence identity enzyme functional annotation could be reliably transferred [344-350]. Similar results have been presented for Gene Ontology annotations [349], for KEGG pathways [350] and FunCat functional categories [350]. Biases in dataset composition were identified as a factor affecting the results of these studies [347], making for a range of different thresholds reported, but the main pattern holds true across studies, in that there exist thresholds of this type for conservation of enzyme function. Notably, it is difficult to say whether or not these thresholds reflect structural necessity or simply the extent to which protein fold space, generated through the stochastic processes of domain recombination and gene family expansion, happens to have become populated.

In analogy to the function versus sequence results, relationships between sequence and structure divergence for homologous proteins have been identified and independently validated [347, 357]. Given that the 3D structure of a protein is required for its function, the general agreement between these studies is expected and reassuring from a methodological perspective.

Sequence identity, as in percentage of identical characters in an alignment, is a crude proxy for sequence divergence, however. It is coarse-grained, and has no way to account for mutational saturation, making it increase in a non-linear fashion as time progresses [358, 359]. Corrections can either be applied to this measure in order to estimate phylogenetic time, or the analysis can be performed using proxies such as BLAST bit scores [12], which are implicitly informed regarding the functional relevance of differences or similarities through the use of a substitution matrix. However, as these are ultimately confidence rather than distance measures, such measures remain at least in theory mere proxies for true evolutionary distance. Evaluations have been performed of how sequence and structure diverges relative to these measures as well [360].

2.3.9 Protein function versus domain architecture

It can easily be shown that protein function annotations correspond to some extent with the presence of particular domains [165, 361, 362] or domain architectures [Paper II, 182, 289]. While there exists some degree of circularity here, as many annotations have been assigned based on the presence of domains or domain architectures, the correspondence also exists for purely experimental functional annotations. That is, the function of a protein clearly is related – in a statistical sense – to its domain architecture [182, 362-364].

From a structural or mechanistic perspective, it is frequently known how protein functions are carried out, and in those descriptions, properties of the constituent domains such as binding affinities or active site characteristics often play important roles [298, 365, 366]. In many isolated cases, thus, it is known that the particular way in which a protein accomplishes its tasks requires the specific domain architecture that it has. Why, then, question the common assumption that proteins possess functions that follow from their domain sets?

The main problem here comes from the fact that proteins are gradually evolved through successive descent, rather than assembled de novo for each organism from some pool of available domains. Orthologs in different species – even distant species – as well as paralogs within the same species frequently share not only the same domain content, but the order of those domains [225, 229, 236]. Reconstructions of domain architecture evolution strongly indicate that most shared architectures are shared, not because of convergent evolution but rather because of common descent [203, 232, 235].

While a certain domain architecture clearly does “cause” a protein to have a particular function, there may be any number of alternative domain architectures, involving different domains, that could be configured to cause the same function to be carried out. That we do not see them in nature would also follow if it were the case that all proteins performing that function were descended from a common ancestor. As such, we cannot conclude from these absences only that a given domain combination is the only way to implement a particular function.

Similarly, domains are often broad classes of related sequences and structures, while functional specificity may follow [367, 368] from just a few residues in some cases. A given domain architecture might potentially be repurposed for very different functions. Again, the fact that this may not have been observed for a given architecture does not make it impossible.

A counter hypothesis to the strong form of the domain grammar hypothesis would then instead state that most domains and domain architectures in theory can be made to accomplish a variety of functions through fine-tuning of a small number of residues, but that only a few of these were realized early on in evolution. Subsequent proteins filling these functional roles have then descended from these ancestral proteins. This counter hypothesis also predicts the pattern of correlations between domain architectures and protein functions that we observe, necessitating more complex approaches in order to test the domain grammar hypothesis.

3 Summary of present investigations

3.1 Objectives

As mentioned in the introduction, the scientific motivation behind this work has been the evaluation of the extent to which commonly used tools and concepts – domain architecture combinations and the distinction of orthologs from paralogs – are useful in order to characterize the functions of novel proteins. This question has been approached both by direct evaluations and indirectly by searching for evidence for those evolutionary mechanisms which justifies the use of these concepts for the purpose of protein function prediction. To enable the furthering of these objectives, some evaluation and improvement of practical methods were also undertaken, and where possible, any applications of the work have been made available to the scientific community at large.

Clearly, there exists a widespread belief that the domain architecture of a protein determines what functions it carries out. As stated in the introduction, I term this the *domain grammar hypothesis*. I further distinguish between its strong form, which is the assumption that the domain architecture of a protein is what causes it to have certain functions, and that particular functions require particular domain architectures, and its weak form, which says that if we learn the domain architecture of a protein, we can use this information to predict its function. The latter will hold even if domain architecture is only indirectly associated with function.

There also exists a widespread belief that an orthologous relationship between two proteins will directly reduce their expected degree of functional divergence relative to the length of time they have been separated, all other things being equal. This has been referred to as the *orthology conjecture* in a previous work [3].

Null hypotheses can be formulated for either of these two beliefs, under which most or all protein functions could evolve regardless of the domain

architectures of the proteins involved, and under which gene duplication does not appreciably change selective pressures so as to change the pattern of functional divergence over time, respectively.

Can we formulate a test for the domain grammar hypothesis? For the weak form, we need only demonstrate that protein domain content can be used to predict protein function. This is accomplished merely by constructing a function prediction algorithm that takes only domain architectures as input. For the strong form, devising a suitable test becomes more difficult. A suitable starting point can be the functional comparison of same-architecture and different-architecture protein pairs. However, since both function and architecture diverge with time, we must compensate for the latter factor. A more complex test, then, would be the functional comparison of same-architecture and different-architecture protein pairs at equivalent separation in time.

Unfortunately, even observing higher functional conservation in same-architecture pairs in such a comparison would not prove a causal link, because there is an alternate model that could also explain the observation. Suppose that the degree of protein function conservation after a given period of time has passed since divergence is dependent not on whether the domain architecture is conserved, but whether or not enough of the protein structure is conserved, regardless of whether this conservation of structure corresponds to conservation of the domain architecture or not. Since conservation of domain architecture leads to conservation of structure, we would in that case expect an indirect link between conservation of domain architecture and conservation of function. As such, not even the test suggested above could prove that conservation of domain architecture is a requirement for conservation of function.

However, there may be indirect evidence to be found for the domain grammar hypothesis, by way of evaluation of the orthology conjecture. Testing that hypothesis should be fairly straightforward in theory, by comparing orthologous and paralogous protein pairs functionally, relative to how long they have diverged. In practice, the feasability and results of such a test will depend on what definition of protein function is used.

Furthermore, if we would see a difference in conservation of some property between orthologous and paralogous protein pairs, and that difference is not an artefact resulting from method or dataset bias, we can relatively safely conclude that it must result from differences in selective pressure on the proteins in the pairs from these categories. That is, if paralogs are less conserved relative to time in some respect, we have no other explanation than that this must be the direct or indirect result of a process such as the

previously introduced models for functional change following a gene duplication event. This in turn tells us that any such differentially conserved properties are likely to be functionally relevant, so that either negative selection on them which otherwise would have taken place is now relaxed, or that positive selection on them now works to optimize some new functional capacity.

As such, demonstrating reliably that some property is better conserved in orthologs over time than in paralogs, assuming that the results are reliable, and that we have no alternative explanation than changes in selective pressure resulting from duplication events for why ortholog and paralog evolution would differ, would provide us with indirect evidence of both the orthology conjecture and of the functional relevance of the conserved property. In the case of conservation of domain architecture, then, demonstrating higher relative conservation of domain architectures in orthologs than in paralogs would provide indirect evidence for the strong form of the domain grammar hypothesis as well. While neither hypothesis would be proven, given that unknown alternative explanations could not be ruled out, such observations would nevertheless provide strong implications for both.

3.2 Paper I: Domain tree-based analysis of protein architecture evolution

3.2.1 Summary

This work had two main objectives. First, it aimed at developing and making available a method for ancestral domain architecture reconstruction which would require no *a priori* species trees or even gene trees, which means that theoretical and practical issues with building a phylogenetic tree for sequences where recombination has occurred are circumvented. Second, it aimed at using this tool to determine how common domain architecture reinventions are, i.e., cases where convergent evolution has given rise to the same domain architecture independently in multiple lineages. Aside from being an interesting question in its own right, this becomes relevant as a reliable inventory of domain architectures which have or have not evolved convergently might provide evidence for the domain grammar hypothesis if there are signs that some domain combinations, but not others, are selected for during domain architecture evolution.

The basic approach was straightforward – given a (domain) phylogeny, a maximum parsimony reconstruction (which is guaranteed to find a set of assignments involving the smallest number of mutation events possible) was used to assign domain architectures to the internal nodes of the tree. For each domain family, a domain tree was computed using neighbour-joining. As an extension of standard phylogenetic bootstrapping [369], for each domain family alignment, a large number of corresponding pseudoalignments were contructed by randomly sampling columns from the original alignment. For each pseudoalignment, tree reconstruction and ancestral domain architecture reconstruction was performed separately. To conclude either polyphyly or monophyly of a multi-domain architecture, then, we required that a majority of pseudoreplicates agreed with the predicted architecture origin events for each of its domains.

Carrying out this analysis for several datasets ranging from 62 to 96 species, both with and without filtering out sequences where there may be missing domains, we concluded that between 5.6% and 12.4% of multi-domain architectures have more than one origin in this dataset. This estimate is higher than one made by another group [235] using SUPERFAMILY [178, 182] rather than Pfam [174] domains. These differing results may stem from differences between the underlying domain databases, or it might reflect differences between the projects in scope and the degree of dataset pruning that is done.

3.2.2 In retrospect

This project suffers from some method limitations. If it were replicated, arguably it should be made to encompass the same pre-filtering for potentially missing or incorrect domain assignment as was done by Gough [235]. While the per-domain analysis is powerful, a maximum-likelihood approach might provide more reliable trees than the distance-based approach used here, but the main problem lies in the interaction between the maximum parsimony framework and the quality of the trees. Since many domain sequences are highly similar, the reconstructed trees will have some very poorly supported branches that are usually very short, or even have zero or negative lengths where the distances between sequences cannot be accurately represented by any tree. These correspond to places where no reliable binary branching order can be shown to win out, and admitting our uncertainty in these cases would correspond to collapsing these branches, creating a multifurcating tree with only trusted branches. When forcing such

data to conform to a binary tree, this means that much of the inferred topology may in fact only reflect dataset noise.

As parsimony does not consider branch lengths at all, or, rather, considers change along one branch as equally important as along any other, the particular topology of a binary tree to which it is applied will impact the results significantly. That is, the combination of binary trees with short, unreliable branches and maximum parsimony [54, 55] applied along its topology, will make the final results sensitive to noise. While our bootstrap approach [369], as applied in the paper, may serve to prevent incorrect conclusions from being drawn as a result of this, it remains fundamentally an *ad hoc* approach.

A more elegant solution would be either to apply an extension of maximum parsimony to multifurcating trees [150], and collapse unreliable branches as per the above, or to replace the parsimony framework with a maximum likelihood framework, which would allow explicit consideration of branch lengths. How much the results of an analysis of this type would shift if this were done is not known.

We have, however, begun construction of a maximum likelihood algorithm for ancestral domain architecture inference. At present, the algorithm allows scoring a set of such assignments along a gene tree. A draft software implementation also exists, and we have tried to use Markov Chain Monte Carlo (MCMC) [370] methods for optimizing ancestral domain architecture assignments. Unfortunately, performance on simulated data is still relatively poor, possibly because the ancestral domain architecture assignment search space is not sampled efficiently enough.

How could we test the domain grammar hypothesis from data of this type? Evidence of selection acting on which domain architectures are reinvented (i.e. produced through mutational shuffling and retained by selection or genetic drift) would imply that not all novel architectures are equally useful in every context, in line with the domain grammar hypothesis. However, to determine that selection takes place requires a suitable neutral model as comparison. The most feasible approach, then, might be to consider a large-scale, reliable set of domain architectures that have arisen one or more times throughout evolution, and test whether or not the distribution can be fitted to a model where all combinations have equal selective fitness. Another approach could be to study the functional similarity of pairs of proteins sharing the same domain architectures. If domain architectures are only indirectly associated with protein function, pairs sharing architecture by convergence should be functionally much less similar than pairs sharing architecture by descent.

3.3 Paper II: Predicting protein function from domain content

3.3.1 Summary

A central question within this thesis as a whole is the extent to which domains and domain architectures are relevant to protein function, i.e. the domain grammar hypothesis. The goal of this work was to test this by proposing a model for how the function of a protein might follow from the set of domains and domain combinations that it contains. Several versions of this model were evaluated, both a rule-based system and a model for assigning a probability that a protein possesses a particular function given its domain architecture. During the course of the project, we discovered that similar work had been carried out by the Interpro [166] team with respect to single domains; we accordingly formulated our rule-based approach as an extension of their method. The probabilistic method also had the advantage of being insensitive to at least a moderate fraction of false negative training examples, something we expected in a large-scale dataset unless it was subjected to significant manual curation prior to the analysis.

To evaluate our model, it was tested on the intersection of the Gene Ontology [252-254] annotations available at the time with a non-redundant subset [371] of UniProt [256], re-predicting annotations under cross-validation. As comparison, a non-domain centric approach, annotation transfer from the closest BLAST [18, 19, 49] hit to an annotated protein, was used. While this approach remained more sensitive on average, it was also less accurate, and more prone to false positives, presumably through matches to only one or a few domains in an architecture. Combinations of multiple domains increased sensitivity beyond use of single domains only, with little loss of accuracy. Last, as expected, the probabilistic method performed better on average than the rule-based method.

