Accepted for publication in *Naturwissenschaften*, Springer Verlag**.**
The original publication will be available at www.springerlink.com.

Homology and Phylogeny and their Automated Inference

Georg Fuellen
Bioinformatics Research Group
Ernst-Moritz-Arndt-University Greifswald
Institute for Mathematics and Computer Science
Jahnstr. 15a, 17487 Greifswald
Phone +49 3834 86 4618
Fax +49 3834 86 80086
fuellen@alum.mit.edu

Abstract. The analysis of the ever-increasing amount of biological and biomedical data can be pushed forward by comparing the data within and among species. For example, an integrative analysis of data from the genome sequencing projects for various species traces the evolution of the genomes and identifies conserved and innovative parts. Here I review the foundations and advantages of this “historical” approach and evaluate recent attempts at automating such analyses. Biological data is comparable if a common origin exists (homology), as is the case for members of a gene family originating via duplication of an ancestral gene. If the family has relatives in other species, we can assume that the ancestral gene was present in the ancestral species from which all the other species evolved. In particular, describing the relationships among the duplicated biological sequences found in the various species is often possible by a phylogeny, which is more informative than homology statements. Detecting and elaborating on common origins may answer *how* certain biological sequences developed, and predict *what sequences are in a particular species and what their function is*. Such knowledge transfer from sequences in one species to the homologous sequences of the other is based on the principle of ‘my closest relative looks and behaves like I do’, often referred to as ‘guilt by association’. To enable knowledge transfer on a large scale, several automated ‘phylogenomics pipelines’ have been developed in recent years, and seven of these will be described and compared. Overall, the

examples in this review demonstrate that homology and phylogeny analyses, done on a large (and automated) scale, can give insights into function in biology and biomedicine.

Keywords: homology search, comparative genomics, protein function, annotation pipelines, phylogenomics

Introduction and terminology. Homology is the relation of biological sequences by way of their common evolutionary origin (Fitch 1970). That is, there once was a piece of DNA, a gene, an interaction between proteins, etc. It was duplicated, and the duplicates evolved separately, gaining, for example, substitutions in sequence. The duplicates are called homologs, no matter how similar they are. Nevertheless, since usually we cannot look back in time, homology is an inference based on similarity. It is a pragmatic yes/no decision that can have an estimate of significance, or probability, attached to it. This estimate is usually based on the quantity of similarity. Thus, two genes can be said to have a high chance of being homologous, or, some *part* of the sequence of one gene can be homologous to another gene. However, two genes should not be called “highly homologous”. Terminology does not allow such a statement; sequences have a common origin, or they do not have one. More importantly, though, it must be recognized that homology is always something we know with limited certainty: Certainty cannot be established since the similarity that we can measure may be due to convergence, where two sequences of different origin become similar because they fulfill a common function. Or, similarity due to common ancestry may get lost in time and no longer be recognizable. Thus, with a few exceptions (using fossil data or evolution in the lab), homology is a concept that must be handled with pragmatism: We cannot be certain about homology, but estimates of homology can nevertheless be used as a foundation of meaningful analyses and valuable predictions, even if the term is misused by many, and the fundamental uncertainty of homology statements is neglected all too often.

The transfer of knowledge about inherited attributes from one homologous sequence to another is at the heart of comparative genomics and phylogenomics, and it will be exemplified in detail below. We will mostly deal with “function” or “functionality”. Defining these concepts precisely is difficult; neither the section “Definition of function” in Watson et al (2005) nor the section “What is function?” in Friedberg (2006) provide a clear-cut definition. A working definition sufficient for our purpose is that function is either a term taken from a controlled vocabulary of biological terms such as the Enzyme Commission (EC, www.hem.qmul.ac.uk/iubmb/enzyme/) classification scheme or the Gene Ontology Consortium (GO) scheme (2006, www.geneontology.org), or a term which is not yet part of such a controlled vocabulary, but which can be added to it by specializing an existing term.

We mentioned “duplicates” of a biological sequence; but we have to distinguish between two scenarios:

1) The standard duplication of a sequence within a single species, e.g. the appearance of two copies of a gene and their subsequent divergence. Whole-genome duplication, segmental duplication, tandem duplication, retrotransposition, and other processes may cause such a duplication.

2) The other common mechanism that brings two copies into existence is speciation, that is the “duplication” of the entire species hosting the gene. Glossing over the speciation process itself (which involves individuals in a population, giving rise to a wide array of complicating factors, see e.g. Maddison and Knowles, 2006), the result is that the gene is found (and usually continues to be found) in the two species, and in their subsequent descendants, and it diverges in these.

Standard duplication gives rise to *paralogs*, speciation gives rise to *orthologs* (Fitch 1970), and a history of duplication and speciation events gives rise to bewildering scenarios. Unfortunately, in this case confusion is all too often heightened by a misuse of terminology which suggests a certainty that does not exist: Two most similar genes found in two species are often called orthologs without any further justification. Even if they are each other’s

reciprocal closest relatives, they need not be orthologs. They may be two paralogs for which the “opposite number” in the other organism does not exist due to differential loss (Fig. _1), a phenomenon called “hidden paralogy” (Martin and Burg, 2002). Here, muddled terminology and an unwillingness to face uncertainty triggers, for example, the problem that the “orthologs” found by methods like Inparanoid (Remm et al, 2001), Orthostrapper (Storm and Sonnhammer, 2002) or OrthoMCL (Li et al, 2003) cannot be used to construct species phylogenies on the assumption of single common origin (see e.g. Theissen 2002). As described by Zmasek and Eddy (2002, their Fig.1), hidden paralogy can impair functional annotation, too: the duplication that is ignored may go together with a change in function.

A detailed discussion of terminology problems with respect to orthology and paralogy, and their interconnection with functional issues, can be found in Jensen (2001) and references therein.

The following text describes the path of the evolutionary analysis of gene/protein families, starting with homology search and alignment, followed by tree inference, and culminating in functional annotation. The paradigm of phylogenomics, which is the superiority of annotation based on trees over annotation based on homology search, is exemplified, and automated phylogenomics pipelines are described and compared in a tabular format.

Homology Search. Homology searches are at the heart of gene/protein family analysis, because they deliver the data to work with. The searches provide the evolutionary relatives of the gene/protein under study. As described above, any homology search can at most find *putative* homologs of a gene/protein in a set or a database of other genes/proteins. Thus, all occurrences of the term “homolog” in the following may be read as “putative homolog”. Using sequence data, similarity is used as a proxy to homology. Then, homology search becomes a string-matching exercise, where matching of similar characters (one from each string) is measured using a scoring matrix that relates the individual characters. Positional homology can be established by an alignment process introducing gaps so that overall, the

matching characters trigger a maximum sum-of-pairs similarity score (for a tutorial see Fuellen 1994).

Using more than one member of a gene/protein family as search input, homology search gains sophistication, finding matches that are closest in similarity to a set of strings. The corresponding methods used to search for protein homologs can be divided into profile-based approaches such as HMMSearch (Eddy 1998) and PSI-Blast (Altschul et al, 1997) and motif-based approaches such as PHI-Blast (Zhang et al, 1998) and MAST (Bailey and Gribskov, 1998). In Alam et al (2004) we added another class that is the combination of existing methods. After struggling with complicated approaches and formulas, we finally adhered to the rule of keeping the approach as simple as possible, combining methods using a simple formula. We were then able to outperform current methods by a good margin. More precisely, our CHASE method combines the ranked lists (Fig. _2) of hits (putative homologs) calculated by the component methods. For each hit, CHASE takes the weighted average of its significance values (to be precise, its E-Values) in each ranking. The new ranking of hits then follows from the weighted average obtained by the hits using the component methods. Two preprocessing steps were necessary, however, for successful combination: The significance values associated with a hit sequence from the database were found to be on a different scale depending on the method that produced the hit list, so they had to be rescaled to render them comparable. Second, the logarithm of the significance values was taken to avoid rounding problems. (A very similar approach to the integration of systems biology data, using P-values, and involving weighting, transformation and scaling, is described in Hwang et al, 2005.)

Validation of sequence-based protein homology search methods is tricky since a standard of truth cannot be obtained directly – there is no way to look back in time. As a proxy, relationships between proteins based on structural data such as SCOP (Murzin et al, 1995) can be taken (Rehmsmeier 2002). Thus, we evaluated in how far our method finds structure-based homologs of a protein family using sequence data of a set of related

proteins from the same family (or superfamily). Indeed we outperformed the component methods; in particular we were able to identify more true positives within the hits at the top of the list, as can be seen from the specificity/sensitivity plot (also known as ROC, receiver-operator-characteristic) of Fig. _3.

CHASE was developed in 2002/2003. It is possible that methods of homology search we did not consider (Spang et al, 2002; Ploetz and Fink, 2005; Kuang et al, 2005) will outperform the specificity / sensitivity we obtained. However, if these methods were incorporated into our scheme, we are confident that the new combination scheme will again be superior to each of its components, thus making it a timeless approach. An important question, which is only answered by anecdotes in the CHASE paper, is, “why does combination work”. Basically, each method has its outliers, and combination suppresses these (cf. Fig. 5 in Alam et al, 2004). However, a rigorous analysis still needs to be conducted to provide more insight. Then again, there exist a lot of combination approaches in bioinformatics, for protein structure prediction (for example, Cuff et al, 1998), protein function prediction not based on sequence homology (Kemmeren et al, 2005), transcription start site calculation (Bajic et al, 2004), gene modeling (Allen et al, 2005) and for data integration in systems biology in general (Hwang et al, 2005), but none of these papers seem to provide genuine insight into why combination is so successful.

