universität wien

DISSERTATION

Titel der Dissertation

„Domains, Proteins, and Evolution“

Verfasserin

Tina Köstler

angestrebter akademischer Grad

Doktorin der Naturwissenschaften (Dr.rer.nat.)

Wien, 2012

Studienkennzahl lt. Studienblatt:	A 091 490
Dissertationsgebiet lt. Studienblatt:	Molekulare Biologie
Betreuer:	Univ.-Prof.Dr. Arndt von Haeseler

Abstract

A major challenge in the post-genomic era is the annotation of functionally uncharacterized proteins that emerge from an ever increasing number of sequencing projects. This task is almost always accomplished by the transfer of information from other known proteins. Here, we evaluate the strengths and weaknesses of established and novel bioinformatics approaches to transfer functional annotations from characterized to yet uncharacterized proteins. Starting with the fundamentals of homology inferred via sequence similarity, we expand the concept of functional inference from homology to functional inference from protein domain structure. We introduce the term feature architecture to summarize the entirety of functional domains, secondary structure elements, and compositional properties, and show that feature architecture similarity serves as a good proxy for the degree of functional similarity between two proteins. With FACT, we provide an implementation of a feature architecture based search algorithm. Subsequently, we evaluate the reliability of domain detection and investigate the evolution of protein domains in a simulation framework. We, therefore, introduce REvolver, a simulator implementing biologically meaningful models of protein sequence evolution by taking domain constraints into account. More precisely, REvolver extracts information from a profile Hidden Markov Model (pHMM) of a domain to automatically parameterize position specific substitution models. Guided by the pHMM it also places insertions and deletions preferentially at positions where they have been observed in other domain instances. In our simulation of protein domain evolution, we identified domains that lose their domain characteristics already after few substitutions. Others preserve their characteristics over large evolutionary distances. Interestingly, some domains repeatedly lose and regain their characteristics in the course of simulated evolution. We discuss this phenomenon in greater detail and suggest a maximum likelihood approach to distinguish between domain detection errors and true evolutionary losses and gains. We then propose how to extend our approach from individual domains to the entire protein and investigate over what evolutionary distances we expect orthologs to be detectable.

Finally, we apply methods to detect orthologs and functional equivalents in the proteomes of microsporidia and zygomycetes. Thereby, we discuss their proposed monophyly and investigate the evolutionary ancestry of sex determination. This analysis illustrates the versatility and complementarity of ortholog inferences and feature architecture similarity searches in the search for functionally equivalent proteins.

Parts of this thesis have been published in the following articles:

(i) T. Koestler, A. von Haeseler, and I. Ebersberger (2012) Modeling sequence evolution under domain constraints. *Molecular Biology and Evolution*, in press

(ii) T. Koestler and I. Ebersberger (2011) Zygomycetes, microsporidia, and the evolutionary ancestry of sex determination. *Genome Biology and Evolution*, (3):186-194.

(iii) T. Koestler, A. von Haeseler, and I. Ebersberger (2010) FACT: Functional annotation transfer between proteins with similar feature architecture. *BMC Bioinformatics*, **11**(1):417.

In preparation:

(i) T. Koestler, B. Q. Minh, I. Ebersberger, and A. von Haeseler
Stability of Protein Domains.

(ii) T. Koestler, A. von Haeseler, and I. Ebersberger
Evolutionary traceability of Proteins.

Contents

Chapter 1

Introduction

A phylogenetic tree is one way to represent the evolutionary relationships of species that all evolved from a single common ancestor (Darwin, 1959). Hence, the genomes of two species underwent independent mutations only for a certain amount of time. Before speciation they shared an ancestral genome and thus the same mutations. From this follows that the more closely related two species are the more similar are their genomes assuming an approximately constant rate of evolutionary sequence change. The same applies for the entire set of proteins encoded in a genome - the proteome. Since the advent of next generation sequencing the number of fully sequenced genomes and available proteomes is increasing rapidly. Meanwhile, sequences from almost all major groups in the eukaryotic tree are available. However, whole genome sequencing efforts typically end with the annotation of the draft genome sequence and little is known about the encoded genes beyond their exon-intron structure and their sequence. Investigating the function of these sequences experimentally in all sequenced species is not feasible, and functional characterizations of genes and proteins are limited to a small number of model organisms. It is, thus, up to computational approaches to transfer insights into the function of individual proteins from model organisms to other species.