We further identified a number of cases where proteins containing particular combinations of domains were significantly enriched for some functional term, but where such enrichment at the same time was not observed for any of its constituent domains or domain subsets. These cases may represent instances of complex function arising through the interplay of multiple domains. While there exists significant risk for circularity in an analysis of this type due to domain information having been used to annotate the proteins in the first place, the central results reported here remained robust even when excluding any Gene Ontology annotations having the Inferred

from Electronic Analysis (IEA) evidence code. This by no means removes all circularity, but it provides an indication that the results are not entirely artefactual.

3.3.2 In retrospect

The basic method used in this work remains sensible, but datasets have evolved since it was performed, particularly through more concerted and consistent efforts by Gene Ontology curators. Applying this method on a larger scale still, but on curated data (which might be possible within the Interpro setting) could potentially yield better and clearer cases of domain combinations which bestow additional functional information beyond that contained in the single domains themselves. Unfortunately, the list presented in the paper may not be particularly reliable, as no correction was performed for the massively multiple hypothesis testing (effectively all combinations of domains versus all Gene Ontology terms). Any reapplication of the algorithm must ensure that this problem is taken care of.

From a theoretical perspective, while the results show in practice that domain architectures may be of use for functional predictions, this does not prove the validity of the presented model as to domain combinations directly causing particular functions. In particular, large protein families, within which both functions and domain architectures are conserved, would produce these results just as much as particular domain combinations being absolutely required for particular functions would. As such, while the study supports the weak form of the domain grammar hypothesis, it cannot say anything about the strong form.

The validity of our approach for functional prediction and identifying functionally relevant domain combinations was recently shown in an independent study. The most recent update of SUPERFAMILY [182] contains a highly similar approach, and yields comparable results.

3.4 Paper III: Benchmarking homology detection procedures with low complexity filters

3.4.1 Summary

This work in fact resulted from preparations for the work reported in Paper IV and Paper VI, but addresses a question of some bioinformatical relevance in its own right. For creating the initial set of homology inferences that InParanoid [Paper IV] is built from, the BLAST program [18, 19, 49] is used. This tool, stated most generally, infers homology from observations of similarity between sequences. However, not all sequence similarity is a consequence of homology. The presence of mobile repeat elements causes similarity between sequences not otherwise homologous, and there are also cases where proteins share uncommon amino acid frequency distributions without being homologs. These factors then cause non-homologous proteins in which they are found to appear more similar than otherwise expected, and again gives rise to false positive homology assignments. In many cases, this causes significant problems for homology-based analyses. Various preprocessing steps or modifications to homology inference methods have been proposed, but their respective trade-offs in terms of false positives and negatives had not been conclusively evaluated.

Benchmarking a homology inference tool or pipeline suffers from the problem that identification of high-fidelity non-homologs can be non-trivial, particularly if the type of problems that algorithms may run into in realistic applications should also appear within the evaluation. Single domain sequences corresponding to known 3D structures have often been used [372], but this biases the dataset away from the complicated cases where we were most interested in performance. Though in Paper I we concluded that domain architecture reinvention was more common than previously reported, it is still rare enough that a useful and arbitrarily large benchmark can be constructed, according to simple classifications of pairs of domain architectures as either clear homologs, unclear cases and clear non-homologs.

Using our benchmark dataset, we evaluated a number of different options, pre- and postprocessing steps for the BLAST algorithm, in a context similar to that used in InParanoid for genome versus genome comparisons. The results showed that the most recent BLAST options incorporating compositional adjustment of scoring matrices [49] achieved the best

performances, though with the side effects that reported alignments were sometimes truncated.

3.4.2 In retrospect

The approach we present for benchmark dataset construction has some problems, notably that it relies on substantially complete domain assignments and on the Pfam higher-level Clan [172] organization to also be complete, in the sense that any two domain sequences that are homologous can be assumed to also be part of the same Clan. While the Clan system is continually being expanded, this goal has not yet been met. It seems doubtful, however, that this fact should have a large effect on the results of the benchmark. Plans are underway, though not yet completed, for applying the strategy described here to generate benchmark datasets, but based on different domain databases, to measure the extent to which classifications of pairs of proteins as homologous, non-homologous or unclear changes with choice of domain database.

3.5 Paper IV: InParanoid 7: new algorithms and tools for eukaryotic orthology analysis

3.5.1 Summary

The InParanoid framework [105-107] is a heuristic approach for inferring orthology and paralogy relationships between the proteins of two species at a time. It is particularly notable in its clear resolution of in-paralogs from out-paralogs, and the relatively swift speed by which it can be built. Its main limitation at this stage, conversely, is that it can only answer questions posed in the context of two species. Paper IV presented an update release (7.0) of the InParanoid database, with all pairwise comparisons between a hundred eukaryotes considered, and with the codebase updated to version 4.0.

The main changes in this update follow directly from the findings in Paper III, in that the underlying homology inferences are now computed using two successive BLAST passes with different filter settings. The base threshold for homology is lowered from 50 bits to 40, but at the same time, hits must

also pass more stringent overlap criteria. All in all, these changes improve both sensitivity and accuracy, resulting in more accurate orthology inferences.

3.5.2 In retrospect

The improvements to InParanoid provided a reliable orthology resource on which the remaining investigations described in this thesis could build. While there is little difference in the case of most species, performance has now improved greatly for some species where high prevalence of coding sequence repeats previously caused many sequences to incorrectly aggregate into very large orthology clusters – the slime mold [373] in particular. As such, the advances made through Papers III and IV allowed the orthology-centered analysis of Papers V and VI, as well as the additional novel results presented in this thesis.

3.6 Paper V: Orthology confers intron position conservation

3.6.1 Summary

Conservation of intron positions between two homologous genes can be thought of as a measure of the extent to which their splicing-related regulation is conserved, which can be seen as a property of function. As such, under the orthology conjecture, we would expect such conservation to be higher between orthologs than between paralogs. Using InParanoid orthology inferences and a simple measure of intron position conservation between a pair of aligned proteins, comparing orthologs and paralogs with respect to this property can be fairly straightforward. However, as divergence of any property is expected to increase with time, the analysis is only meaningful if somehow stratified for differences in divergence. Therefore pairs of orthologous or paralogous proteins were divided into bins based on sequence identity, and the average intron position conservation was compared between the groups within each such bin and the difference evaluated for statistical significance.

Overall, the analysis showed significantly higher intron position conservation for orthologs than for paralogs (in support of the orthology

conjecture), but also some anomalies, notably that inparalogs exhibited significantly higher conservation than outparalogs, as well as a large fraction of even close homologs with no conservation of intron position at all.

3.6.2 In retrospect

During the course of Paper VI as well as the other investigations described here, a number of possible pitfalls for an analysis of these types became apparent, notably that a binning analysis encompassing too few bins (and hence, too wide-ranging bins) coupled with an uneven distribution of the binned variable will cause artificially significant differences to appear between categories. It may also be that protein sequence identity is a poor proxy for evolutionary time, such that saturation of mutations will lump together almost all pairs above a certain divergence into the same bin.

Attempts were later made at applying the same binning strategy and distance measures as in Paper VI to the homologous pairs from Paper V, where we expected that the difference between in- and outparalogs should turn out to derive from the artefacts mentioned above and thus disappear with these changes. Strangely enough, that difference remained significant, whereas the difference between orthologs and outparalogs lost significance. This might imply biases in intron position recognition such that a protein with a same-species homolog is more likely to have been annotated as having introns at the same positions as its inparalog.

However, very recently at the time of writing, a bug was found in the parsing scripts used to build the dataset for Paper V. This is likely to have caused many pairs of orthologs as well as paralogs to appear much more dissimilar in intron structure than they actually are. This explains the larger-than-expected fraction of homolog pairs lacking any intron position conservation at all, and while there are no *a priori* reasons to expect that this artefact should affect ortholog and paralog pairs differently, it does make drawing strong conclusions from the results inadvisable at this point.

3.7 Paper VI: Domain architecture conservation in orthologs

3.7.1 Summary

In this project, like in Paper V, we aimed at evaluating the orthology conjecture by comparing orthologous versus paralogous pairs of homologs with respect to similarity of domain architecture. We have not been able to find any previous attempts to study this; while Hulsen and co-workers [90] investigated the degree to which orthologs shared Interpro domain families, they did not contrast this with the corresponding conservation in paralogs. The approach is fundamentally very similar to that of Paper V – InParanoid was used to identify pairs of orthologs and paralogs, Pfam domain architectures were assigned and compared, and differences between the categories were measured and tested for statistical significance within bins corresponding to different ranges of sequence divergence. To get around the problem of saturation, Jukes-Cantor corrected protein sequence identity computed from the aligned regions was used in place of raw sequence identity. Bin borders were also optimized so as to avoid forming bins which were very sparsely populated from one category but not the other.

Overall, we could demonstrate that, above a certain degree of divergence, orthologs were relatively better conserved than paralogs at the same evolutionary distance, whereas no such trend was seen when comparing in- and outparalogs. Changing the number of bins had at most a small impact on the results, implying that the observations are, in fact, reliable. Furthermore, considering the average conservation of domain architecture when comparing human with successively more distant species, we observed a larger drop in conservation than expected when going outside the vertebrates. This is consistent with the model that two consecutive rounds of whole-genome duplication occurred early in the vertebrate lineage (the *2R hypothesis*) [374].

3.7.2 In retrospect

Of the publications included in this thesis, Paper VI provides the most reliable support for the hypotheses investigated, with observations matching predictions from a combination of the orthology conjecture and the domain grammar hypothesis. A recent study (performed later but published earlier)

using a tree-based method [142] also found that changes in domain architectures were around twice as common following gene duplications as following speciations, which is consistent with our results.

One potential problem, which we discuss in the paper, is that domain architecture changes conceivably could make recognition of some orthologous relationship more difficult. This is also supported by recent benchmarking efforts [96], and could potentially bias the set of orthology inferences towards orthologs with more similar architectures and outparalogs with less similar architectures. As such bias lies in the direction of our results, it is a relevant problem, though one difficult to circumvent. While indirect properties of orthology, such as expectations of conserved synteny, may not be affected by the biases that influence sequence similarity, they are not generally applicable across large evolutionary ranges. Phylogenetic reconstruction generally depends on sequence similarity, and as such remains vulnerable. Careful curation of the ortholog set could be used to minimize its potential impact, but makes constructing a dataset large enough to perform statistical analysis difficult and may also introduce bias. However, the fact that our observations hold for comparison of orthologous pairs versus inparalogous pairs is promising, as the proteins in these pairs exhibited reduced domain architecture conservation for some ranges of separation while still being considered ortholog cluster members.

Further testing of the robustness of these conclusions could certainly be carried out, notably by evaluating more species comparisons, the use of phylogenetically derived orthology inferences, and by repeating the analysis using structural domains (such as SCOP/SUPERFAMILY) in place of Pfam domains. Likewise, other proxies for evolutionary time, such as the number of synonymous substitutions, could be used to further validate the results. It would be interesting to see how species comparisons at different evolutionary distances or from different lineages affect the outcome, though this is also problematic as it becomes more difficult to sample enough pairs from the various categories at equivalent divergence times to be able to conclude that any observed differences are significant.

3.8 Additional work: Direct functional conservation analysis of orthologs and paralogs

3.8.1 Introduction

The obvious continuation of the analyses performed in Papers V and VI would be, of course, to evaluate the ortholog conjecture directly by comparing the functional similarity of orthologs and paralogs at equivalent levels of divergence. The exact same framework can be used, and we applied such an analysis using as definition of function experimental annotations from the Gene Ontology.

3.8.2 Materials and methods

We compared the degree of conservation of Gene Ontology annotation of ortholog and non-ortholog protein pairs acquired by pooling data from all pairwise comparisons of nine well-studied model organisms. Their species were chosen because they are common model organisms, are part of the Gene Ontology Annotation project [375], have sequences of high quality available, and because the resulting set allows a spread across a large evolutionary range. They include the plant *Arabidopsis thaliana*, the fungi *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, the amoeba *Dictyostelium discoideum*, the worm *Caenorhabditis elegans*, the fly *Drosophila melanogaster*, the fish *Danio rerio*, and the mammals *Mus musculus* and *Homo sapiens*.

As the Gene Ontology sub-ontologies are potentially very different with regards to completeness of annotation or conservation of function, we separated them throughout the analysis. To avoid drawing any conclusions from terms assigned based on orthology, only terms assigned using the following Gene Ontology evidence codes were retained: EXP (Inferred from Experiment), IDA (Inferred from Direct Assay), IPI (Inferred from Physical Interaction), IMP (Inferred from Mutant Phenotype), IGI (Inferred from Genetic Interaction), IEP (Inferred from Expression Pattern).

The Gene Ontology annotated datasets were obtained either from the GO website, from ENSEMBL BioMart [376], or from the respective model organism websites. Any GO annotations with a classifier, such as 'NOT', were excluded from the analysis. The file

http://archive.geneontology.org/latest-termdb/go_daily-termdb.obo-xml.gz was downloaded on July 7 2009. Proteins lacking annotation within a given category were excluded from analysis of that category. For all proteins, any Gene Ontology terms that were strictly implied from the existing annotation, given the structure of the GO, were added, excepting the non-informative ontology root terms.

The InParanoid 7 clusters between all the species in the dataset were constructed using the latest version of the InParanoid software [Paper IV]. Sequences were retrieved from the datasets used to build the InParanoid database. NCBI BLAST version 2.2.18 was used to generate the underlying set of homology inferences, with the MSPcrunch BLAST parser [Paper III] used to merge compatible HSPs. As with InParanoid 7, a two-pass BLAST approach was used where all nonredundant proteins in one species dataset were used as queries against those of another species dataset, using compositional adjustment and SEG low complexity filtering. Any hit achieving a bit score above 40 at this step was realigned with compositional adjustment and low complexity filtering turned off. This approach was taken because of our findings that compositional adjustment, while much more accurate for homology assignment, produces shorter alignments [Paper III].