Genomic Homology Search. The search for evolutionary relatives of a gene/protein should not be limited to protein database searches. Genomic databases containing nucleotide data of whole chromosomes or genomes may contain relatives that have not yet made it into the protein databases because the nucleotide data has not yet been analyzed, or because the gene was missed. Here, the piecemeal intron/exon structure of many eukaryotic genes adds a complicating factor, and a straightforward analysis of all six-frame translations of a genomic sequence is not enough. Instead, using gene modeling approaches we need to

predict the intron/exon structures of the genes in the genomic sequence. As we just saw, combinative approaches have also been developed for this task. In particular, Jigsaw (Allen et al, 2005) combines evidence from cDNA, transcript and gene data collected from databases and from ab-initio gene finders as well as from gene finders that use a sequence-conservation approach, and from phylogenetic analysis. The EnsEMBL pipeline can also be thought of as a combinative approach, see Curwen et al, 2004. Finally, FIGENIX (Gouret et al, 2005) also includes a combinative approach to genomic searches, before starting the phylogenomics pipeline that we discuss in more detail below. (See also Electronic Supplementary Material S1 for *Genomic Homology Search - a method combination approach.*)

Phylogenetic tree inference. A gene tree is calculated given the gene sequences as input, aligned for positional homology. The tree-shaped arrangement is then based on the similarity between the aligned sequences, evaluated position by position of the alignment. Some clever algorithms like maximum parsimony, Bayesian inference, maximum likelihood and neighbor joining (Felsenstein 2003) have been developed to calculate the tree by which the similarity / inheritance relationships among the sequences are best reflected. Most of them work for both nucleotide and amino acid sequences. Excellent reviews and textbooks exist on these topics (Thornton and DeSalle 2000, Felsenstein 2003). For very large amounts of sequence data, only distance-based methods such as neighbor-joining are fast enough to deliver a tree in reasonable time (see Mailund et al 2006 for a fast implementation). However, distance-based methods fail to fully consider the column-wise pattern of similarity provided by the positional homology of the sequence alignment. Instead, they perform pairwise sequence comparisons to calculate pairwise distances and only then they move on towards a multi-species analysis. In contrast, parsimony, likelihood and Bayesian approaches are time-consuming because they take each column of the multiple alignment into consideration, and, in turn, they very often they yield more plausible trees. Parallelized and speed-optimized versions of the Bayesian approach (MrBayes, Ronquist

and Huelsenbeck, 2003, Altekar at al, 2004) and of maximum likelihood (RaXML, Stamatakis 2006) are the best option if accurate trees are to be estimated from large amounts of data (e.g. up to several thousand average-length protein sequences). In any case, a tree contains more useful information than a tabular listing of gene sequence similarity. There are also methods that can calculate a network, instead of a tree, allowing for the representation of recombination events, gene conversion, horizontal gene transfer, hybridization, and/or simple uncertainty (Bandelt and Dress 1994, Bryant and Moulton 2004, Huson and Bryant 2006).

A combined gene/species tree is depicted in Fig. _4. (A similar evolutionary scenario was already used in Fig. _1 to describe “hidden paralogy”). Following the tree from the root to the leaves, a specific scenario of gene evolution by duplication and speciation can be read off the tree. At the root a duplication took place, and the red and the orange copy of the gene evolved without further duplication or loss; both genes went through the speciation events as indicated. Based on the red or the orange copy alone, the correct species tree can be inferred; all orthologs are present and there is no hidden paralogy: If sequences were lost as in Figure _1, a tree based on closeness of relationship would place together paralogs, in that case resulting in an incorrect species tree. If only a few sequences were lost, we can take note of the problem, and return that species tree that is concordant with the gene tree assuming a minimum of gene duplication and loss (Page 1998, Chen at al, 2000). To achieve this goal, we can use parsimony as an optimization criterion to find the species tree with which the given gene tree reconciles best. Such a reconciled gene tree illustrates all putative speciations, duplications and losses, and we can infer all orthology and paralogy relationships between genes. Since the history of duplications and speciations of a gene may be quite complex, rooting of gene trees is not straightforward; some ideas are presented in Chen et al (2000) and Stechmann and Cavalier-Smith (2002).

Knowledge transfer based on comparative genomics: Function prediction using phylogenies of GPCR and NR proteins. The following examples of the use of phylogenetic trees for the functional characterization of protein sequences illustrate some of the most important issues encountered in phylogenomics; cases of success are described as well as problems such as missing conservation, and convergence. An early paper by Fryxell (1996) reports success in correlating phylogenetic tree structure and functional annotation for paralogous G protein α chains, suggesting that "each pharmacological class of G_α genes share a single, ancient evolutionary origin", and convergence can be ruled out (Fryxell 1996). Later, G-protein-coupled receptors (GPCRs) have been studied intensively. Communi et al (2001) identify a novel GPCR, and their phylogenetic tree of paralogous GPCRs (figure 2 in their paper) shows that a protein of high affinity to ADP is its closest relative. Indeed, their experimental work confirms high ADP affinity of their novel GPCR. Joost and Methner (2002) suggest that their phylogenetic analysis of 277 human G- protein-coupled receptors is a "tool for the prediction of orphan receptor ligands". For example, their tree gives a valuable hint regarding the function of the GPR12 protein (Ignatov et al, 2003). Furthermore, Metpally and Sowdhamini (2005) describe a very exhaustive study of GPCRs, noting "unexpected levels of evolutionary conservation across human and Drosophila GPCRs". Many papers have been published studying the evolution of single amino acids (or small sets of amino acids) with respect to function (see Yao et al, 2003; Thornton and Kelley, 1998, for examples.)

Guilt by association (that is, phylogenetic closeness) does not always work. For example, Escriva et al (1997) study NRs (nuclear receptors; orthologs as well as paralogs) and they report that they found "no relationship between the position of a given liganded receptor in the tree and the chemical nature of its ligand". They propose that the various nuclear receptors "have gained the ability to bind their ligands independently and that the ancestral NR was an orphan receptor". However, homodimerization versus heterodimerization correlates with the different groups in the NR tree: Laudet (1997) proposes that the ability to heterodimerize evolved once, in a gene tree of orthologs and paralogs. Convergence is a

frequent explanation for failure of knowledge transfer based on phylogeny. For example, Kornegay et al (1994) describe a case of species-specific convergence for stomach lysozymes.

The question whether orthologs or (closest) paralogs are better suited for function prediction is debated, see below and Jensen (2001).

Homology search versus phylogenetic tree inference for functional annotation. Homology search can be used for functional annotation in two ways: The sequences found to be related can be used to predict attributes of the sequence(s) used for the search, and vice versa. For example, an uncharacterized sequence can be used as search input, and the hits give hints regarding functionality, if something is known about these. In turn, all proteins known to be encoded by a given genome can be put into a database, and homology searches with known proteins can be used to annotate this protein database. Such a knowledge transfer is done implicitly if data from KOG (eukaryotic clusters of orthologous groups / eukaryotic COG, Tatusov et al, 2003) or Pfam (Sonnhammer et al, 1998) are used for annotation of a new genome.

Just performing database searches, accuracy of functional annotation can be compromised, as discussed in Brown and Sjölander (2006) (earlier papers are Koski and Golding (2001), Devos and Valencia (2002), Galperin and Koonin (1998), Eisen (1998) and Eisen and Wu (2002)). In particular, the relationships between search input and hits, and the subsequent knowledge transfer, are devoid of the structure that is inherent to biological data, namely the tree-shaped or network-shaped relationship due to common evolutionary history. And indeed, it was shown that it is worth employing the fine-grained tree structure for homology search itself (Rehmsmeier and Vingron, 2001), and for the annotation or functional characterization of sequences. In particular, rate variation (possibly combined with

duplication and hidden paralogy) can trigger incorrect functional annotation by homology search alone (Eisen, 1998; but see Zmasek and Eddy, 2002, page 17). Of course, rate variation can also yield incorrect phylogenetic trees (see e.g. Philippe et al (2005) who discussed this issue quite recently). This effect is pronounced if fast distance-based tree inference methods are employed, because distances may be inflated by substitutions that occur exclusively in one sequence (so-called autapomorphies), or distances may be reduced artificially between the sequences that evolved slowly, triggered by the leftover unsubstituted character sites (so-called symplesiomorphies) which they share (Thornton and DeSalle 2000; Fuellen et al, 2001).