All comparisons of two species from those included were considered, yielding a total of 36 species comparisons. In each comparison, we considered two types of protein pairs. Ortholog pairs are any pair of proteins, one from each species, that are part of the same InParanoid cluster. Non-ortholog pairs are any pair of proteins, one from each species, with a BLAST bit score [18] above 40 (the cutoff used in InParanoid) that are not part of the same InParanoid cluster, and that, additionally, fulfill InParanoid's sequence overlap criteria; that is, that the sequence range spanned by the aligned regions must be at least 50% of the length of each protein in the pair, and the residues aligned must comprise at least 25% of each protein. To avoid falsely classifying proteins from clusters that for some reason were missed as non-orthologs, only proteins that are part of an InParanoid cluster are included in the analysis. To allow robust statistical analysis, pairs of each type were pooled between all species comparisons for each category separately.

Two proteins can be said to be functionally identical if they have identical sets of Gene Ontology annotation terms. When two proteins share some, but not all, of their Gene Ontology terms, this may reflect either a difference in function or insufficient annotation. Mistry & Pavlidis [271] introduced several measures for comparing Gene Ontology annotations of proteins. Their basic approach was the Term Overlap (TO), which is simply the number of annotation terms, ancestors included, that two proteins have in

common. Expanding on some other options they evaluated, we also implemented a Jaccard-normalized version of TO (dividing the number of shared terms with the number of terms possessed by either protein) and a version of the Jaccard-normalized score weighted by the information content of the terms, so that rare terms contribute more to how functionally similar the proteins are considered to be. We found the Jaccard-normalized TO (JTO) score to be the most relevant for our purpose, as it corresponds to the accuracy when assigning any annotation term through transfer from an ortholog or a homolog. Because of this, those scores were used for this analysis. Weighting by information content affected the results only slightly and we observed the same trends in the data when doing so during early trials.

As orthologous protein pairs are usually more similar in sequence than other homologs, it is non-trivial to separate the effect of orthology from the effect of close homology. The approach we took was to compare orthologous and non-orthologous protein pairs at the same sequence distance, in order to isolate the specific contribution of orthology. The data was divided into "bins" by sequence divergence, using Jukes-Cantor corrected sequence distance (computed using the same procedure used in Paper VI). The sizes of the bins were selected using an iterative procedure (identical to that used in Paper VI) to ensure each contained sufficient numbers of pairs from both categories under comparison. Within each bin, we considered the functional similarity of orthologs versus non-orthologous homologs, i.e., paralogs.

The number of ortholog and non-ortholog pairs each protein is involved in will vary considerably. In particular, the number of ortholog pairs defined by an InParanoid cluster is roughly proportional to the square of the number of cluster members. Because we were interested in the functional similarity of two arbitrary proteins, avoiding bias to the analysis from a relatively small number of large gene families was desirable. Because of this, in the statistical analysis, the statistical contribution from each protein pair was weighted inversely to the frequencies of its constituent proteins in the protein pair dataset.

Within each bin, the weighted means for each category were tested for statistically significant difference under a randomized permutation test [377]. The protein pairs from both categories were pooled, and new sets of protein pairs, equal in size to the two categories, were drawn from this pool. This process was repeated 1000 times, and the number of times when the difference in weighted mean between the two sampled set was at least as large as that between the two categories in the original dataset was recorded. The frequency with which that occurred then corresponds to the P-value for the null hypothesis, that the two categories would stem from a distribution

with the same mean value. Within each experiment, these raw P-values were corrected for multiple tests (across the bins) using Bonferroni correction. The categories were considered to be significantly different if the corrected P-values were below 0.05.

3.8.3 Results

Figure 1A–C shows the mean JTO score for orthologous versus cross-species outparalogous protein pairs across sequence identity bins, for the three Gene Ontology sub-ontologies, using only proteins with experimentally determined annotations. The error bars show the standard error of the means, and the upwards-facing triangle markers indicate whether the means are significantly different between the categories ($P < 0.05$, permutation test, Bonferroni corrected for twenty observations per plot). On average, ortholog pairs show higher JTO than cross-species outparalog pairs, though it is only consistently significant for the Cellular Localization sub-ontology. That sub-ontology further displays an oddity in that conservation in either category does not decrease appreciably with sequence divergence. In the other two categories, out-paralog JTO decreases, but not ortholog JTO. Notably, overall JTO is much lower for the Biological Process sub-ontology than for the other two.

Figure 2A–C shows the equivalent plot for orthologous versus inparalogous protein pairs, which are also expected to show a significant difference under the orthology conjecture. We do see a difference, but it is the opposite of what was expected: inparalog pairs are strongly and significantly more similar than ortholog pairs at all degrees of divergence. The inparalog pairs also do not show lower conservation in the Biological Process sub-ontology, which is seen for all other pair categories.

Figure 3A–C shows the equivalent plot for inparalogous versus same-species outparalogous protein pairs. This is intended as a negative control, as the orthology conjecture predicts that inparalogs and same-species outparalogs should not differ if the assumptions of the analysis are valid, i.e. that the divergence measure or function conservation measure is not biased in any way. The inparalog pairs are significantly more similar than the same-species outparalog pairs, though the separation is not as large as in some of the other comparisons. There is less overall conservation in the Biological Process sub-ontology than in the other two sub-ontologies.

Figure 4A–C shows the equivalent plot for cross-species versus same-species outparalogous protein pairs. This is intended as a negative control in the same sense as the previous plot. The same-species outparalogs are

consistently significantly more similar than the cross-species outparalog pairs at all degrees of divergence. There is again less overall conservation in the Biological Process sub-ontology than in the other two sub-ontologies.

3.8.4 Discussion

There are several anomalies in the results that are consistent neither with the orthology conjecture nor the underlying assumptions of the experiment.

Notably, cross-species outparalog pairs are much less similar than same-species outparalog pairs. A likely explanation for this is that there is a strong species bias in the experimental Gene Ontology annotations, such that different model organism database curators have focused on different properties or chosen different terms for describing similar properties. Furthermore, different model organisms may form model systems for different functional classes, concentrating research there. This is likely also the cause of the unexpected and vast difference between orthologs and inparalogs observed. It could also be that many proteins have gained species-specific functions, by becoming integrated in species-unique contexts and processes.

It is disconcerting that so many of the categories show no clear decrease in functional similarity with evolutionary divergence. Given the overall low similarity within cross-species pairs, it may be that the data is generally so sparse that such a trend is obscured through biases with respect to which proteins that have nevertheless been annotated with particular terms. Alternatively, this could reflect issues with using a protein-based divergence measure rather than actual evolutionary time estimates from nucleotide sequences.

While there does seem to be support for a somewhat greater functional conservation in orthologous than cross-species outparalogous genes, consistent with the recommendation that orthologs should be chosen over other homologs when transferring annotations from model organisms, the issues with this analysis as a whole indicates that the experimental framework, and perhaps the current state of the Gene Ontology annotation set itself, are insufficient for drawing any reliable conclusions on the orthology conjecture. This could be due to sampling effects and to biases in which genes have been evaluated for which potential annotations.

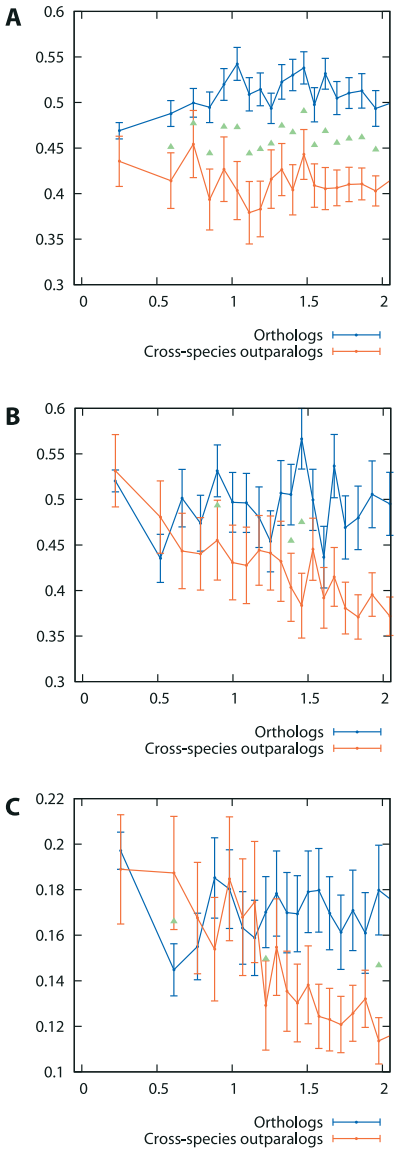

Figure 1

JTO conservation of orthologs versus cross-species outparalogs, relative to separation measured as expected number of amino acid substitutions per site. Results shown separately for the three Gene Ontology sub-ontologies. Error bars show standard error of the means (SEM). Green triangles indicate significant differences ($P < 0.05$ after Bonferroni correction).

A. Cellular Localization B. Molecular Function C. Biological Process

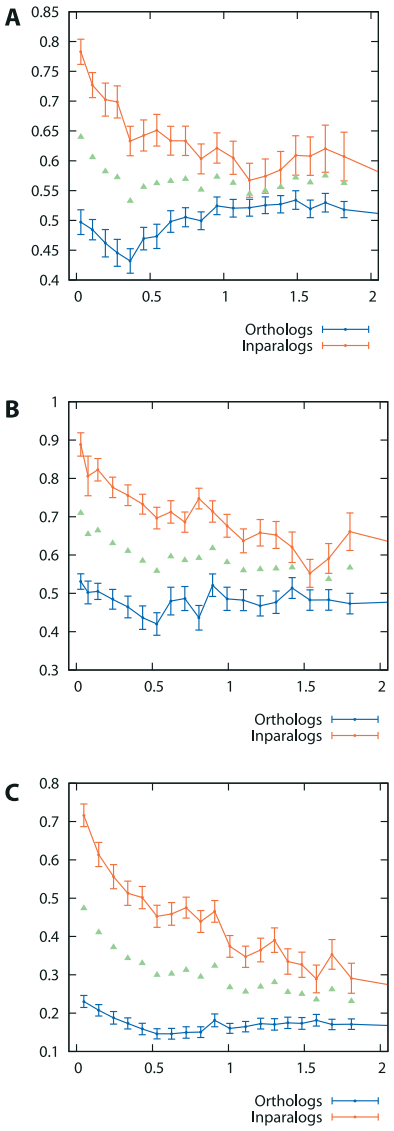

Figure 2

JTO conservation of orthologs versus inparalogs, relative to separation measured as expected number of amino acid substitutions per site. Results shown separately for the three Gene Ontology sub-ontologies. Error bars show standard error of the means (SEM). Green triangles indicate significant differences ($P < 0.05$ after Bonferroni correction).

A. Cellular Localization B. Molecular Function C. Biological Process

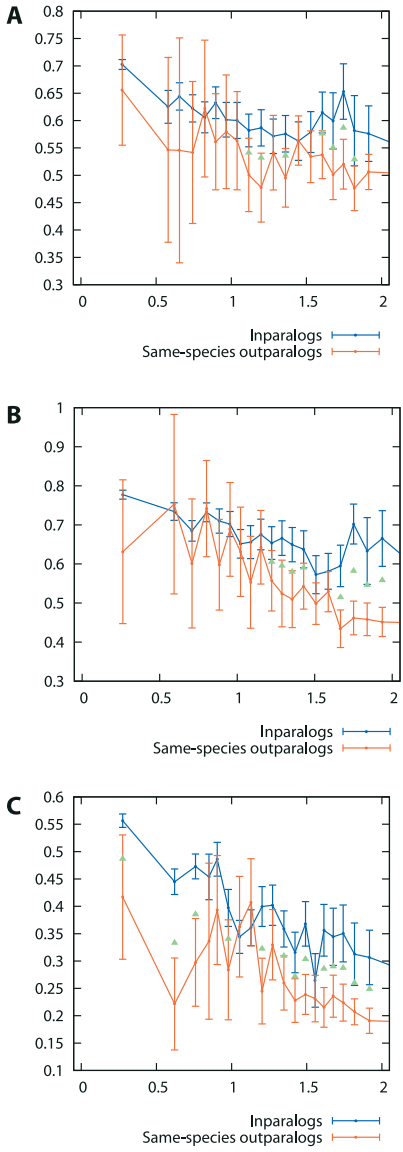

Figure 3

JTO conservation of inparalogs versus same-species outparalogs, relative to separation measured as expected number of amino acid substitutions per site. Error bars show standard error of the means (SEM). Green triangles indicate significant differences ($P < 0.05$ after Bonferroni correction, under a permutation test for mean JTO within each bin).