Automated pipelines for homology search, phylogenetic tree inference and functional annotation. As described, one major reason to do phylogenomics is the quest for more accurate functional annotations (Eisen 1998, Eisen and Wu, 2002). Naturally, phylogenomics is done on a large scale: we wish to annotate not just a single protein family, and we want to include as much data as possible to maximize accuracy of the analysis. A larger dataset not only improves chances that some sequences are annotated based on experiment. We can also assume that the more homologous sequences are included, the better the tree structure (Rannala et al, 1998). Large-scale analysis calls for automation, exemplified by the seven pipelines compared in Table _2. Automation started with the *pyphy* tool by Sicheritz-Ponten and Andersson (2001). They introduced crude tree structure schemata called „phylogenetic connections". Using these, for each gene in a genome the user of *pyphy* can then determine e.g. whether it features nearest neighbors only from the archaeal kingdom. Around the same time, Zmasek and Eddy (2002) developed RIO, Resampled Inference of Orthologs, with an emphasis on the estimation of orthology and paralogy given complex gene histories, including confidence values of the estimates. RIO is tightly connected to the Pfam database, restricting input options. Its output consists of lists of orthologs and paralogs; no phylogenetic tree is provided. Plewniak et al (2003) calculate no phylogenetic tree either, but they do provide a clustering of the sequences found to be related to the query. Their *PipeAlign* tool already automates retrieval of related sequences from databases, as well as the generation and curation of the multiple alignment. Frickey and Lupas (2004) describe an automated "phylome generation and analysis" tool called *PhyloGenie* that is inspired by *pyphy* and includes improvements on the generation and the post-processing of the

multiple alignments. These are not based on the full sequences; instead, the homologous regions are written underneath the query sequence, an approach called “stacking of high-scoring segment pairs (HSPs)”. Moreover, they maintain a database of the gene trees constructed and enable extraction of all phylogenies that match specific constraints on tree structure. Gouret et al. (2005) report an “intelligent automation of genomic annotation” called *FIGENIX*, which calculates gene trees, guided by an expert system. For each protein family, three different phylogeny reconstruction methods (neighbor joining, maximum parsimony, maximum likelihood) are used, and a consensus is calculated. ProteinUniverse (Brown and Sjolander 2006; Krishnamurthy et al, 2007) provides a suite of tools that taken together implement a phylogenomics pipeline. Special care is taken to deal with domain organization issues. Input options are very flexible, and a sophisticated functional analysis can be performed towards the end of the pipeline. Most recently, tree construction based on homology search output has been added to BLAST itself (Wheeler et al, 2007), by stacking of HSPs to provide the multiple alignment. Only two distance-based tree reconstruction methods are available (neighbor joining and a variant of minimum evolution, Fitch and Farris (1974)), and no confidence estimates (bootstrap values) are calculated. To a varying degree, all these pipelines attempt to automate four tasks: Collect useful sequence information, align it, generate a tree or a set of trees, and analyze the evolutionary information in some manner motivated by the biological question that was the starting point. In the next section, we will use the RiPE pipeline to exemplify these tasks, sometimes with reference to one of the other pipelines just described.

The RiPE pipeline for automated phylogenetic analysis. We designed a pipeline (Fuellen et al, 2005; Spitzer 2006) called Retrieval-induced phylogeny estimation (RiPE). RiPE automates phylogenomic analyses, in order to annotate a protein family as accurately as possible using as much information as possible, as summarized in Fig. _5. Collecting this information is **task 1**, so we conduct a homology search with a search profile (derived from the protein family) as query. Ideally, the query corresponds to what we call a “maximum unit of common evolutionary heritage”, that is a repeat-free concatenation of domains, which evolved together in the members of a protein family (Spitzer 2006, chapter 3). Optimally, we use a combination of homology searches as in CHASE, in as large a dataset as reasonable.

In the study reported in Fuellen et al (2005) we restricted ourselves to proteins known from completely sequenced genomes, and we used PSI-Blast (Altschul et al, 1997) for searching. The former restriction made it a bit easier to analyze results. We obtained a tree of 1138 sequences; a preliminary analysis using the NR database (restricted in size only by setting the number of bacteria and archaea to a representative subset) yielded an unmanageable tree of more than 4000 sequences. Moreover, searching in the proteomes of *completely* sequenced genomes, the analysis should not be impaired by missing (yet unsequenced) paralogs. PSI-Blast was used because CHASE was still in development.

We stack the homology search results (high-scoring segment pairs, HSPs) in the form of blockwise local pairwise alignments between the profile (the already aligned set of query sequences) and the homologous sequences from the database (cf. Figure _5). Thus, only the homologous parts of the homologous database sequences are retained, and the position-by-position homology as defined by the alignment is the result of the homology search itself. Thus, **task 2** is accomplished, namely the multiple alignment of the sequences, here defined by the positional homology assumed for each position in the alignment. Our approach focuses on the more reliable (less noisy) regions of the alignment, as suggested in Eisen (1998) and Sjölander (2003), and it is an alternative to alignment masking (Frickey and Lupas, 2004) that is the exclusion of alignment positions deemed unreliable in a post-processing step. As explained by Sjölander, such masking has the downside that functionally important regions outside of the conserved core may be neglected; this downside is avoided by our approach. Moreover, stacking is much faster than any true multiple alignment, so that we can analyze much larger data sets.

Task 3 is the phylogenetic tree reconstruction; we do not do anything special here, using standard software like neighbor joining to establish a tree-shaped fine-grained relationship among the homologous sequences. Despite their shortcomings already discussed, using fast

distance-based methods (such as neighbor-joining using Quickjoin, Mailund et al, 2006), makes it possible, even for very large datasets, to calculate confidence values for subtrees based on bootstrap re-sampling. Then again, for the neighbor-joining tree of the 1138 sequences we analyzed, we find subtree support of 0% for branches close to the root of the tree. This is to be expected (Thornton and DeSalle, 2000); the phenomenon is caused by sequences that “wander around” in the tree because they do not really belong to any of the subtrees that branch off close to the root (see also Thornton and DeSalle 2000, page 54). Nevertheless, the tree features large subtrees that correspond to the ABC subfamilies known from the literature, and it features many smaller subsubtrees with high bootstrap support that correspond to known subsubfamilies. The subtrees represent all known subfamilies and they contain almost exactly the sequences known to belong to these based on published inventories, with only 10 exceptions (out of 264 sequences classified in the literature), and 6 of these 10 exceptions are most likely an error in the literature (Fuellen et al, 2005, supplementary data).

Task 4 is the functional analysis of the sequences. For this task, we collected functional annotation for all sequences in the tree. This is unfortunately a highly manual task because the sources of experimental annotation information are often dispersed. For ABC transporters, the GO annotation (Gene Ontology Consortium, 2006) is insufficient to assign precise substrate specificity (transport capacity). Thus, we obtained precise substrate specificities from databases and from the literature. Given a gene tree with annotated and unannotated sequences as leaves, we then use the simple idea that knowledge transfer should be done from an annotated leaf to every leaf in the tree to which it is the closest annotated leaf. This idea is a variant of the ‘guilt by association’ principle; in a phylogenetic context, this association is common evolutionary history. The function transfer rule as defined in Fuellen et al (2005) and illustrated in Fig. _6 is a formalization of this idea. A closely related formalization used by the RIO pipeline is the definition of “subtree neighbors” of a sequence *s* (Zmasek and Eddy, 2002), denoting by default all other sequences that originate

from the grandparent *p* of *s* in the tree, no matter whether the path from *p* to *s,* and from *p* to the subtree neighbors, features duplication or speciation events. More generally, *p* may be the *k*-level parent of *s*, e.g. the great grandparent for *k=3*. The difference between this concept and the function transfer rule lies in the arbitrary threshold employed to define subtree neighbors: the level *k* must be fixed in advance. Zmasek and Eddy suggest function annotation transfer is best for proteins which are subtree neighbors, and at the same time deemed orthologous by their method. Alternatively, they suggest that superorthologs (no duplication in the path from the annotated to the unannotated protein) and ultraparalogs (no speciation in that path) are good candidates for function annotation transfer. As described, RIO provides bootstrap-based significance values for orthology, superorthology and subtree neighborhood, and report rankings based on orthology. They do not integrate their concepts in an automated way, yielding e.g. a combined ranking. A recent Bayesian approach to functional inference from gene trees is outlined in Engelhardt et al, 2005.

It has been put forward that orthology should be the single criterion for validity of function annotation transfer (Eisen, 1998). As described, Zmasek and Eddy (2002) suggest functional annotation transfer based on different criteria, and they point out problems if the ortholog is not also a subtree neighbor. Moreover, we may add that orthology assignments are often based on similarity, and hidden paralogy is possible. Jensen et al (2003) claim superiority of using orthologs based on phylogenetic profiling but not sequence homology, for predicting cellular function. They acknowledge that for 3D protein structures, there is no difference; paralogs as well as orthologs are conserved. In any case, closeness in the gene tree is a very plausible, if not the most plausible, justification for transferring an annotation (Thornton and DeSalle 2000, page 50), even if the presence of paralogs almost always hints at sub- or neofunctionization (Prince and Pickett 2002).

Our RiPE pipeline automatically performs tasks 1) to 3), and we used it to analyze the evolution of ABC proteins, with a focus on their function. As described in Fuellen et al (2005),

functional predictions based on phylogeny (summarized in Table _1 and Fig. _7) were superior to functional predictions based on a Blast search. Presumably, the tree put the sequences into a relationship structure that is more accurate than homology search; phylogeny reconstruction exploits the complex interplay of the position-by-position similarity data given by the alignment of the sequences. As described, our conclusion on the superiority of the phylogenetic approach is in line with many other publications.

We applied RiPE not just to ABC proteins, but also to FinGER proteins (Stolle et al, 2005), DNA-directed RNA polymerases (Klenk et al, 2004) as well as S100 proteins and tyrosine kinases (Spitzer, 2006). In particular, in Stolle et al (2005), we calculate a gene tree that divides the human FinGER proteins into 6 subfamilies, and we generate a plausible prediction of what the FinGER protein under study, FinGER-5 (also known as SMAP-5) may be doing. Based on its closest characterized relative, the yeast protein Yip1p, it may be part of the ER (Endoplasmatic Reticulum) to Golgi transport pathway. In case of Klenk et al (2004) our trees confirmed standard phylogenies based on RNA and protein data.