A. Cellular Localization B. Molecular Function C. Biological Process

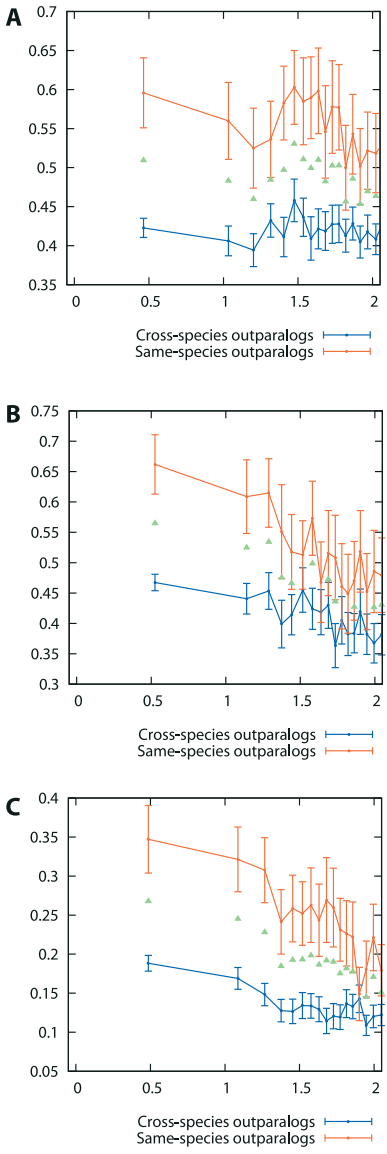

Figure 4

JTO conservation of cross-species versus same-species outparalogs, relative to separation measured as expected number of amino acid substitutions per site. Error bars show standard error of the means (SEM). Green triangles indicate significant differences ($P < 0.05$ after Bonferroni correction, under a permutation test for mean JTO within each bin).

A. Cellular Localization B. Molecular Function C. Biological Process

4 Conclusions

From the research that has been described in this thesis, by myself and by others, is there anything that can conclusively be said about the relationship between orthology, protein domain architecture, and protein function? Specifically, can anything be said regarding the validity of the orthology conjecture and the domain grammar hypothesis?

From my work in Paper II, as well as from the subsequent independent validation in SUPERFAMILY [182], it is clear that protein domain architecture, whether as the presence of single domains or as the presence of particular combinations of domains, provide functional information and can be used to predict at least some classes of protein function. As such, the weak form of the domain grammar hypothesis certainly holds.

During this work, I have discovered several factors that complicate validation of the strong form of the domain grammar hypothesis. Most particularly, the problem remains that homologs conserved in function may independently be conserved in domain architecture.

There is however some indirect validation of the strong form of the domain grammar hypothesis as well, from the results of Paper VI. Assuming that the shortcomings described in my evaluation of the connection between orthology and function are related to biases in the Gene Ontology only and not to the use of a protein sequence based measure for evolutionary time, Paper VI supports increased domain architecture conservation in orthologs after a certain time has passed. Since selective pressures ultimately depend on the effects a gene product has on phenotype fitness, such increased conservation is most likely a consequence of some form of functional conservation, perhaps one insufficiently represented by current functional annotation efforts. In comparison with the Gene Ontology, domain assignments are relatively complete and free of bias, so it is not inconceivable that they might reveal patterns otherwise not visible. As such, our demonstration of increased domain architecture conservation in orthologs imply that the orthology conjecture is indeed valid for some aspects of protein function, and that conservation of those aspects does depend on domain architecture conservation. As such, it implies that the

strong form of the domain grammar hypothesis is true for at least some classes of functions and domain architectures.

One problem with the above interpretation lies in the fact that the study in Paper V did not yield results for gene structure that matched those reported in Paper VI for domain architecture. That is, while Paper V demonstrated higher relative conservation of intron positions for orthologs than for cross-species paralogs, it also demonstrated higher relative conservation for inparalogs than for same-species paralogs. Unpublished follow-up analyses showed that this latter trend remained even with the same sequence divergence measure and significance testing strategy as used in Paper VI and the functional conservation in orthologs analysis. However, intron structure information is less complete, and more prone to bias, than domain architecture assignments, which may mean that effects similar to those affecting the latter analysis also affect the data used in Paper V. Furthermore, we very recently found that an error was made when constructing the data set for Paper V, such that many protein pairs appeared to have much more divergent intron structures than they truly have. As such, those results also appear too uncertain for me to be able to draw any conclusions either way from them regarding the orthology conjecture.

We were not able to directly show the validity of the orthology conjecture, which matches results recently reported by others [3]. However, we were able to identify anomalies in the framework used to test this hypothesis, and from these, we conclude that it is not possible to confirm or reject the orthology conjecture from presently available data.

Overall, then, I was partly successful in my efforts. I have shown that the weak form of the domain grammar hypothesis holds, and that there are indications that the orthology conjecture and the strong form of the domain grammar hypothesis hold, at least for some classes of proteins. I have also identified obstacles that further research into these questions must surmount to be able to come up with a more conclusive answer.

5 Possible directions of future work

Concerning the issues of potential biases in existing protein function annotation resources toward more similar annotations for inparalogs and other same-species protein pairs, one approach might be to use functional annotation that is entirely high-throughput in origin, systematically generated in the same fashion for all genes in a genome. Do such datasets exist? While there are whole-genome knockout studies, there may still exist a species bias in that different model organism biologists might tend to focus on different aspects, and, as such, might describe phenotypes in a species-specific manner. This source of bias should not favour inparalogs over same-species outparalogs, however, and it may be possible to normalize functional similarity measures with respect to it, producing an unbiased dataset. However, since the relationship between sequence divergence and conservation of other properties may be highly non-linear, such normalization will not be trivial.

It would be highly desirable to validate our results from Paper VI under different settings – changing the underlying domain definition scheme, changing the species set, changing the measures used to estimate evolutionary divergence, and changing the orthology inference strategy. The easiest way to do this might be to exploit tree-based orthology resources where branch lengths are available. Most crucially, the general reliance in most orthology resources on protein sequence divergence as a proxy for evolutionary time is problematic, since it is potentially strongly affected by the very selection processes we seek to analyze. Moving towards relatively more neutral measures of synonymous substitution distances is thus desirable, though it requires an underlying data set with clear mappings between nucleotide sequences and their encoded proteins, which are not always stored in the protein databases, and which suffer from poor synchronization between databases such as version differences.

One way to evaluate the domain grammar hypothesis might be to investigate how the degree of functional variation differs within groups of same-architecture proteins that are monophyletic and polyphyletic, respectively. That is, do architectures that have reliably originated many times through convergent evolution exhibit greater variation of function than architectures

that all stem from a common same-architecture ancestor? In all likelihood, the degree of this variation will differ between architectures and functional classes.

There may exist a set of architecture-function correspondences that are equally valid for monophyletic and polyphyletic architectures, and these would form evidence for the strong form of the domain grammar hypothesis being valid at least for that class of architectures. However, concluding this requires us to be able to rule out insufficient sampling of examples, and as such, a large dataset of reliable cases of convergent evolution of domain architectures, coupled with reliable functional annotation neither directly nor indirectly inferred from domain architecture, would be needed. At present, no such data set is available. Available parsimony-based methods for domain architecture reconstruction are most likely too uncertain to construct one. A likelihood-based framework may however be able to, and if applied to high-quality large-scale gene trees, spanning genomes with extensive and unbiased experimental functional characterization, it may be possible to do so. For this purpose, I have also begun work on a probabilistic method for domain architecture reconstruction along a gene or domain phylogeny. An additional potential problem lies in how to compensate for differences in divergence time in an analysis of this type, as it may not be well-defined for pairs of proteins that convergently evolved the same architecture.

Another approach, though costly and time-consuming, might be directed experiments, constructing synthetic proteins from domains taken from various sources and evaluating their function experimentally. Its value would lie in direct evaluation on the utility of domain architecture as a guide to designing protein function, perhaps the central selling point of the domain grammar hypothesis.

6 Sammanfattning på svenska

Ett centralt problem i den nya biologin är att förstå vilken roll varje protein spelar i den organism där det återfinns. Att avgöra det genom laboratorieexperiment är i många fall inte möjligt av etiska, ekonomiska eller tekniska skäl. Därför har man utvecklat metoder för att försöka förutsäga vad ett protein egentligen gör med hjälp av databehandling, statistik och databaser över redan känd information. De flesta metoder för att förutsäga ett proteins funktion bygger på att hitta andra, tillräckligt liknande proteiner vars funktion redan är känd. Det här projektet har ytterst syftat till att förbättra dessa metoder.

Ett fall där likhet mellan proteiner kan antas medföra att de antagligen också gör ungefär samma sak är när de har likartad molekylär struktur. Proteiners funktion bygger på vilka kemiska egenskaper de har, något som följer just av deras struktur. Det är känt att proteiner är sammansatta av återkommande strukturella byggstenar kallade *domäner*, som finns i många olika proteiner i olika kombinationer. Proteiner med likartad uppsättning domäner borde alltså också vara funktionellt lika, så att man kan dra slutsatser utifrån dem om motsvarande geners funktion. Ingen har dock testat systematiskt hur mycket vi egentligen kan säga om ett proteins funktion enbart utifrån dess domäner, eller ifall sambandet är direkt eller indirekt.

En annan viktig faktor är geners eller motsvarande proteiners släktskap med varandra. Många gener är varandras homologer, så att de stammar från samma ur-gen i någon nu utdöd art. De proteiner de kodar för kan därmed inte vara mer olika funktionellt än vad som medges av hur mycket vart och ett av dem har förändrats under de tusentals generationer som gått sedan de gick skilda vägar. Om den ursprungliga genen kopierades till två gener, som båda blev kvar inom samma art, sägs homologerna vara *paraloger* till varandra. Det finns då skäl att tro att deras respektive proteiners funktioner snabbare skall börja skilja sig åt, eftersom organismen inte behöver fler än ett av dem. Gener vars ur-gen skildes åt genom att en art blev till flera arter är istället *ortologa* med varandra. Om de blir alltför olika kommer ur-genens roll inte längre att fyllas av någon gen alls, så att en individ med gener förändrade på det sättet dör eller inte får lika stor avkomma. På så vis borde det naturliga urvalet se till att proteiner som är ortologa med varandra förblir funktionellt lika över längre tid än proteiner som är paraloga med varandra.

Detta evolutionära scenario ligger till grund för flera metoder för att förutsäga proteinfunktion, men hur väl det stämmer har inte heller testats uttömmande.

Inom bioinformatiken finns alltså en hel del verktyg för att förutsäga proteiners funktion, endera från vilka andra proteiner som är deras ortologer eller från deras strukturella domäner. Vår bild av både evolution och proteinkemi gör antagandena rimliga, men de har inte testats systematiskt, och har ibland ifrågasatts. Eftersom det är både teoretiskt och praktiskt viktigt att ha svar i de här frågorna har jag på olika sätt försökt utvärdera relationen mellan proteinfunktion, domänuppsättning, och huruvida ett protein utvecklats som ortolog eller som paralog.

Andra problem har behövt lösas längs vägen. I den första artikeln i avhandlingen tog vi fram en metod för att avgöra hur domänarkitekturen hos en grupp proteiner troligast har utvecklats över tid. Vi var särskilt intresserade av hur ofta domänarkitekturer hade uppstått flera gånger i olika släktlinjer, oberoende av varandra. Vi kom fram till att detta visserligen var ovanligt, men vanligare än vad tidigare studier kommit fram till.

I den andra artikeln undersökte jag sambandet mellan domänarkitektur och funktion mer direkt. Jag satte upp flera möjliga modeller för hur dessa båda fenomen kunde vara sammankopplade, och undersökte sedan hur väl modellerna kunde återskapa vad som redan är känt om proteiners funktion. Det visade sig att väldigt mycket av vår kunskap om proteiners funktion kan återskapas bara från kunskap om deras domänuppsättning, och dessutom att vissa funktioner bara kan förutsägas från kombinationer av domäner, inte från enskilda domäner. Därmed kan jag säga att domänuppsättningen helt tydligt är användbar för att förutsäga proteiners funktion. Däremot säger resultaten inget om huruvida sambandet är direkt eller indirekt. Till exempel skulle ortologa proteiner kunna vara mer lika både funktionellt sett och i sin domänarkitektur, som två oberoende effekter av den ortologa relationen.

För att kunna undersöka den här möjligheten, och kanske utesluta felkällan, behövde jag undersöka hur ortolog respektive paralog evolution påverkar domänarkitektur och funktion. Jag deltog i ett projekt vid Stockholms Bioinformatikcentrum för att förbättra en metod för att avgöra om gener och motsvarande proteiner är ortologa eller inte, kallat InParanoid. Den tidigare versionen av InParanoid gav felaktiga resultat för vissa arter, på grund av egenheter i deras proteinsekvenser. Genom att använda en del av resultaten från avhandlingens första artikel kunde jag pröva olika metoder för att komma runt den här felkällan. Resultaten från den studien ingår som den tredje artikeln i avhandlingen. Utifrån de resultaten kunde vi förbättra InParanoid och ge ut en ny, mycket mer omfattande version, vilket beskrivs i

avhandlingens fjärde artikel. Därmed kunde jag börja utvärdera vilken roll ortolog respektive paralog evolution egentligen spelar.

Gener i komplicerade organismer är uppdelade i mindre fragment, som kallas *exoner* och *introner*. Efter transkription klipps intronerna bort, och exonerna klistras ihop till en sekvens som sedan translateras till motsvarande protein. Genom att styra vilka exoner som skall användas och vilka som skall ignoreras, bland annat baserat på kontrollsekvenser i intronerna, kan cellen skapa många olika varianter av ett protein från samma gen. Den femte artikeln i avhandlingen beskriver vår utvärdering av frågan om de här möjliga kontrollsekvenserna var bättre bevarade mellan ortologa gener än mellan paraloga gener. Genom att använda InParanoid och kunskap om intron-exon-struktur i olika arter kunde vi visa att intronernas läge i genen verkar förändras snabbare över tid mellan ortologer än mellan paraloger, vilket är vad vi förväntade oss från teorin. Andra delar av resultatet var däremot sådana att de varken kan förklaras med standardmodellen eller dess nollhypotes, något som antyder att det kan finnas dolda brister i de data och metoder vi använt.

Den viktigaste delstudien var att se hur evolutionen av domänarkitektur beror på ortologi respektive paralogi. Vi använde återigen InParanoid, nu på många fler arter, och kunde se hur paralogers domänarkitektur förändras snabbare över tid än ortologers, resultat som vi beskriver i den sjätte artikeln i avhandlingen. Vår tolkning av de här resultaten är att det visar både hur funktion bevaras bättre i proteiner som inte har en kopia i samma art och hur den större graden av bevarad funktion visar sig genom bättre bevarad domänarkitektur. Det antyder i sin tur att sambandet mellan funktion och domänarkitektur inte enbart är indirekt, även om det är svårt att säga något om dess exakta omfattning.

Slutligen ville jag undersöka hur funktion bevaras eller förändras mellan proteiner beroende på om de är ortologer eller paraloger till varandra. Metoden jag använde var nära besläktad med den som användes i de två sista artiklarna. På samma sätt som i den femte artikeln visade sig resultaten varken vara förenliga med den testade modellen eller dess antites. I stället menar jag att de tyder på brister eller vinklingar i hur databaserna för proteiners funktion har tagits fram, så att det i nuläget inte går att använda data och metoder av dessa slag för att utvärdera kopplingen mellan proteiners funktion och ortologi. Jag diskuterar också möjliga sätt att komma runt de här bristerna, så att man i framtiden tydligare kan pröva antagandet att funktion bevaras bättre mellan ortologer än mellan paraloger.

Sammantaget har jag alltså velat utvärdera några viktiga modeller för hur proteiners funktion och struktur förändras av evolutionen över tid, och vad

resultaten av de utvärderingarna har för betydelse för vår möjlighet att förutsäga funktion där den annars är okänd. Jag har delvis lyckats med detta, genom att visa att domänarkitektur låter oss förutsäga funktion samt hur en indirekt observerad högre grad av bevarad funktion mellan ortologa proteiner visar sig i en högre grad av bevarad domänarkitektur. Jag har inte kunnat direkt påvisa någon högre grad av bevarad funktion mellan ortologa proteiner, och jag har hittat möjliga felkällor som måste hanteras innan någon studie helt kan utröna sanningen.

7 Acknowledgements

There are too many people that I want to acknowledge to remember. Thus, consider this a random sampling, in no particular order. Persons that by rights should be in multiple paragraphs are listed only once, and any omissions are wholly unintended.

My parents, Annika Åhs-Forslund and Bengt-Åke Forslund, literally for everything. Likewise thanks to the rest of my (super-)family; my amazing sisters, their partners and to other members of the Åhs and Forslund clans, for continual and unconditional support and inspiration. I would also like to thank Andreas Isberg and Kerstin Fredlund for artistic and practical support surrounding the production of this thesis.

My supervisors – thanks to Erik Sonnhammer for taking me in and allowing me to pursue the research directions I wanted as well as for being accepting of my eccentricities. I have felt that we have been united in the pursuit of a common goal, and been able to accomplish a lot in that. Also thanks for providing many opportunities for me to learn new things and to test my skills; manuscript reviews, conferences and summits, and responsibilities for various tools or resources. Thanks to Cristina al-Khalili Szigyarto for valuable discussions and feedback, and for providing additional perspective on the research community and on potential career paths. Also thanks to my previous supervisors – Vincent Moulton, Daniel Huson and Mats Gustafsson – for giving me a view early on regarding what the scientific process is like; it made me decide that I wanted to stay.

My friends – Project Nyaga members, Nietzsche Day and Crowleymass celebrants, and past denizens of Härnösand and of Flogsta, Uppsala: Leo Correia de Verdier, Eli Gladnikoff, Andreas “Inky” Bäckström, Henrik Kjölstad, Karl Arvinder, Daniel Visén, Sara Lilja-Visén, Mattias Göransson, Dionisis Granas, Cecilia Halling (and Tigris), Anders Nilsson, Julia Holm, Magnus Johansson, Lisa Hägglund, Marija Kramar, Kim Dahlin, Mathias “Tias” Lans, and all the others. Thank you for brightening the non-work side of my life, and providing occasional inspiration for the work part. Thanks to Lars-Anders Carlson for inspiration, guidance and valuable feedback; I look forward already to our next mountain hike. Particular thanks also to Olof

Kjölstad, not only for politics, music and horror scenarios, but also for considerable help with graphics and layout during the making of this thesis.

The Sonnhammer group, for providing assistance, feedback and support despite my increasingly asocial and neurotic tendencies as the project grew closer to its end. Thanks to Dave Messina for helping me navigate the real world in general and the US in particular, to Gabriel Östlund for many enjoying discussion of how to apply optimization techniques; I would have wanted to hear much more, and am grateful at how you have coped with my overstressed demeanor. Thanks also to Thomas Schmitt, Oliver Frings and Dimitri Guala Fernandovich for help in various practical or impractical situations. The systems administrators, Erik Sjölund and Romans Valls Guimerà have also been of tremendous help on many occasions.

My apologies to Stockholm Södra Kendo and Sensei Gilles Lahoundere for my many absences during the completion of this work. Masakatsu agatsu!

Thanks to the AlbaNova Voices choir, and to the AlumniX board; in the overlap between these, particularly to Jesper Gantelius for many amazing discussions and explorations. Thanks also to Malin Sandström for wisdom and perspective, and Åsa Innings for her cynical yet cheerful demeanor.

Thanks to my direct collaborators: Anna Henricson, Volker Hollich, Isabella Pekkari, Nattaphon “Joe” Thanintorn, Fabian Schreiber (as well as for pointing out crucial literature to me), Andreas Tjärnberg (ever patient) and Torbjörn Nordling, and also to those students I have had the pleasure to co-supervise: Alessia Peviani, Emil Kölbeck, Dan Berglund and Ganapathi Varma.

To various colleagues and collaborators: the Arrhenius, House 16 and SciLifeLab denizens, and the crazy crew of Cambridge and Janelia Farm for enjoyable meetings. Thanks particularly to Rob Finn and John Tate for infinite patience, Sean Eddy and Alex Bateman for valuable discussion, Pawel Herman for helping me find ways to determine statistical significance and for valuable feedback, Lars Arvestad and Jens Lagergren for providing phylogenetic insight, Hossein Farahani for his enlightened and empowering perspective, and Gunnar von Heijne for advice and support. Thanks also to Erik Aurell and Karin Julenius for providing inspiration.

To Elin – thank you for existing, and for providing a relief from the solipsist bubble of my mind. The fact that someone understands means the world. Let us change the world. Break the world’s shell!

8 References

[1] O'Brien KP, Westerlund I, Sonnhammer EL. OrthoDisease: a database of human disease orthologs. Human Mutation 2004;24:112-9.

[2] Forslund K, Schreiber F, Thanintorn N, Sonnhammer EL. OrthoDisease: tracking disease gene orthologs across 100 species. Briefings in Bioinformatics 2011; in press.

[3] Nehrt NL, Clark WT, Radivojac P, Hahn MW. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Computational Biology 2011;7:e1002073.

[4] Studer R, Robinson-Rechavi M. How confident can we be that orthologs are similar, but paralogs differ? Trends in Genetics 2009;25:210-6.

[5] Dessailly BH, Redfern OC, Cuff A, Orengo C. Exploiting structural classifications for function prediction: towards a domain grammar for protein function. Current Opinion in Structural Biology 2009;19:349-56.

[6] Owen R. Lectures on the comparative anatomy and physiology of the invertebrate animals. Longman, Brown, Green, & Longmans, London 1843.

[7] Darwin C. The Origin of Species by Means of Natural Selection. John Murry, London, 1859.

[8] Huxley THH. The Origin of Species. Westminster Reviews 1860;17:541-70.

[9] Patterson C. Homology in classical and molecular biology. Molecular Biology and Evolution 1988;5:603-25.

[10] Fitch WM. Distinguishing Homologous from Analogous Proteins. Systematic Zoology 1970;19:99-113.

[11] Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO Journal 1986;5:823-6.

[12] Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. PNAS 1990;87:2264-8.

[13] Mitrophanov Y. Statistical significance in biological sequence analysis. Briefings in Bioinformatics 2006;7:2-24.

[14] Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology 1981;147:195-7.

[15] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 1970;48:443-53.

[16] Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science 1985;227:1435-41.

[17] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. PNAS 1988;85:2444-8.

[18] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology 1990;215:403-10.

[19] Altschul SF, Madden TL, Schäffer A, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 1997;25:3389-402.

[20] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 1994;22:4673-80.

[21] Larkin M, Blackshields G, Brown NP, Chenna R, McGettigan P, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23:2947-8.

[22] Lassmann T, Sonnhammer EL. Kalign--an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics 2005;6:298.

[23] Lassmann T, Frings O, Sonnhammer EL. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Research 2009;37:858-65.

[24] Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 2002;30:3059-66.

[25] Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research 2005;33:511-8.

[26] Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics 2008;9:286-98.

[27] Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004;5:113.

[28] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 2004;32:1792-7.

[29] Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 2001;17:700-12.

[30] Goldenberg O, Erez E, Nimrod G, Ben-Tal N. The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. Nucleic Acids Research 2009;37:323-7.

[31] Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. PNAS 1987;84:4355-8.

[32] Schäffer A, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Research 2001;29:2994-3005.

[33] Biegert A, Söding J. Sequence context-specific profiles for homology searching. PNAS 2009;106:3770-5.

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

[35] Eddy SR. Profile hidden Markov models. Imagine 1998;755-763.

[36] Eddy SR. A new generation of homology search tools based on probabilistic inference. International Conference on Genome Informatics 2009;23:205-11.

[37] Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research 2011;39 Supplement 2:29-37.

[38] Koski LB, Golding GB. The closest BLAST hit is often not the nearest neighbor. Journal of Molecular Evolution 2001;52:540-2.

[39] Tatusov RL, Galperin MY, Natale D, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 2000;28:33-6.

[40] Tatusov RL, Natale D, Garkavtsev IV, Tatusova T, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Research 2001;29:22-8.

[41] Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale D. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 2003;4:41.

[42] Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research 2003;13:2178-89.

[43] Alexeyenko A, Tamas I, Liu G, Sonnhammer EL. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 2006;22:9-15.

[44] Karlin S, Altschul SF. Applications and statistics for multiple high-scoring segments in molecular sequences. PNAS 1993;90:5873-7.

[45] Altschul SF. A protein alignment scoring system sensitive at all evolutionary distances. Journal of Molecular Evolution 1993;36:290-300.

[46] Wall DP, Fraser HB, Hirsh E. Detecting putative orthologs. Bioinformatics 2003;19:1710-1.

[47] Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. Computers & Chemistry 1993;17:149-63.

[48] Shin SW, Kim SM. A new algorithm for detecting low-complexity regions in protein sequences. Bioinformatics 2005;21:160-70.

[49] Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, Schäffer A, Yu Y-K. Protein database searches using compositionally adjusted substitution matrices. The FEBS journal 2005;272:5101-9.

[50] Yu Y-K, Wootton JC, Altschul SF. The compositional adjustment of amino acid substitution matrices. PNAS 2003;100:15688-93.

[51] Yu Y-K, Gertz EM, Agarwala R, Schäffer A, Altschul SF. Retrieval accuracy, statistical significance and compositional similarity in protein sequence database searches. Nucleic Acids Research 2006;34:5966-73.

[52] Yu Y-K, Altschul SF. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. Bioinformatics 2005;21:902-11.

[53] Camin JH, Sokal RR. A Method for Deducing Branching Sequences in Phylogeny. Evolution 1965;19:311-26.

[54] Fitch WM. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Systematic Zoology 1971;20:406-6.

[55] Sankoff D, Morel C, Cedergren RJ. Evolution of 5S RNA and the non-randomness of base replacement. Nature New Biology 1973;245:232-4.

[56] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 1981;17:368-76.

[57] Makarova KS, Sorokin AV, Novichkov PS, Wolf YI, Koonin EV. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. Biology Direct 2007;2:33.

[58] Wang M, Caetano-Anollés G. Global phylogeny determined by the combination of protein domains in proteomes. Molecular Biology and Evolution 2006;23:2444-54.

[59] Page RD, Charleston M. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Molecular Phylogenetics and Evolution 1997;7:231-40.

[60] Daubin V, Gouy M, Perrière G. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Research 2002;12:1080-90.

[61] Grünewald S, Forslund K, Dress A, Moulton V. QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets. Molecular Biology and Evolution 2007;24:532-8.

[62] Dutilh BE, Snel B, Ettema TJG, Huynen M. Signature genes as a phylogenomic tool. Molecular Biology and Evolution 2008;25:1659-67.

[63] Wu M, Eisen J. A simple, fast, and accurate method of phylogenomic inference. Genome Biology 2008;9:151.

[64] Kupczok A, Schmidt H, von Haeseler A. Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. Algorithms for Molecular Biology 2010;5:37.

[65] Stark M, Berger S, Stamatakis A, von Mering C. MLTreeMap--accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. BMC Genomics 2010;11:461.

[66] Fitch WM. Homology: A personal view on some of the problems. Trends in Genetics 2000;16:227-31.

[67] Syvanen M. Horizontal gene transfer: evidence and possible consequences. Annual Review of Genetics 1994;28:237-61.

[68] Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification. Annual Reviews in Microbiology 2001;55:709-42.

[69] Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nature Reviews Microbiology 2005;3:711-21.

[70] Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics 2008;9:605-18.

[71] Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. Nature Reviews Microbiology 2005;3:679-87.

[72] Choi I-G, Kim S-H. Global extent of horizontal gene transfer. PNAS 2007;104:4489-94.

[73] Doolittle WF. Phylogenetic Classification and the Universal Tree. Science 1999;284:2124-8.

[74] Kurland CG, Canback B, Berg OG. Horizontal gene transfer: a critical view. PNAS 2003;100:9658-62.