Criteria for comparing phylogenomics pipelines. Following up on Gouret et al (2005), we collected criteria in Table _2 that highlight different features of phylogenomics pipelines. Beyond such a tabular comparison of features, there is no straightforward way to compare or benchmark them. The pipelines were designed with different aims, and since we cannot look back in time, benchmarking the quality of phylogenetic trees they return is particularly difficult.

Auxiliary tools for homology search and phylogeny. See Electronic Supplementary Material S1 for *Auxiliary tools for homology search and phylogeny*.

Conclusions. Phylogenetic analysis of whole-genome data across organisms is still in its infancy. The pipelines currently available all cover just a small portion of an all-encompassing

evolutionary (or call it historical) analysis of genes and genomes. They are limited by the scope as well as the depth of the analysis. Ultimately, given biological data of all sorts from a large range of organisms, one would like to trace back the evolution of all data, how it started from a few ancestral precursors by ways of duplication and speciation, giving rise to the complexity of life that we observe. In other words, a generalization of phylogenetics to all levels of biological organization is needed (Serb and Oakley, 2005). For biological pathways, such a generalization is difficult but not impossible, given their relatively low level of evolutionary coherence (Gabaldón 2005). It is an open question how far back one can trace with acceptable certainty, given improvements in methods as well as in data availability, with a maximum range in species diversity (including fossil data) and data diversity. It is also an open question how much of this knowledge can be put to use to cure human disease, or, more generally, how much of it is helpful in applied research.

Acknowledgement. The author wishes to thank the following people for their feedback on (parts of) the manuscript: Etienne Danchin, Tancred Frickey, Iddo Friedberg, Philippe Gouret, Jake Gunn-Glanville, Claus Kerkhoff, Stefan Lorkowski, Frederic Plewniak, Michael Rebhan, Kimmen Sjölander, Michael Spitzer and Dion Whitehead.

References.

Alam I, Dress A, Rehmsmeier M, Fuellen G (2004) Comparative homology agreement search: an effective combination of homology-search methods. Proc Natl Acad Sci U S A 101: 13814-13819

Allen JE, Salzberg SL (2005) Jigsaw: integration of multiple sources of evidence for gene prediction. Bioinformatics 21: 3596-3603

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F (2004) Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20: 407-415

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402

Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. Bioinformatics 14: 48-54

Bajic VB, Tan SL, Suzuki Y, Sugano S (2004) Promoter prediction analysis on the whole human genome. Nat Biotechnol 22: 1467-1473

Bandelt HJ, Dress AW (1992) Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol Phylogenet Evol 1: 242-252

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL (2000) The Pfam protein families database. Nucleic Acids Res 28: 263-266

Brown D, Sjölander K (2006) Functional classification using phylogenomic inference.. PLoS Comput Biol 2: e77

Brown NP, Leroy C, Sander C (1998) MView: a web-compatible database search or multiple alignment viewer. Bioinformatics 14: 380-381

Bryant D, Moulton V (2004) Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21: 255-265

Chen K, Durand D, Farach-Colton M (2000) Notung: a program for dating gene duplications and optimizing gene family trees. J Comput Biol 7: 429-447

Communi D, Gonzalez NS, Detheux M, Brezillon S, Lannoy V, Parmentier M, Boeynaems JM (2001) Identification of a novel human ADP receptor coupled to G(i). J Biol Chem 276: 41479-41485

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) Jpred: a consensus secondary structure prediction server. Bioinformatics 14: 892-893

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M (2004) The Ensembl automatic gene annotation system. Genome Res 14: 942-950

Devos D, Valencia A (2001) Intrinsic errors in genome annotation. Trends Genet 17: 429-431

Eddy SR (1998) Profile hidden markov models. Bioinformatics 14: 755-763

Edgar RC (2004) Muscle: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113

Edgar RC, Sjölander K (2003) Satchmo: sequence alignment and tree construction using hidden markov models.. Bioinformatics 19: 1404-1411

Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8: 163-167

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

Engelhardt BE, Jordan MI, Muratore KE, Brenner SE (2005) Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol 1: e45

Escriva H, Safi R, Hanni C, Langlois MC, Saumitou-Laprade P, Stehelin D, Capron A, Pierce R, Laudet V (1997) Ligand binding was acquired during evolution of nuclear receptors. Proc Natl Acad Sci U S A 94: 6803-6808

Felsenstein J (2003) Inferring phylogenies. Sinauer, Sunderland MA, USA.

Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19: 99-113

Fitch WM, Farris JS (1974) Evolutionary trees with minimum nucleotide replacements from amino acid sequences.. J Mol Evol 3: 263-278

Frickey T, Lupas AN (2004) Phylogenie: automated phylome generation and analysis. Nucleic Acids Res 32: 5231-5238

Friedberg I (2006) Automated protein function prediction--the genomic challenge. Brief Bioinform 7: 225-242

Fryxell KJ (1996) The coevolution of gene family trees. Trends Genet 12: 364-369

Fuellen G (1994) A gentle guide to multiple alignment. Complexity International 4.

Fuellen G, Spitzer M, Cullen P, Lorkowski S (2005) Correspondence of function and phylogeny of ABC proteins based on an automated analysis of 20 model protein data sets. Proteins 61: 888-899

Fuellen G, Wagele JW, Giegerich R (2001) Minimum conflict: a divide-and-conquer approach to phylogeny estimation. Bioinformatics 17: 1168-1178

Gabaldón T (2005) Evolution of proteins and proteomes: a phylogenetics approach. Evolutionary Bioinformatics Online 1: 51-61

Galperin MY, Koonin EV (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. In Silico Biol 1: 55-67

Gene Ontology Consortium (2006) The gene ontology (GO) project in 2006. Nucleic Acids Res 34: D322-6

Gouret P, Vitiello V, Balandraud N, Gilles A, Pontarotti P, Danchin EGJ (2005) Figenix: intelligent automation of genomic annotation: expertise integration in a new software platform. BMC Bioinformatics 6: 198

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23: 254-267

Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, Weston AD, de Atauri P, Aitchison JD, Hood L, Siegel AF, Bolouri H (2005) A data integration methodology for systems biology. Proc Natl Acad Sci U S A 102: 17296-17301

Ignatov A, Lintzel J, Hermans-Borgmeyer I, Kreienkamp H, Joost P, Thomsen S, Methner A, Schaller HC (2003) Role of the G-protein-coupled receptor GPR12 as high-affinity receptor for sphingosylphosphorylcholine and its expression and function in brain development. J Neurosci 23: 907-914

Jensen LJ, Ussery DW, Brunak S (2003) Functionality of system components: conservation of protein function in protein feature space. Genome Res 13: 2444-2449

Jensen RA (2001) Orthologs and paralogs - we need to get it right. Genome Biol 2: INTERACTIONS1002

Joost P, Methner A (2002) Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands. Genome Biol 3: RESEARCH0063

Katoh K, Kuma K, Toh H, Miyata T (2005) Mafft version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511-518

Kemmeren P, Kockelkorn TTJP, Bijma T, Donders R, Holstege FCP (2005) Predicting gene function through systematic analysis and quality assessment of high-throughput data. Bioinformatics 21: 1644-1652

Klenk H, Spitzer M, Ochsenreiter T, Fuellen G (2004) Phylogenomics of hyperthermophilic archaea and bacteria. Biochem Soc Trans 32: 175-178

Kornegay JR, Schilling JW, Wilson AC (1994) Molecular adaptation of a leaf-eating bird: stomach lysozyme of the hoatzin. Mol Biol Evol 11: 921-928

Koski LB, Golding GB (2001) The closest Blast hit is often not the nearest neighbor. J

Mol Evol 52: 540-542

Krishnamurthy N, Brown D, Sjolander K (2007) Flowerpower: clustering proteins into domain architecture classes for phylogenomic inference of protein function. BMC Evolutionary Biology 7: S12

Kuang R, Ie E, Wang K, Wang K, Siddiqi M, Freund Y, Leslie C (2005) Profile-based string kernels for remote homology detection and motif extraction. J Bioinform Comput Biol 3: 527-550

Laudet V (1997) Evolution of the nuclear receptor superfamily: early diversification from an ancestral orphan receptor. J Mol Endocrinol 19: 207-226

Li L, Stoeckert CJJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes.. Genome Res 13: 2178-2189

Maddison WP, Knowles LL (2006) Inferring phylogeny despite incomplete lineage sorting. Syst Biol 55: 21-30

Mailund T, Brodal GS, Fagerberg R, Pedersen CNS, Phillips D (2006) Recrafting the neighbor-joining method. BMC Bioinformatics 7: 29

Martin AP, Burg TM (2002) Perils of paralogy: using HSP70 genes for inferring organismal phylogenies. Syst Biol 51: 570-587

Metpally RPR, Sowdhamini R (2005) Cross genome phylogenetic analysis of human and drosophila G protein-coupled receptors: application to functional annotation of orphan receptors. BMC Genomics 6: 106

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540

Page RD (1998) Genetree: comparing gene and species phylogenies using reconciled trees. Bioinformatics 14: 819-820

Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F (2005) Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5: 50

Plewniak F, Bianchetti L, Brelivet Y, Carles A, Chalmel F, Lecompte O, Mochel T, Moulinier L, Muller A, Muller J, Prigent V, Ripp R, Thierry J, Thompson JD, Wicker N, Poc (2003) PipeAlign: a new toolkit for protein family analysis. Nucleic Acids Res 31: 3829-3832

Plotz T, Fink GA (2005) Robust remote homology detection by feature based profile hidden markov models. Stat Appl Genet Mol Biol 4: 1

Prince VE, Pickett FB (2002) Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet 3: 827-837

Rannala B, Huelsenbeck JP, Yang Z, Nielsen R (1998) Taxon sampling and the accuracy of large phylogenies. Syst Biol 47: 702-710

Rehmsmeier M (2002) Phase4: automatic evaluation of database search methods. Brief Bioinform 3: 342-352

Rehmsmeier M, Vingron M (2001) Phylogenetic information improves homology detection. Proteins 45: 360-371

Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314: 1041-1052

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574

Serb JM, Oakley TH (2005) Hierarchical phylogenetics as a quantitative analytical framework for evolutionary developmental biology. Bioessays 27: 1158-1166

Sicheritz-Ponten T, Andersson SG (2001) A phylogenomic approach to microbial evolution. Nucleic Acids Res 29: 545-552

Sjolander K (2004) Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics 20: 170-179

Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 26: 320-322

Spang R, Rehmsmeier M, Stoye J (2002) A novel approach to remote homology detection: jumping alignments. J Comput Biol 9: 747-760

Spitzer M (2006) Automating the analysis of protein family evolution. PhD-Thesis. University of Muenster. 2006.