[75] Koonin EV, Puigbò P, Wolf YI. Comparison of phylogenetic trees and search for a central trend in the "forest of life". Journal of Computational Biology 2011;18:917-24.

[76] Snel B, Bork P, Huynen M. Genome evolution. Gene fusion versus gene fission. Trends in Genetics 2000;16:9-11.

[77] Patthy L. Genome evolution and the evolution of exon-shuffling--a review. Gene 1999;238:103-14.

[78] Patthy L. Modular assembly of genes and the evolution of new functions. Genetica 2003;118:217-31.

[79] Roy SW. Recent evidence for the exon theory of genes. Genetica 2003;118:251-66.

[80] Chan CX, Darling AE, Beiko RG, Ragan M. Are protein domains modules of lateral genetic transfer? PloS One 2009;4:e4524.

[81] Song N, Sedgewick RD, Durand D. Domain architecture comparison for multidomain homology identification. Journal of Computational Biology 2007;14:496-516.

[82] Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Research 2008;36:250-4.

[83] Dutilh BE, van Noort V, van der Heijden RTJM, Boekhout T, Snel B, Huynen M. Assessment of phylogenomic and orthology approaches for phylogenetic inference. Bioinformatics 2007;23:815-24.

[84] Dufayard J-F, Duret L, Penel S, Gouy M, Rechenmann F, Perrière G. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics 2005;21:2596-603.

[85] Akerborg O, Sennblad B, Arvestad L, Lagergren J. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. PNAS 2009;106:5714-9.

[86] Zuckerkandl E, Pauling L. In Bryson V and Vogel HJ (eds), Evolving Genes and Proteins. Academic Press, New York 1965;97-166.

[87] Sonnhammer EL, Koonin EV. Orthology, paralogy and proposed classification for paralog subtypes. Trends in Genetics 2002;18:619-20.

[88] Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annual Reviews in Genetics 2005;39:309-38.

[89] Storm CEV, Sonnhammer EL. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Genomics 2002;18:92-9.

[90] Hulsen T, Huynen M, de Vlieg J, Groenen PM. Benchmarking ortholog identification methods using functional genomics data. Genome Biology 2006;7:31.

[91] Marron M, Swenson KM, Moret BME. Genomic Distances under Deletions and Insertions. Theoretical Computer Science 2004;325:347-60.

[92] Fang G, Bhardwaj N, Robilotto R, Gerstein MB. Getting started in gene orthology and functional analysis. PLoS Computational Biology 2010;6:e1000703.

[93] Storm CEV, Sonnhammer EL. Comprehensive analysis of orthologous protein domains using the HOPS database. Genome Research 2003;13:2353-62.

[94] Gabaldón T. Large-scale assignment of orthology: back to phylogenetics? Genome Biology 2008;9:235.

[95] Kuzniar A, van Ham RCHJ, Pongor S, Leunissen JM. The quest for orthologs: finding the corresponding gene across genomes. Trends in Genetics 2008;24:539-51.

[96] Trachana K, Larsson TA, Powell S, Chen WH, Doerks T, Muller J, Bork P. Orthology prediction methods: A quality assessment using curated protein families. Bioessays 2011.

[97] Rivera MC, Jain R, Moore JE, Lake J. Genomic evidence for two functionally distinct gene classes. PNAS 1998;95:6239-44.

[98] Zheng XH, Lu F, Wang Z-Y, Zhong F, Hoover J, Mural R. Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. Bioinformatics 2005;21:703-10.

[99] Jun J, Mandoiu II, Nelson CE. Identification of mammalian orthologs using local synteny. BMC Genomics 2009;10:630.

[100] Hachiya T, Osana Y, Popendorf K, Sakakibara Y. Accurate identification of orthologous segments among multiple genomes. Bioinformatics 2009;25:853-60.

[101] Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, Powell S, von Mering C, Doerks T, Jensen LJ, Bork P. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Research 2010;38:190-5.

[102] Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J, Holt I, Liang F, Quackenbush J. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). Genome Research 2002;12:493-502.

[103] Hubbard TJP, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E. Ensembl 2007. Nucleic Acids Research 2007;35:610-7.

[104] Vilella AJ, Severin J, Ureta-vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Research 2009;327-35.

[105] Remm M, Storm CEV, Sonnhammer EL. Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. Journal of Molecular Biology 2001;314:1041-52.

[106] O'Brien KP, Remm M, Sonnhammer EL. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Research 2005;33:476-80.

[107] Berglund A-C, Sjölund E, Ostlund G, Sonnhammer EL. InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Research 2008;36:263-6.

[108] Wheeler DL, Barrett T, Benson D, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova T, Wagner L, Yaschenko E. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 2007;35:5-12.

[109] Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Computational Biology 2009;5:e1000262.

[110] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 2000;28:27-30.

[111] Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Research 2002;30:42-6.

[112] Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Research 2004;32:277-80.

[113] Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Research 2006;34:354-7.

[114] Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. KEGG for linking genomes to life and the environment. Nucleic Acids Research 2008;36:480-4.

[115] Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Research 2010;38:355-60.

[116] Pryszcz LP, Huerta-Cepas J, Gabaldón T. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Research 2011;39:e32.

[117] Roth ACJ, Gonnet GH, Dessimoz C. Algorithm of OMA for large-scale orthology inference. BMC Bioinformatics 2008;9:518.

[118] Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Research 2011;39:289-94.

[119] Datta RS, Meacham C, Samad B, Neyer C, Sjölander K. Berkeley PHOG: PhyloFacts orthology group prediction web server. Nucleic Acids Research 2009;37:84-9.

[120] Dessimoz C, Boeckmann B, Roth ACJ, Gonnet GH. Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. Nucleic Acids Research 2006;34:3309-16.

[121] van Dongen S. Graph clustering by flow simulation. PhD Thesis, University of Utrecht, The Netherlands 2000.

[122] Chen F, Mackey AJ, Stoeckert CJ, Roos DS. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Research 2006;34:363-8.

[123] Dehal PS, Boore JL. A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. BMC Bioinformatics 2006;7:201.

[124] Goodstadt L, Ponting CP. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Computational Biology 2006;2:e133.

[125] Deluca TF, Wu I-H, Pu J, Monaghan T, Peshkin L, Singh S, Wall DP. Roundup: a multi-genome repository of orthologs and evolutionary distances. Bioinformatics 2006;22:2044-6.

[126] Li H, Coghlan A, Ruan J, Coin LJ, Hériché J-K, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK-S, Zheng W, Dehal P, Wang J, Durbin R.

TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Research 2006;34:572-80.

[127] Ruan J, Li H, Chen Z, Coghlan A, Coin LJM, Guo Y, Hériché J-K, Hu Y, Kristiansen K, Li R, Liu T, Moses A, Qin J, Vang S, Vilella AJ, Ureta-Vidal A, Bolund L, Wang J, Durbin R. TreeFam: 2008 Update. Nucleic Acids Research 2008;36:735-40.

[128] Zmasek CM, Eddy SR. A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics 2001;17:821-8.

[129] Zmasek CM, Eddy SR. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics 2002;3:14.

[130] Alexeyenko A, Lindberg J, Pérez-Bercoff Å, Sonnhammer EL. Overview and comparison of ortholog databases. Drug Discovery Today: Technologies 2006;3:137-143.

[131] Schmitt T, Messina DN, Schreiber F, Sonnhammer EL. SeqXML and OrthoXML: standards for sequence and orthology information. Briefings in Bioinformatics 2011;1-4.

[132] Sonnhammer EL, Kahn D. Modular arrangement of proteins as inferred from analysis of homology. Protein Science 1994;3:482-92.

[133] Doolittle RF. The multiplicity of domains in proteins. Annual Reviews in Biochemistry 1995;64:287-314.

[134] Sonnhammer EL. Protein family databases for automated protein domain identification. Database 1998;9:68-78.

[135] Apic G, Gough J, Teichmann S. An insight into domain combinations. Bioinformatics 2001;17 Supplement 1:S83-9.

[136] Han J-H, Batey S, Nickson A, Teichmann S, Clarke J. The folding and evolution of multidomain proteins. Nature Reviews Molecular Cell Biology 2007;8:319-30.

[137] Bru C, Courcelle E, Carrère S, Beausse Y, Dalmar S, Kahn D. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Research 2005;33:212-5.

[138] Heger A, Wilton CA, Sivakumar A, Holm L. ADDA: a domain database with global coverage of the protein universe. Nucleic Acids Research 2005;33:188-91.

[139] Tordai H, Nagy A, Farkas K, Bányai L, Patthy L. Modules, multidomain proteins and organismic complexity. The FEBS journal 2005;272:5064-78.

[140] Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann S. Structure, function and evolution of multidomain proteins. Current Opinion in Structural Biology 2004;14:208-16.

[141] Schmidt EE, Davies CJ. The origins of polypeptide domains. Bioessays 2007;29:262-70.

[142] Buljan M, Bateman A. The evolution of protein domain families. Biochemical Society Transactions 2009;37:751-5.

[143] Weiner J, Beaussart F, Bornberg-Bauer E. Domain deletions and substitutions in the modular protein evolution. The FEBS Journal 2006;273:2037-47.

[144] Moore AD, Björklund AK, Ekman D, Bornberg-Bauer E, Elofsson A. Arrangements in the modular evolution of proteins. Trends in Biochemical Sciences 2008;33:444-51.

[145] Ponting CP, Russell RR. The natural history of protein domains. Annual Reviews in Biophysics 2002;31:45-71.

[146] Ekman D, Björklund AK, Frey-Skött J, Elofsson A. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. Journal of Molecular Biology 2005;348:231-43.

[147] Winstanley HF, Abeln S, Deane CM. How old is your fold? Bioinformatics 2005;21 Supplement 1:449-58.

[148] Vogel C, Chothia C. Protein family expansions and biological complexity. PLoS Computational Biology 2006;2:e48.

[149] Ranea JG, Sillero A, Thornton JM, Orengo C. Protein superfamily evolution and the last universal common ancestor (LUCA). Journal of Molecular Evolution 2006;63:513-25.

[150] Itoh M, Nacher JC, Kuma K, Goto S, Kanehisa M. Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. Genome Biology 2007;8:121.

[151] Orengo CA, Pearl FM, Bray JE, Todd E, Martin a C, Lo Conte L, Thornton JM. The CATH Database provides insights into protein structure/function relationships. Nucleic Acids Research 1999;27:275-9.

[152] Orengo CA, Bray JE, Buchan DWA, Harrison A, Lee D, Pearl FMG, Sillitoe I, Todd AE, Thornton JM. The CATH protein family database: a resource for structural and functional annotation of genomes. Structure 2002;11-21.

[153] Pearl FM, Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, Sillitoe I, Thornton J, Orengo CA. The CATH database: an extended protein family resource for structural and functional genomics. Nucleic Acids Research 2003;31:452-455.

[154] Pearl FM, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Research 2005;33:247-51.

[155] Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo C. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Research 2007;35:291-7.

[156] Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J, Orengo C. The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Research 2009;37:310-4.

[157] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Research 2000;28:235-42.

[158] Buchan DWA, Shepherd AJ, Lee D, Pearl FMG, Rison SCG, Thornton JM, Orengo CA. Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. Genome Research 2002;12:503-14.

[159] Yeats C, Lees J, Reid A, Kellam P, Martin N, Liu X, Orengo C. Gene3D: comprehensive structural and functional annotation of genomes. Nucleic Acids Research 2008;36:414-8.

[160] Lees J, Yeats C, Redfern O, Clegg A, Orengo C. Gene3D: merging structure and function for a Thousand genomes. Nucleic Acids Research 2010;38:296-300.

[161] Yeats C, Lees J, Carter P, Sillitoe I, Orengo C. The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences. Nucleic Acids Research 2011;39:546-50.

[162] Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert C, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Tasneem A, Thanki N, Yamashita R, Zhang D, Zhang N, Bryant SH. CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Research 2009;37:205-10.

[163] Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita R, Zhang D, Zhang N, Zheng C, Bryant SH. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research 2011;39:225-9.

[164] McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research 2004;32:20-5.

[165] Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J, Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJ, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. New developments in the InterPro database. Nucleic Acids Research 2007;35:224-8.

[166] Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. InterPro: the integrative protein signature database. Nucleic Acids Research 2009;37:211-5.

[167] Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 1997;28:405-20.

[168] Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Research 1998;26:320-2.

[169] Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Research 1999;27:260-2.

[170] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Research 2002;30:276-80.

[171] Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Research 2004;32:138-41.

[172] Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Research 2006;34:247-51.

[173] Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz H-R, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. The Pfam protein families database. Nucleic Acids Research 2008;36:281-8.

[174] Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A. The Pfam protein families database. Nucleic Acids Research 2010;38:211-22.

[175] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 1995;247:536-40.

[176] Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin a G, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Research 2000;28:257-9.

[177] Shakhnovich BE, Max Harvey J. Quantifying structure-function uncertainty: a graph theoretical exploration into the origins and limitations of protein annotation. Journal of Molecular Biology 2004 Apr 2;337:933-49.

[178] Gough J, Chothia C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Research 2002;30:268-72.

[179] Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J. The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Research 2004;32:235-9.

[180] Wilson D, Madera M, Vogel C, Chothia C, Gough J. The SUPERFAMILY database in 2007: families and functions. Nucleic Acids Research 2007;35:308-13.

[181] Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J. SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Research 2009;37:380-6.

[182] de Lima Morais DA, Fang H, Rackham OJ, Wilson D, Pethica R, Chothia C, Gough J. SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Research 2011;39:427-34.

[183] Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. PNAS 1998;95:5857-64.

[184] Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Research 2000;28:231-4.

[185] Ponting CP, Schultz J, Milpetz F, Bork P. SMART: identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Research 1999;27:229-32.