Spitzer M, Fuellen G, Cullen P, Lorkowski S (2004) VisCoSe: visualization and comparison of consensus sequences. Bioinformatics 20: 433-435

Stamatakis A (2006) RaXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690

Stechmann A, Cavalier-Smith T (2002) Rooting the eukaryote tree by using a derived gene fusion.. Science 297: 89-91

Stolle K, Schnoor M, Fuellen G, Spitzer M, Engel T, Spener F, Cullen P, Lorkowski S (2005) Cloning, cellular localization, genomic organization, and tissue-specific expression of the TGFbeta1-inducible smap-5 gene. Gene 351: 119-130

Storm CEV, Sonnhammer ELL (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics 18: 92-99

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 Y (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41

Theissen G (2002) Secret life of genes. Nature 415: 741

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680

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

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

Thornton JW, Kelley DB (1998) Evolution of the androgen receptor: structure-function implications. Bioessays 20: 860-869

Watson JD, Laskowski RA, Thornton JM (2005) Predicting protein function from sequence and structural data. Curr Opin Struct Biol 15: 275-284

Wheeler DL, Barrett T, Benson DA, 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 T (2007) Database resources of the national center for biotechnology information. Nucleic Acids Res. 35: D5-12

Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. J Mol Biol 326: 255-261

Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF (1998) Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res 26: 3986-3990

Zmasek CM, Eddy SR (2002) Rio: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics 3: 14

Table and Figure Legends.

Table 1. Predictions of ABC protein function by subfamily.

Table 2. Comparison of phylogenetic analysis pipelines. HSP, High-Scoring Segment Pair; HMM, Hidden Markov Model.

Fig. _1. Gene tree with hidden paralogy. A gene duplicated in the ancestor of *Malus, Citrus, Oryza* and *Hordeum*, yielding a red and an orange copy. After a period of co-existence, differential loss occurred (marked by daggers). The red copy in *Malus* now has an orange copy as its reciprocal closest match in *Citrus*, even though these copies are not orthologs. Correct orthology cannot be determined without further information such as a correct species tree. Viewing these genes in isolation, it looks like *Malus* is more closely related to *Hordeum*, and *Citrus* goes with *Oryza*. The inadvertent comparison of “apples” (red gene) and “oranges” (orange gene) puts the apple tree (*Malus*) apart from the orange tree (*Citrus*), and places one grass each (rice (*Oryza*) and barley (*Hordeum*), respectively), next to them as their closest relatives.

Fig. _2. Ranked list of hits that are returned by a homology search method (here: PSI-Blast), given an input query and a database to be searched. Alignment scores and E-Values are listed on the right.

Fig. _3. Receiver-Operator characteristic plot detailing CHASE accuracy on the detection of remote homologs, in comparison to standard methods. The plot displays the coverage (percentage of true positives) as a function of the false positives that must be tolerated to allow such coverage. The true positives are *remote* homologs, members of a protein family that is from the same *super*family as the query sequences. More precisely, all but one family of a given superfamily provide the search input, and members of the one family left out must be found. Results are averaged over a large test set of protein superfamilies.

Fig. _4. A gene tree with red and orange edges, embedded into a species tree of plants (white). The gene duplicated before the first speciation event, with no further duplications or losses.

Fig. _5. Flowchart of the RiPE pipeline. Starting with the profile of a protein family, a database search is conducted and the search results are taken directly to create a multiple alignment by stacking. The alignment is then used to infer the phylogenetic tree. (The green bars represent the profile and the pink bars the homologous parts of the database hits. Red vertical bars symbolize conserved regions).

Fig. _6. The function transfer rule described in Fuellen et al (2005) can be used to annotate the human proteins in a protein family under study, using annotation from non-vertebrate species. The rule transfers annotation between proteins in sister subtrees. In the simplest case, for a human protein H, its sister subtree just features the non-human ortholog N of H. In more complex cases, the rule transfers annotation from a larger sister subtree to a subtree that includes H. The sister subtree may contain uncharacterized paralogs P of the protein N from which the annotation is transferred, and it may contain human proteins HO that are putative orthologs of N. The subtree that contains H may include other human proteins J. Annotation transfer is successful if, nevertheless, function is conserved across the thick lines in the gene tree.

Fig. _7. Simplified tree of the ABCB subfamily. Domain arrangements found are the full-transporter arrangement transmembrane-ABC-transmembrane-ABC and the half-transporter arrangement transmembrane-ABC. The balloon labeled “Peptides” refers to human ABCB2/B3/B9 and yeast MDL1. The experimental annotation for MDL1 is “peptide transport”, and it matches the one for ABCB2/B3, which are also known as “TAP”, transporter associated with antigen processing. The balloon labeled “Fe / S” refers to human ABCB6/B7 and plant/yeast ATM3/ATM1. The experimental annotation for ATM1/ATM3 is “involvement in iron/sulfur cluster protein metabolism”, and it matches the one of at least human ABCB7. Another correspondence is “Hydrophobic compounds, colchicines”. “Lipids ?!” is a prediction that awaits confirmation.

ABC Subfamily	Correspondence (True positive prediction)	No correspondence (False positive prediction)	Prediction (no validation possible)
A	2	-	-
B	4	1	1
C	8	2	6
D	1	-	1
F	1	-	2
G	3	-	-
Total:	19	3	10
Examples	ABCB7: Fe/S-Cluster protein metabolism	ABCC7: channel, gluthathione, organic anions, bicarbonate (known) vs. amino acids (predicted, among other substrates)	ABCC10: glutathione-conjugates

Table 1

Criterion	NCBI treeview (Wheeler et al, 2007)	RiPE (Fuellen et al, 2005)	FIGENIX (Gouret et al, 2005)	ProteinUniverse (Brown and Sjolander 2006; Krishnamurthy et al, 2007)	PhyloGenie (Frickey and Lupas, 2004)	PipeAlign (Plewniak et al, 2003)	RIO (Zmasek and Eddy, 2002)
Input	Single sequence (PSSM [position-specific scoring matrix] via advanced options)	Single sequence or aligned set of sequences	Single sequence	Single sequence, set of sequences, general search terms (e.g. Gene ontology phrases)	Single sequence	Single sequence or set of sequences	Single sequence and corresponding Pfam (Bateman et al, 2000) alignment
Choice of searchable databases	Any NCBI database	Any set of NCBI-formatted protein databases	Any set of NCBI-formatted protein databases	NR database, organism-specific databases, protein-family-specific databases, more	Any set of NCBI-formatted protein databases	Swissprot and TrEMBL, Varsplic (spicing variants), PDB	Swissprot and TrEMBL
Choice of organisms	User-based choice available	User-based choice by selection of database	User-based choice available	User-based choice available	User-based choice available	-	-
Filtering of sites or subsequences of the sequences	Homologous-regions-only data (stacked HSPs)	Use of full-length sequences or homologous-regions-only data (stacked HSPs)	Elimination of sites that do not evolve neutrally; automatic detection and compaction of alignment columns with too many gaps	Filtering by family- *and* subfamily-specific conservation using a crude gene family tree and HMMs (by PSI-PHY and Flowerpower); elimination of columns with too many gaps	Use of homologous-regions-only data corrected by multiple sequence re-alignment	None, but re-ranking of search results according to presence of local conservation motifs	-
Handling of query sequence domain organization	-	None; relies on user-defined "maximum unit of common evolutionary heritage" that is domain-repeat-free	Automatic domain detection	-	-	-	Manual input of Pfam domain
Handling of database sequence domain organization	-	Processing of HSPs to yield reassembled sequences that match the domain organization of the "maximum unit"	Selection of domains / repeats with congruent evolution based on an expert system	Low quality of global alignment and conflicting Pfam domain organization detects domain organization outliers.	-	-	No further consideration of domain issues
Basis for the multiple alignment	Stacking of HSPs	Stacking of HSPs by Mview (Brown et al, 1998); optional realignment by Mafft (Katoh et al 2005), or Muscle (Edgar, 2004)	ClustalW (Thompson et al, 1994)	Clusters of globally alignable homologs (provided by Flowerpower) realigned with Muscle (Edgar, 2004)	Stacking of HSPs and realignment of problematic regions	DBClustal (Thompson et al, 2000), using anchors derived by the "Ballast" component of PipeAlign	Pfam (Bateman et al, 2000)