[186] Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P. Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Research 2002;30:242-4.

[187] Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P. SMART 4.0: towards genomic data integration. Nucleic Acids Research 2004;32:142-4.

[188] Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P. SMART 5: domains in the context of genomes and networks. Nucleic Acids Research 2006;34:257-60.

[189] Letunic I, Doerks T, Bork P. SMART 6: recent updates and new developments. Nucleic Acids Research 2009;37:229-32.

[190] Kimura M. Fixation of a deleterious allele at oxie of two "duplicate" loci by mutation pressure and random drift. Genetics 1979;76:2858-2861.

[191] Orozco-Mosqueda MC, Altamirano-Hernandez J, Farias-Rodriguez R, Valencia-Cantero E, Santoyo G. Homologous recombination and dynamics of rhizobial genomes. Research in Microbiology 2009;160:733-41.

[192] Heyer W-D, Ehmsen KT, Liu J. Regulation of homologous recombination in eukaryotes. Annual Reviews in Genetics 2010;44:113-39.

[193] van Rijk A, Bloemendal H. Molecular mechanisms of exon shuffling: illegitimate recombination. Genetica 2003;118:245-9.

[194] Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nature Reviews Genetics 2009;10:691-703.

[195] Gogvadze E, Buzdin A. Retroelements and their impact on genome evolution and functioning. CMLS 2009;66:3727-42.

[196] Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annual Reviews in Genetics 2007;41:331-68.

[197] Babushok DV, Ostertag EM, Kazazian HH. Current topics in genome evolution: molecular mechanisms of new gene formation. CMLS 2007;64:542-54.

[198] Weiner J, Bornberg-Bauer E. Evolution of circular permutations in multidomain proteins. Molecular Biology and Evolution 2006;23:734-43.

[199] Vogel C, Morea V. Duplication, divergence and formation of novel protein topologies. Bioessays 2006;28:973-8.

[200] Ekman D, Björklund AK, Elofsson A. Quantification of the elevated rate of domain rearrangements in metazoa. Journal of Molecular Biology 2007;372:1337-48.

[201] Fong JH, Geer LY, Panchenko AR, Bryant SH. Modeling the evolution of protein domain architectures using maximum parsimony. Journal of Molecular Biology 2007;366:307-15.

[202] Farris JS. Phylogenetic Analysis Under Dollo's Law. Systematic Zoology 1977;26:77-88.

[203] Przytycka T, Davis G, Song N, Durand D. Graph theoretical insights into evolution of multidomain proteins. Journal of Computational Biology 2006;13:351-63.

[204] Björklund AK, Ekman D, Light S, Frey-Skött J, Elofsson A. Domain rearrangements in protein evolution. Journal of Molecular Biology 2005;353:911-23.

[205] Forslund K, Sonnhammer EL. Evolution of protein domain architectures. In Anisimova M (ed), Evolutionary Genomics: statistical and computational methods. Springer-Humana 2011; in press.

[206] Huynen M, van Nimwegen E. The frequency distribution of gene family sizes in complete genomes. Molecular Biology and Evolution 1998;15:583-9.

[207] Qian J, Luscombe NM, Gerstein M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. Journal of Molecular Biology 2001;313:673-81.

[208] Luscombe NM, Qian J, Zhang Z, Johnson T, Gerstein M. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome Biology 2002;3:r0040.

[209] Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV. Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evolutionary Biology 2002;2:18.

[210] Clauset A, Shalizi CR, Newman M. Power-law distributions in empirical data. SIAM Review 2009;51:661-703.

[211] Albert R, Jeong H, Barabasi AL. The diameter of the world wide web. Nature 1999; 401:130-1.

[212] Zipf GK. Human Behavior and the Principle of Least Effort. Addison-Wesley, Boston 1949.

[213] Newman M. Power laws, Pareto distributions and Zipf's law. Contemporary Physics 2005;46:323-51.

[214] Evlampiev K, Isambert H. Modeling protein network evolution under genome duplication and domain shuffling. BMC Systems Biology 2007;1:49.

[215] Fell D, Wagner A. The small world of metabolism. Nature Biotechnology 2000;18:1121-2.

[216] Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi A-L. The large-scale organization of metabolic networks. Nature 2000;651-654.

[217] Li W-H, Yang J, Gu X. Expression divergence between duplicate genes. Trends in Genetics 2005;21:602-7.

[218] Wuchty S. Scale-free behavior in protein domain networks. Molecular Biology and Evolution 2001;18:1694-702.

[219] Kuznetsov VA, Knott GD. Analysis of the evolving proteomes: Predictions of the number of protein domains in nature and the number of genes in eukaryotic organisms Journal of Biological Systems 2002; 10:381-407.

[220] Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature 2002;420:218-23.

[221] Erdös P, Rényi A. The Evolution of Random Graphs. Magyar Tud. Akad. Mat. Kutató Int. Közl. 1950; 5:17-61.

[222] Barabási A, Albert R. Emergence of Scaling in Random Networks. Science 1999;286:509-12.

[223] Apic G, Gough J, Teichmann S. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. Journal of Molecular Biology 2001;310:311-25.

[224] Apic G, Huber W, Teichmann S. Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. Journal of Structural and Functional Genomics 2003;4:67-78.

[225] Vogel C, Berzuini C, Bashton M, Gough J, Teichmann S. Supra-domains: evolutionary units larger than single protein domains. Journal of Molecular Biology 2004;336:809-23.

[226] Yanai I, Camacho CJ, DeLisi C. Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Physical Review Letters 2000;85:2641-4.

[227] van Nimwegen E. Scaling laws in the functional content of genomes. Trends in Genetics 2003;19:479-84.

[228] Ranea JG, Buchan DW, Thornton JM, Orengo C. Evolution of protein superfamilies and bacterial genome size. Journal of Molecular Biology 2004;336:871-87.

[229] Kummerfeld SK, Teichmann S. Protein domain organisation: adding order. BMC Bioinformatics 2009;10:39.

[230] Marcotte EM, Pellegrini M, Ng H, Rice DW, Yeates TO, David Eisenberg D. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. Science 1999;285:751-3.

[231] Basu MK, Poliakov E, Rogozin IB. Domain mobility in proteins: functional and evolutionary implications. Briefings in Bioinformatics 2009;10:205-16.

[232] Vogel C, Teichmann S, Pereira-Leal J. The relationship between domain duplication and recombination. Journal of Molecular Biology 2005;346:355-65.

[233] Weiner J, Moore AD, Bornberg-Bauer E. Just how versatile are domains? BMC Evolutionary Biology 2008;8:285.

[234] Basu MK, Carmel L, Rogozin IB, Koonin EV. Evolution of protein domain promiscuity in eukaryotes. Genome Research 2008;18:449-61.

[235] Gough J. Convergent evolution of domain architectures (is rare). Bioinformatics 2005;21:1464-71.

[236] Bashton M, Chothia C. The geometry of domain combination in proteins. Journal of Molecular Biology 2002;315:927-39.

[237] Geer LY, Domrachev M, Lipman DJ, Bryant SH. CDART: Protein Homology by Domain Architecture. Genome Research 2002;1619-23.

[238] Lin K, Zhu L, Zhang D-Y. An initial strategy for comparing proteins at the domain architecture level. Bioinformatics 2006;22:2081-6.

[239] Lee B, Lee D. DAhunter: a web-based server that identifies homologous proteins by comparing domain architecture. Nucleic Acids Research 2008;36:60-4.

[240] Lee B, Lee D. Protein comparison at the domain architecture level. BMC Bioinformatics 2009;10 Supplement 1:S5.

[241] Fraser AG, Marcotte EM. A probabilistic view of gene function. Nature Genetics 2004;36:559-64.

[242] Bard JBL, Rhee SY. Ontologies in biology: design, applications and future challenges. Nature Reviews Genetics 2004;5:213-22.

[243] Tipton K, Boyce S. History of the enzyme nomenclature system. Bioinformatics 2000;16:34-40.

[244] Mewes HW, Hani J, Pfeiffer F, Frishman D. MIPS: a database for protein sequences and complete genomes. Nucleic Acids Research 1998;26:33-7.

[245] Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, Lemcke K, Mannhaupt G, Pfeiffer F, Schüller C, Stocker S, Weil B. MIPS: a database for genomes and protein sequences. Nucleic Acids Research 2000;28:37-40.

[246] Mewes HW, Frishman D, Güldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Münsterkötter M, Rudd S, Weil B. MIPS: a database for genomes and protein sequences. Nucleic Acids Research 2002;30:31-4.

[247] Mewes HW, Amid C, Arnold R, Frishman D, Güldener U, Mannhaupt G, Münsterkötter M, Pagel P, Strack N, Stümpflen V, Warfsmann J, Ruepp A. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Research 2004;32:41-4.

[248] Mewes HW, Frishman D, Mayer KF, Münsterkötter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stümpflen V. MIPS: analysis and annotation of proteins from whole genomes in 2005. Nucleic Acids Res. 2006;34:D169-72.

[249] Mewes HW, Dietmann S, Frishman D, Gregory R, Mannhaupt G, Mayer KFX, Münsterkötter M, Ruepp A, Spannagl M, Stümpflen V, Rattei T. MIPS: analysis and annotation of genome information in 2007. Nucleic Acids Research 2008;36:196-201.

[250] Mewes HW, Ruepp A, Theis F, Rattei T, Walter M, Frishman D, Suhre K, Spannagl M, Mayer KFX, Stümpflen V, Antonov A. MIPS: curated databases and comprehensive secondary data resources in 2010. Nucleic Acids Research 2011;39:220-4.

[251] Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Güldener U, Mannhaupt G, Münsterkötter M, Mewes HW. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Research 2004;32:5539-45.

[252] Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genetics 2000;25:25-9.

[253] Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. Genome Research 2001;11:1425-33.

[254] Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Research 2010;38:331-5.

[255] Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. Nature Reviews Genetics 2008;9:509-15.

[256] UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Research 2010;38:142-8.

[257] Gilbert W. Why genes in pieces? Nature. 1978;271:501.

[258] Stajich JE, Dietrich FS, Roy SW. Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. Genome Biology 2007;8:223.

[259] Zambelli F, Pavesi G, Gissi C, Horner DS, Pesole G. Assessment of orthologous splicing isoforms in human and mouse orthologous genes. BMC Genomics 2010;11:534.

[260] Neverov AD, Artamonova II, Nurtdinov RN, Frishman D, Gelfand MS, Mironov A. Alternative splicing and protein function. BMC Bioinformatics 2005;6:266.

[261] Kim E, Goren A, Ast G. Alternative splicing: current perspectives. Bioessays 2008;30:38-47.

[262] Ho M-R, Jang W-J, Chen C, Ch'ang L-Y, Lin W. Designating eukaryotic orthology via processed transcription units. Nucleic Acids Research 2008;36:3436-42.

[263] Liu M, Grigoriev A. Protein domains correlate strongly with exons in multiple eukaryotic genomes--evidence of exon shuffling? Trends in Genetics 2004;20:399-403.

[264] Roy SW, Fedorov A, Gilbert W. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. PNAS 2003;100:7158-62.

[265] Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. Conservation versus parallel gains in intron evolution. Nucleic Acids Research 2005;33:1741-8.

[266] Carmel L, Rogozin IB, Wolf YI, Koonin EV. Patterns of intron gain and conservation in eukaryotic genes. BMC Evolutionary Biology 2007;7:192.

[267] Nurtdinov RN, Artamonova II, Mironov AA, Gelfand MS. Low conservation of alternative splicing patterns in the human and mouse genomes. Human Molecular Genetics 2003;12:1313-20.

[268] Nurtdinov RN, Mironov A, Gelfand MS. Rodent-specific alternative exons are more frequent in rapidly evolving genes and in paralogs. BMC Evolutionary Biology 2009;9:142.

[269] Wang W, Zheng H, Yang S, Yu H, Li J, Jiang H, Su J, Yang L, Zhang J, McDermott J, Samudrala R, Wang J, Yang H, Yu J, Kristiansen K, Wong GK, Wang J. Origin and evolution of new exons in rodents. Genome Research 2005;15:1258-64.

[270] Carmel L, Wolf YI, Rogozin IB, Koonin EV. Three distinct modes of intron dynamics in the evolution of eukaryotes. Genome Research 2007;1034-44.

[271] Mistry M, Pavlidis P. Gene Ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics 2008;9:327.

[272] Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Computational Biology 2009;5:e1000443.

[273] Lord PW, Stevens RD, Brass A, Goble C. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 2003;19:1275-1283.

[274] Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics 2007;23:1274-81.

[275] Pesquita C, Faria D, Bastos H, Ferreira AEN, Falcão AO, Couto FM. Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 2008;9 Supplement 5:S4.

[276] Jiang JJ, Conrath DW. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. Computational Linguistics 1997.

[277] Sheehan B, Quigley A, Gaudin B, Dobson S. A relation based measure of semantic similarity for Gene Ontology annotations. BMC Bioinformatics 2008;9:468.

[278] Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J. Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. Nature 2000;408:325-30.

[279] Beutner EH. Immunofluorescent Staining: the Fluorescent Antibody Method. Bacteriological Reviews 1961;25:49-76.

[280] Phizicky EM, Fields S. Protein-protein interactions: methods for detection and analysis. Microbiological Reviews 1995;59:94-123.

[281] Coin L, Bateman A, Durbin R. Enhanced protein domain discovery by using language modeling techniques from speech recognition. PNAS 2003;100:4516-20.

[282] Hahn MW, Kern AD. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Molecular Biology and Evolution 2005;22:803-6.

[283] Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Molecular Systems Biology 2007;3:88.

[284] Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifier. Machine Learning 1997;29:131-163.