Filtering database sequences for inclusion in the tree/ clustering	-	Elimination of sequen-ces without the most prominent motifs possible; data can be filtered using (sub)family boundaries based on E-Value; filtering of splice variants; filtering of sequences with ≥ 95% similarity that belong to the same species; filtering of fragments	Elimination of sequences with divergent amino acid composition; of doubles (with threshold parameter); of sequences reducing alignment quality; of sequences with unusual length.	None; all filtering done before tree reconstruction	Elimination of sequences with above-threshold similarity (that belong either to the same strain, species or genus)	Elimination of sequences without local conserva-tion motifs, or without the core blocks of the multiple alignment	-
Choice of phylogeny reconstruc tion method	Only neighbor Joining and Minimum Evolution are available	Manual choice of method is possible	Choice of 3 methods and projection on consensus tree	Choice of 4 methods; additionally a SATCHMO (Edgar and Sjölander 2003) analysis is possible	Only neighbor joining is available	No phylo-geny, but choice of two clustering methods, one of which uses neighbor joining	-
Tree recon-ciliation	-	-	Automatic detection of speciation/ duplication / orthology / paralogy	-	Automatic detection of speciation/ duplication / orthology / paralogy	-	-
Functional annotation	Links to NCBI protein database incl. annotation	-	Automatic extraction of functional annotation using a multi-agent system	Detection of enrichment in GO (Gene Ontology) and EC (Enzyme Class) annotation; Pfam domain annotation; input from experts and text mining tools	-	-	-

Table 2

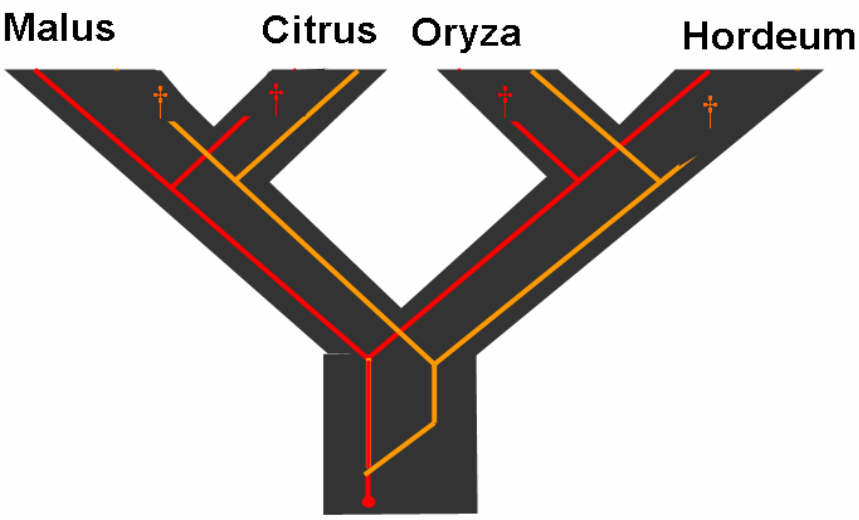


Figure 1.

BLASTP 2.2.17 (Aug-26-2007)

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

Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding
environmental samples from WGS projects
5,507,867 sequences; 1,907,794,466 total letters

Query= gi|11321634|ref|NP_036252.1| CD2-associated protein [Homo sapiens]

Sequences producing significant alignments:	Score (Bits)	E Value
ref|NP_036252.1| CD2-associated protein [Homo sapiens]	1029	0.0
emb|CAH91263.1| hypothetical protein [Pongo pygmaeus]	1023	0.0
ref|XP_527616.2| PREDICTED: CD2-associated protein [Pan troglody	1020	0.0
....		
emb|CAG31983.1| hypothetical protein [Gallus gallus]	633	1e-179
gb|AAI02655.1| LOC533188 protein [Bos taurus]	562	3e-158
ref|NP_001086432.1| hypothetical protein LOC445851 [Xenopus l...	448	6e-124
emb|CAJ83986.1| CD2-associated protein [Xenopus tropicalis]	374	1e-101
emb|CAG13005.1| unnamed protein product [Tetraodon nigroviridis]	367	1e-99

Figure 2.

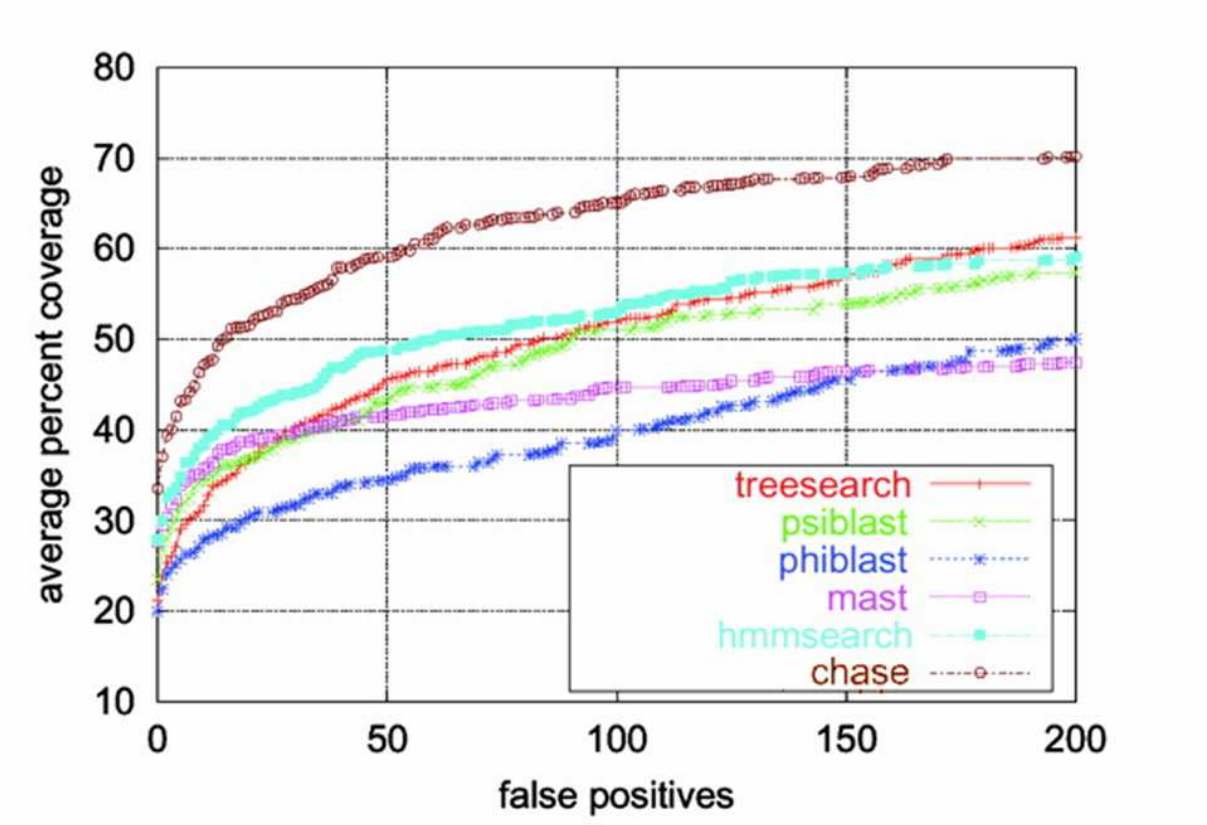


Figure 3.

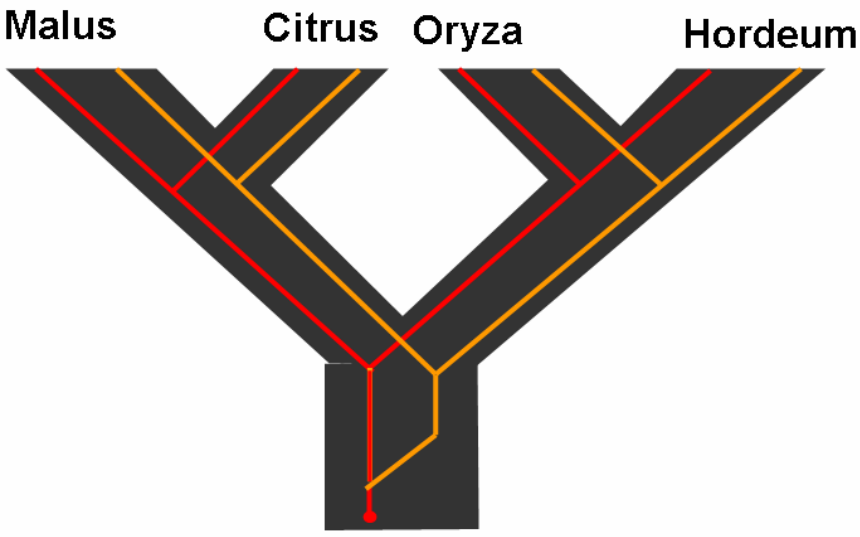
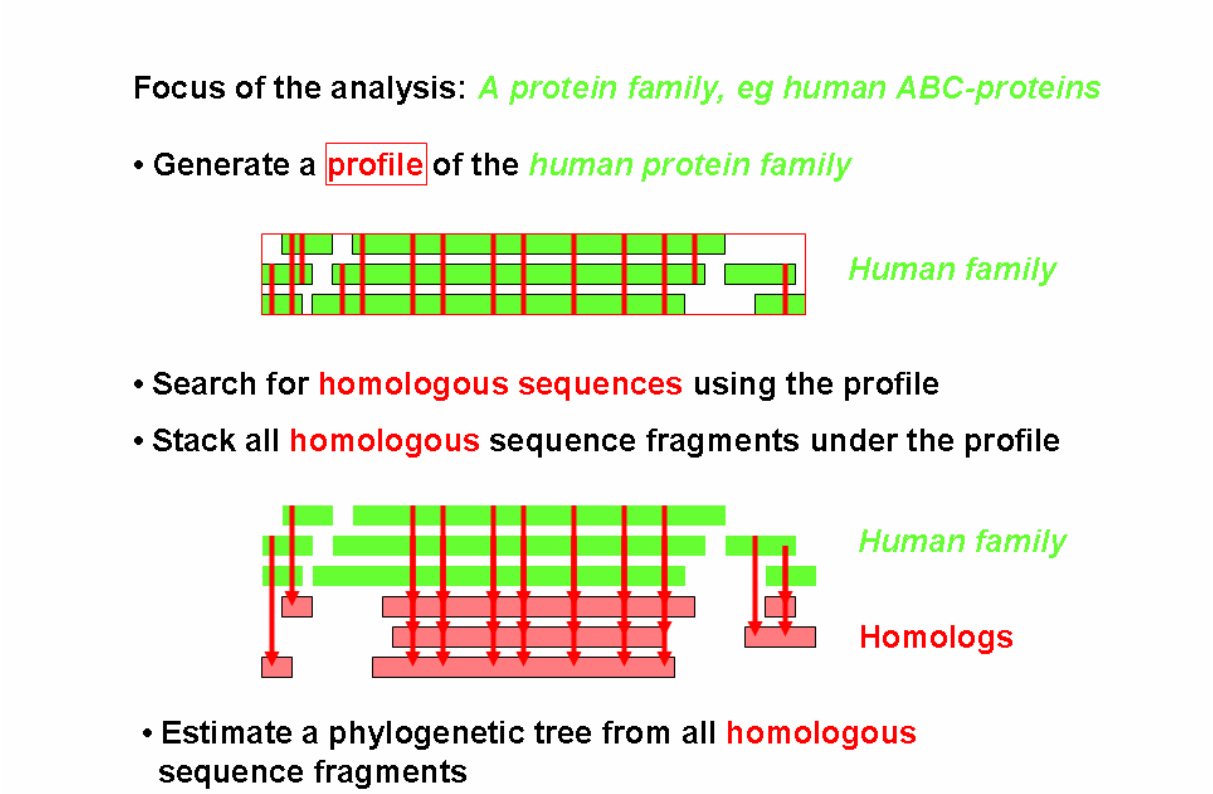


Figure 4.



70

Figure 5.

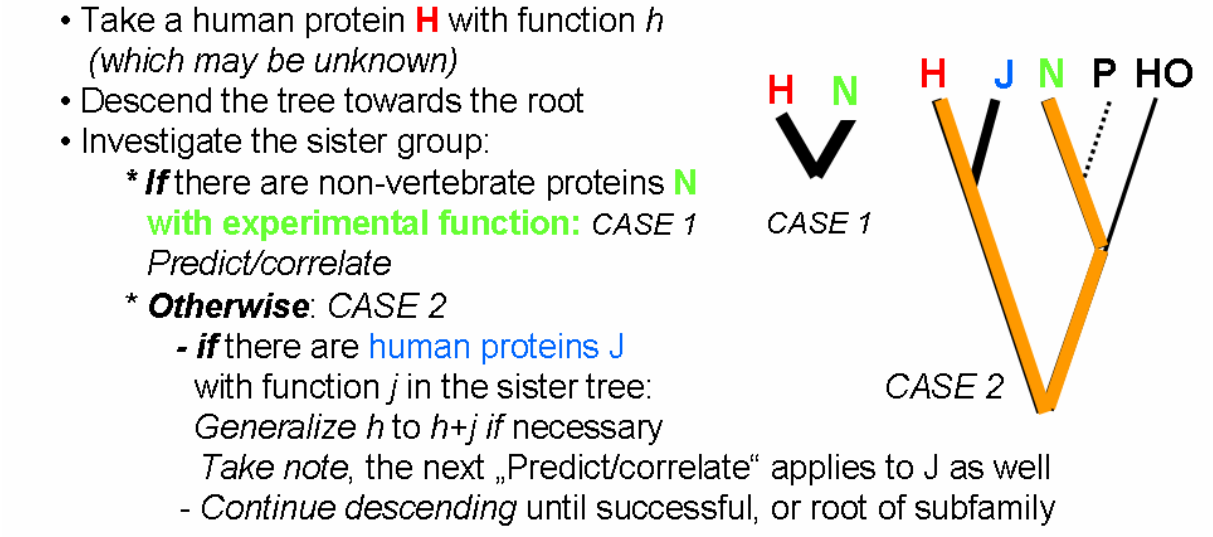
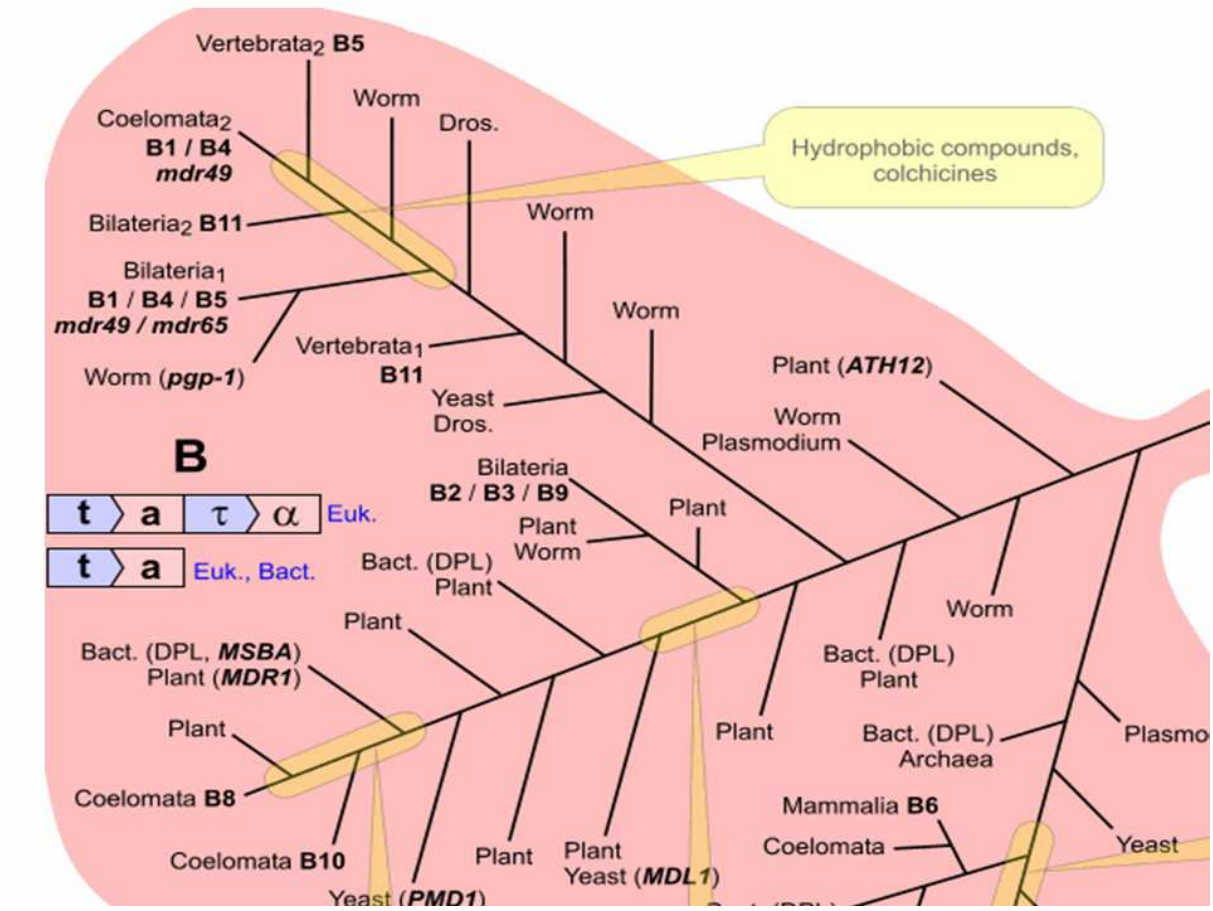


Figure 6.


Vertebrata2 B5
Worm
Dros.
Coelomata2
B1 / B4
mdr49
Hydrophobic compounds,
colchicines
Bilateria2 B11
Worm
Bilateria1
B1 / B4 / B5
mdr49 / mdr65
Worm
Worm (pgp-1)
Vertebrata1
B11
Yeast
Dros.
Plant (ATH12)
Worm
Plasmodium
B
Bilateria
B2 / B3 / B9
Plant
t a τ α Euk.
Plant
Worm
t a Euk., Bact.
Bact. (DPL)
Plant
Worm
Plant
Bact. (DPL, MSBA)
Plant (MDR1)
Bact. (DPL)
Plant
Plant
Bact. (DPL)
Archaea
Plasmo
Coelomata B8
Mammalia B6
Coelomata
Yeast
Coelomata B10
Plant
Plant
Yeast (MDL1)
Yeast (PMD1)

75

Figure 7.