[285] Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. Protein molecular function prediction by Bayesian phylogenomics. PLoS Computational Biology 2005;1:e45.

[286] Vinayagam A, König R, Moormann J, Schubert F, Eils R, Glatting K-H, Suhai S. Applying Support Vector Machines for Gene Ontology based gene function prediction. BMC Bioinformatics 2004;5:116.

[287] Dobson PD, Doig AJ. Predicting enzyme class from protein structure without alignments. Journal of Molecular Biology 2005;345:187-99.

[288] Kretschmann E, Fleischmann W, Apweiler R. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. Genome 2001;17:920-6.

[289] Syed U, Yona G. Using a mixture of probabilistic decision trees for direct prediction of protein function. In the proceedings of RECOMB 2003;224-234.

[290] Hayete B, Bienkowska JR. Gotrees: predicting GO associations from protein domain composition using decision trees. Pacific Symposium on Biocomputing 2005;10:127-138.

[291] Jones CE, Schwerdt J, Bretag TA, Baumann U, Brown AL. GOSLING: a rule-based protein annotator using BLAST and GO. Bioinformatics 2008;24:2628-9.

[292] Hawkins T, Chitale M, Luban S, Kihara D. PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins 2009;74:566-82.

[293] Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ. The PROSITE database. Nucleic Acids Research 2006;34:227-30.

[294] Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche B, de Castro E, Lachaize C, Langendijk-Genevaux PS, Sigrist CJ. The 20 years of PROSITE. Nucleic Acids Research 2008;36:245-9.

[295] Laskowski R, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Research 2005;33:89-93.

[296] Pal D, Eisenberg D. Inference of protein function from protein structure. Structure 2005;13:121-30.

[297] Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nature Reviews Molecular Cell Biology 2007;8:995-1005.

[298] Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Research 2004;32:129-33.

[299] Walker MG, Volkmuth W, Sprinzak E, Hodgson D, Klingler T. Prediction of Gene Function by Genome-Scale Expression Analysis: Prostate Cancer-Associated Genes. Genome Research 1999;9:1198-1203.

[300] Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. Science 2002;296:750-2.

[301] Daub CO, Sonnhammer EL. Employing conservation of co-expression to improve functional inference. BMC Systems Biology 2008;2:81.

[302] Gabaldón T, Huynen M. Prediction of protein function and pathways in the genome era. CMLS 2004;61:930-44.

[303] Mao F, Su Z, Olman V, Dam P, Liu Z, Xu Y. Mapping of orthologous genes in the context of biological pathways: An application of integer programming. PNAS 2006;103:129-34.

[304] Hulsen T, de Vlieg J, Groenen PM. PhyloPat: phylogenetic pattern analysis of eukaryotic genes. BMC Bioinformatics 2006;7:398.

[305] Snitkin ES, Gustafson AM, Mellor J, Wu J, DeLisi C. Comparative assessment of performance and genome dependence among phylogenetic profiling methods. BMC Bioinformatics 2006;7:420.

[306] Ranea JG, Yeats C, Grant A, Orengo C. Predicting protein function with hierarchical phylogenetic profiles: the Gene3D Phylo-Tuner method applied to eukaryotic genomes. PLoS Computational Biology 2007;3:e237.

[307] Cokus S, Mizutani S, Pellegrini M. An improved method for identifying functionally linked proteins using phylogenetic profiles. BMC Bioinformatics 2007;8 Supplement 4:S7.

[308] Kensche PR, van Noort V, Dutilh BE, Huynen M. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. Journal of the Royal Society 2008;5:151-70.

[309] Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science 2003;302:249-55.

[310] Huynen M, Snel B, Mering CV, Bork P. Function prediction and protein networks. Current Opinion in Cell Biology 2003;15:191-8.

[311] De Las Rivas J, Fontanillo C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Computational Biology 2010;6:e1000807.

[312] Alexeyenko A, Sonnhammer EL. Global networks of functional coupling in eukaryotes from comprehensive data integration. Genome Research 2009;19:1107-16.

[313] von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Research 2003;31:258-61.

[314] von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Research 2005;33:433-7.

[315] Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research 2009;37:412-6.

[316] Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Research 2011;39:561-8.

[317] Song J, Singh M. How and when should interactome-derived clusters be used to predict functional modules and protein function? Bioinformatics 2009;25:3143-50.

[318] Campillos M, von Mering C, Jensen LJ, Bork P. Identification and analysis of evolutionarily cohesive functional modules in protein networks. Genome Research 2006;16:374-82.

[319] Dutkowski J, Tiuryn J. Identification of functional modules from conserved ancestral protein-protein interactions. Bioinformatics 2007;23:149-58.

[320] Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. IEEE Computer Society Bioinformatics Conference 2002;1:197-206.

[321] Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. Nature Biotechnology 2003;21:697-700.

[322] Chua HN, Sung W-K, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics 2006;22:1623-30.

[323] Chua HN, Sung W-K, Wong L. Using indirect protein interactions for the prediction of Gene Ontology functions. BMC Bioinformatics 2007;8 Supplement 4:S8.

[324] Sun S, Zhao Y, Jiao Y, Yin Y, Cai L, Zhang Y, Lu H, Chen R, Bu D. Faster and more accurate global protein function assignment from protein interaction networks using the MFGO algorithm. FEBS letters 2006;580:1891-6.

[325] Zhu M, Gao L, Guo Z, Li Y, Wang D, Wang J, Wang C. Globally predicting protein functions based on co-expressed protein-protein interaction networks and ontology taxonomy similarities. Gene 2007;391:113-9.

[326] Jaeger S, Gaudan S, Leser U, Rebholz-Schuhmann D. Integrating protein-protein interactions and text mining for protein function prediction. BMC Bioinformatics 2008;9 Supplement 8:S2.

[327] Jaeger S, Sers CT, Leser U. Combining modularity, conservation, and interactions of proteins significantly increases precision and coverage of protein function prediction. BMC Genomics 2010;11:717.

[328] Eisen JA. Phylogenomics : Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. Genome Research 1998;163-7.

[329] Sjölander K. Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics 2004;20:170-9.

[330] Brown D, Sjölander K. Functional classification using phylogenomic inference. PLoS Computational Biology 2006;2:e77.

[331] Krishnamurthy N, Brown DP, Kirshner D, Sjölander K. PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. Genome Biology 2006;7:83.

[332] Friedberg I, Harder T, Godzik A. JAFA: a protein function annotation meta-server. Nucleic Acids Research 2006;34:379-81.

[333] Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L. The distributed annotation system. BMC Bioinformatics. 2001;2:7.

[334] Ohno S. Evolution by gene duplication. Springer-Verlag, New York 1970.

[335] Li WH. Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. Genetics 1980;95:237-58.

[336] Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 1999;151:1531-45.

[337] Stoltzfus A. On the possibility of constructive neutral evolution. Journal of Molecular Evolution 1999;49:169-81.

[338] Hahn MW. Distinguishing among evolutionary models for the maintenance of gene duplicates. The Journal of Heredity 2009;100:605-17.

[339] Zhang J. Evolution by gene duplication: an update. Trends in Ecology and Evolution 2003;18:292-8.

[340] Kondrashov F, Rogozin IB, Wolf YI, Koonin EV. Selection in the evolution of gene duplications. Genome Biology 2002;3:r0008.

[341] Kondrashov F, Kondrashov AS. Role of selection in fixation of gene duplications. Journal of Theoretical Biology 2006;239:141-51.

[342] Gibson T, Goldberg DS. Questioning the ubiquity of neofunctionalization. PLoS Computational Biology 2009;5:e1000252.

[343] Dolinski K, Botstein D. Orthology and functional conservation in eukaryotes. Annual Reviews in Genetics 2007;41:465-507.

[344] Wilson C, Kreychman J, Gerstein M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. Journal of Molecular Biology 2000;297:233-49.

[345] Devos D, Valencia A. Practical limits of function prediction. Proteins 2000;41:98-107.

[346] Todd a E, Orengo C, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. Journal of Molecular Biology 2001;307:1113-43.

[347] Rost B. Enzyme function less conserved than anticipated. Journal of Molecular Biology 2002;318:595-608.

[348] Tian W, Skolnick J. How Well is Enzyme Function Conserved as a Function of Pairwise Sequence Identity? Journal of Molecular Biology 2003;333:863-882.

[349] Sangar V, Blankenberg DJ, Altman N, Lesk AM. Quantitative sequence-function relationships in proteins based on gene ontology. BMC Bioinformatics 2007;8:294.

[350] Addou S, Rentzsch R, Lee D, Orengo C. Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. Journal of Molecular Biology 2009;387:416-30.

[351] Bromham L, Penny D. The modern molecular clock. Nature Reviews Genetics 2003;4:216-24.

[352] Li WH, Gojobori T. Rapid evolution of goat and sheep globin genes following gene duplication. Molecular Biology and Evolution 1983;1:94-108.

[353] Scannell DR, Wolfe KH. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome Research 2008;137-147.

[354] Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC, Sali A. Evolutionary constraints on structural similarity in orthologs and paralogs. Protein Science 2009;18:1306-15.

[355] Davis JC, Petrov D. Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biology 2004;2:e55.

[356] Gerlt J, Babbitt PC. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annual Reviews in Biochemistry 2001;70:209-46.

[357] Rost B. Twilight zone of protein sequence alignments. Protein Engineering 1999;12:85-94.

[358] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 1980;16:111-20.

[359] Takahata N, Kimura M. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics 1981;98:641-57.

[360] Wood TC, Pearson WR. Evolution of protein sequences and structures. Journal of Molecular Biology 1999;291:977-95.

[361] Schug J, Diskin S, Mazzarelli J, Brunk BP, Stoeckert CJ. Predicting gene ontology functions from ProDom and CDD protein domains. Genome Research 2002;12:648-55.

[362] Lopez D, Pazos F. Gene ontology functional annotations at the structural domain level. Proteins 2009;76:598-607.

[363] Hegyi H, Gerstein M. Annotation Transfer for Genomics: Measuring Functional Divergence in Multi-Domain Proteins Our Approach to Functional. Genome Research 2001;1632-40.

[364] Bashton M, Chothia C. The generation of new protein functions by the combination of domains. Structure 2007;15:85-99.

[365] Pegg SC-H, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC. Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. Biochemistry 2006;45:2545-55.

[366] Redfern OC, Dessailly B, Orengo C. Exploring the structure and function paradigm. Current Opinion in Structural Biology 2008;18:394-402.

[367] Mirny L, Gelfand MS. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. Journal of Molecular Biology 2002;321:7-20.

[368] Capra J, Singh M. Characterization and prediction of residues determining protein functional specificity. Bioinformatics 2008;24:1473-80.

[369] Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. PNAS 1996;93:13429-34.

[370] Andrieu C. An Introduction to MCMC for Machine Learning. Science 2003;5-43.

[371] Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 2007;23:1282-8.

[372] Chandonia J-M, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. Nucleic Acids Research 2004;32:189-92.

[373] Eichinger L, Pachebat J, Glöckner G, Rajandream M-a, Sucgang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q, Tunggal B, Kummerfeld S, Madera M, Konfortov B, Rivero F, Bankier a T, Lehmann R, Hamlin N, Davies R, Gaudet P, Fey P, Pilcher K, Chen G, Saunders D, Sodergren E, Davis P, Kerhornou A, Nie X, Hall N, Anjard C, Hemphill L, Bason N, Farbrother P, Desany B, Just E, Morio T, Rost R, Churcher C, Cooper J, Haydock S, van Driessche N, Cronin A, Goodhead I, Muzny D, Mourier T, Pain A, Lu M, Harper D, Lindsay R, Hauser H, James K, Quiles M, Madan Babu M, Saito T, Buchrieser C, Wardroper A, Felder M, Thangavelu M, Johnson D, Knights A, Loulseged H, Mungall K, Oliver K, Price C, Quail M, Urushihara H, Hernandez J, Rabbinowitsch E, Steffen D, Sanders M, Ma J, Kohara Y, Sharp S, Simmonds M, Spiegler S, Tivey A, Sugano S, White B, Walker D, Woodward J, Winckler T, Tanaka Y, Shaulsky G, Schleicher M, Weinstock G, Rosenthal A, Cox EC,

Chisholm RL, Gibbs R, Loomis WF, Platzer M, Kay RR, Williams J, Dear PH, Noegel A, Barrell B, Kuspa A. The genome of the social amoeba *Dictyostelium discoideum*. Nature 2005;435:43-57.

[374] Dehal PS, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biology 2005;3:e314.

[375] Gaudet P, Chisholm R, Berardini T, Dimmer E, Engel SR, Fey P, Hill DP, Howe D, Hu JC, Huntley R, Khodiyar VK, Kishore R, Li D, Lovering RC, McCarthy F, Ni L, Petri V, Siegele DA, Tweedie S, Van Auken K, Wood V, Basu S, Carbon S, Dolan M, Mungall CJ, Dolinski K, Thomas P, Ashburner M, Blake JA, Cherry JM, Lewis SE, Balakrishnan R, Christie KR, Costanzo MC, Deegan J, Diehl AD, Drabkin H, Fisk DG, Harris M, Hirschman JE, Hong EL, Ireland A, Lomax J, Nash RS, Park J, Sitnikov D, Skrzypek MS, Apweiler R, Bult C, Eppig J, Jacob H, Parkhill J, Rhee S, Ringwald M, Sternberg P, Talmud P, Twigger S, Westerfield M. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. PLoS Computational Biology 2009; 5:e1000431.

[376] Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A. BioMart--biological queries made easy. BMC Genomics 2009;10:22.

[377] Nichols TE, Holmes AP. Nonparametric Permutation Tests For Functional Neuroimaging: A Primer with Examples. Human Brain Mapping 2001;15:1-25.