Electronic Supplementary Material S1

Genomic Homology Search - a method combination approach, *Knowledge transfer based on comparative genomics: Genomic Homology Search for mouse S100A12* and *Auxiliary tools for homology search and phylogeny.*

Genomic Homology Search - a method combination approach. Using the CHASE idea, we started working on a scheme called GenCHASE that combines genomic homology search methods such as TPSI-BlastN (Altschul et al, 1997) and TFASTY (Pearson et al, 1997), striving for better input data for gene modeling, i.e. more precise data on the regions that, if translated, feature similarity, in part, to known protein sequence and that may therefore be part of the exons of homologous genes. Indeed, GenCHASE combined with gene modeling software such as GeneWise (Birney et al, 2004) and GenScan (Burge and Karlin, 1997) was able to detect a novel human ABC transporter that was subsequently shown to be expressed (see Alam, 2005), and it delivered a promising yet ambiguous candidate for the DAHP gene in the malaria agent Plasmodium falciparium. The DAHP gene is important because it is part of the Shikimate pathway that does not exist in vertebrates. Therefore, inhibiting this protein may harm Plasmodium, but neither human nor other vertebrates. GenCHASE is currently described best in the PhD thesis of Alam (2005). A genomic homology search for the mouse S100A12 protein (Fuellen et al, 2003, 2004) is described in the next section. Here, genome browser websites were used to do the searches. GenCHASE searches later on did not reveal anything new.

Knowledge transfer based on comparative genomics: Genomic Homology Search for mouse S100A12. The study of model organisms such as mouse gives us important information that can be compared to information gained from studies involving human. Furthermore, predictions for human can be made using mouse as a model. More generally, computational comparative analyses can relate genomic, expression and other data for many species, and knowledge can be transferred. In our case study on ABC protein function prediction (described in the main text), we confirmed the hypothesis that such transfer is more successful using a phylogenetic tree instead of simple homology searches. Computational predictions are often the only information available if studies in humans are impossible. However, there are many reasons for caution regarding any knowledge transfer from animal to human, and it would be preferable to have confidence values of some sort. The closer the data to be interpreted and the underlying molecular biology match between model organism and human, the higher the confidence that homologous phenomena are studied, and the better the chance that knowledge transfer is valid.

For example, some pathways involved in inflammation are highly conserved between mouse and human. Both species share the RAGE protein, the Receptor of Advanced Glycation End products (Deane et al, 2003). In humans, the S100A12 protein was shown to interact with human RAGE, fostering inflammation (Hofmann et al, 1999). In the literature, a homologous scenario was reported for mouse, proposing that mice are a good model for investigating this kind of inflammation (Hofmann et al, 1999; Schmidt et al, 2001). We called this proposal into question by performing homology searches in mouse, failing to find a mouse *S100A12* gene in the first place (Fuellen et al, 2003, 2004). Since this work predates the development of GenCHASE, we performed the homology searches using standard WWW tools, in particular homology searches at the Jackson lab (Blake et al, 2003) and gene locus investigations using the UCSC genome browser (Kent et al, 2002). Searching with human *S100A12*, we found a single candidate in mouse that matches closely but only partially; it is a sequence covering the first (noncoding) exon and a few hundred nucleotides of intronic sequence before and after, with 60% similarity. The matching TATA box in mouse is non-canonical, and the remaining exons 2 and 3 of the human *S100A12* have no match whatsoever in mouse, so we can assume that the gene is not functional in mouse. Moreover, the partial match in mouse is immediately followed by sequence that matches human sequence many kilobases away from the human *S100A12* gene, so that we can assume that a large segment was deleted in mouse. In fact, since the situation is similar in rat, all murinae may share the deletion. In this case, knowledge transfer failed: the earlier

reports were based on hybridization assays that obviously were not rigorously validated, as the hybridizing protein that was supposed to be mouse S100A12 was not sequenced.

Auxiliary tools for homology search and phylogeny. In the context of the RiPE pipeline, we developed a few associated tools that ease various steps, or give additional insight. At the start of the pipeline, we would like to have a search profile, even though our pipeline would also accept a single sequence. The search is based on some or all members of the protein family known beforehand, e.g. all human ABC protein sequences. To simplify sequence retrieval, we wrote a small web-based tool (Mersch and Fuellen, 2003) that takes a table of sequence names and accession codes directly from the PDF (or HTML) of a publication, and returns the sequences from public databases. The tool, paper2sequences, is based on the Bioperl package (Stajich et al, 2002) and it features some heuristics to maximize the chances of finding the sequences in question.

In the middle of the pipeline, we face the issue of splice variants, also known as isoforms. Many genes can be decomposed into two or more exons, between which non-coding introns are located. Often, a gene can give rise to different proteins by way of alternative splicing; the different protein products differ in their exon composition: some exons may be missing in some products, may be shortened, etc. These splice variants can clutter the analysis pipeline, slowing it down and expanding the size of the resulting phylogenetic tree. Generally, they do not add value; there are very few specific functional annotations for splice variants (see, however, Searls 2003, Figure 2 for an example). Therefore, we designed and implemented a method to filter out splice variants, without access to the genomic sequence, since we work with protein databases. Our method, IsoSVM (Spitzer et al, 2006), uses state of the art machine learning technology in the form of support vector machines. SVMs are used to achieve best possible accuracy; we do not want to miss a single member of the protein family under investigation just because it is mistakenly classified as a splice variant while in fact it is a paralog.

Towards the end of the pipeline, we have to handle very large gene trees, and we wrote a tool called TreeSimplifier (Lott et al, 2006) to simplify these to some degree, in particular by collapsing subtrees where a gene evolved according to species phylogeny without any duplication. In other words, all genes in a collapsed subtree are orthologs. Given a species tree, we can for example summarize a subtree of ((ABCC10_human, ABCC10_mouse), ABCC10_fugu) to a single leaf “ABCC10_vertebrata”. Further, our tool allows blurring the distinction between closely related species, e.g. summarizing different yeast species by a common label. This allows simplifying the gene tree further. Optionally, the tool even allows fixing a very limited amount of error in tree topology. Altogether, we were able to simplify an ABC protein tree of 1138 leaves to 397 leaves, and a POU transcription factor tree from 185 to 98 leaves.

Our approach to function prediction does not yet consider the domain structure of proteins, i.e. the decomposition of the entire sequence into smaller subsequences conserved across families (and across species) that may in part be the source of functionality. In case of ABC proteins, ignoring domain structure does not seem to cause problems; specific function is associated with the entire sequence, and the recognition of two domains, an ATP-binding-cassette and a transmembrane region, causes internal-repeat issues instead. However, a step towards a more detailed analysis was taken by developing VisCoSe (Spitzer et al, 2003), which allows us to visualize domains as well as motifs. More specifically, VisCoSe performs a multiple alignment of consensus sequences. These are color-coded by conservation (based on the underlying alignment). Thus, we can delineate domains that are color-coded as conserved subsequences, identifying e.g. the subdomains of the ABC cassette (Walker 1, signature sequence, Walker 2). These can then be compared for different subtrees, or for different groups of species, and an evolutionary analysis can be performed.

References

Alam I (2005) Integrative Approaches to Homology Search, (PhD), University of Bielefeld, 2005

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14: 988-995.

Blake JA, Richardson JE, Bult CJ Kadin JA, Eppig JT; Mouse Genome Database Group (2003) MGD: the Mouse Genome Database. Nucleic Acids Res 31: 193-195.

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268: 78-94.

Deane R, Du Yan S, Submamaryan RK et al. (2003) RAGE mediates amyloid-beta peptide transport across the blood-brain barrier and accumulation in brain. Nat Med 9: 907-913.

Fuellen G, Foell D, Nacken W, Sorg C, Kerkhoff C. (2003) Absence of S100A12 in mouse: implications for RAGE-S100A12 interaction. Trends Immunol 24: 622-624.

Fuellen G, Nacken W, Sorg C, Kerkhoff C (2004) Computational searches for missing orthologs: the case of S100A12 in mice. OMICS 8: 334-340.

Hofmann MA, Drury S, Fu C et al. (1999) RAGE mediates a novel proinflammatory axis: a central cell surface receptor for S100/calgranulin polypeptides. Cell 97: 889-901.

Lott P, Mundry M, Sassenberg C, Lorkowski S, Fuellen G (2006) Simplifying gene trees for easier comprehension. BMC Bioinformatics 7: 231.

Kent WJ, Sugnet CW, Furey TS, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.

Mersch H, Fuellen G (2003) Paper2sequences: retrieval of sequences listed in a publication. Appl Bioinformatics 2: 113-116.

Pearson WR, Wood T, Zhang Z, Miller W (1997) Comparison of DNA sequences with protein sequences. Genomics 46: 24-36.

Schmidt AM, Yan SD, Yan SF, et al. (2001) The multiligand receptor RAGE as a progression factor amplifying immune and inflammatory responses. J Clin Invest 108: 949-955.

Searls DB (2003) Pharmacophylogenomics: genes, evolution and drug targets. Nat Rev Drug Discov 2: 613-623.

Stajich JE, Block D, Boulez K, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12: 1611-1618.

Spitzer M, Lorkowski S, Cullen P,Sczyrba A, Fuellen G (2006) IsoSVM--distinguishing isoforms and paralogs on the protein level. BMC Bioinformatics 7: 110.