For the simple transfer of information from one species to another pairwise comparisons are often sufficient, however, in order to learn more about the evolution of the proteins in question we need to consider larger phylogenetic trees. The presence and absence information of a protein in a phylogeny is called phylogenetic profile. It was shown that proteins of similar phylogenetic profiles strongly tend to be functionally linked (Pellegrini et al., 1999). Thus, the observation that two proteins are either both present or both absent suggests that they cooperate in some biological process i.e. part

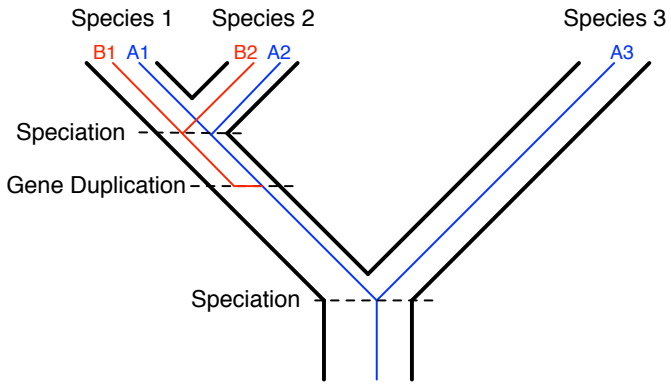


Figure 1.1: Evolutionary relationships of species and their genes. The black tree shows the phylogeny of species 1, 2, and 3. All genes are homologous. A1 and A3 are orthologs whereas B1 and A2 are paralogs.

of the same pathway or building a protein complex (Pellegrini et al., 1999). Moreover, the shared presence or absence of proteins belonging to the same functional module indicates when during evolution the module emerged or where it was, for example, lost. Bioinformatics approaches to search for functionally equivalent proteins together with the available proteome data from almost all major groups in the eukaryotic tree facilitate now to investigate in what eukaryotic species a functional module is present, when during evolution it emerged, and how it evolved. As a result, a more refined picture of organismal and functional evolution will emerge.

1.1 Homology

Homology is a central concept in biology implying a common origin of traits. In other words, pairs of genes that share a common ancestry are called *homologs*. Depending on the evolutionary event that separated the extant genes, we distinguish between *orthologs* and *paralogs*. Orthologous genes descended from a single gene in their ancestral species and were separated by a speciation event. Paralogous genes, on the other hand, were separated by a duplication event (Fitch, 1970). Figure 1.1 shows a phylogenetic tree of 3 species. Species 1 has one copy of both genes B1 and A1, species 2 of B2 and A2, and species 3 only a copy of gene A3. All extant genes share a common ancestor in the ancestral species of all three species. Thus, all genes are homologous. More specifically, genes A1 and A2 are orthologs because their lineages were separated by a speciation event. The same applies to the gene pairs B1/B2,

A1/A3, and A2/A3. In contrast, the gene pairs A1/B2 and B1/A2 are paralogous: In the common ancestor of species 1 and 2 a gene duplication gave rise to genes A and B. Please note that orthology is not transitiv. For example, A2 is orthologous to A3, A3 is orthologous to B1, however B1 is paralogous to A2.

Later, two additional terms were introduced to further distinguish between gene pairs in different species that were duplicated i) before the speciation (*out-paralogs*) and ii) after the speciation (*in-paralogs*) (Remm, Storm and Sonnhammer, 2001). Figure 1.1 illustrates a scenario where a gene duplication has happened after the speciation that gave rise to species 3 and the last common ancestor of species 1 and 2. This event has resulted in two genes in the last common ancestor of species 1 and 2 that were both retained in the contemporary species 1 and 2. Both genes in species 1 (A1 and B1) were separated from the gene A3 in species 3 via a speciation event. Thus, each of these genes is orthologous to A3. However, with respect to each other, genes A1 and B1 are paralogous. To indicate that the gene duplication event post-dates the speciation of species 3, A1 and B1 are called in-paralogs with respect to this speciation event. In contrast, genes B1 and A2 are an example for out-paralogs with respect to the split of species 1 and 2.

Inference of Evolutionary Relationships

The introduced terms to describe the evolutionary relationships between genes also apply to proteins and were defined already more than four decades ago by Fitch (1970). However, to decide whether two proteins are orthologous or paralogous is still challenging (Dessimoz et al., 2012). To infer their evolutionary relationships, we commonly assume that the degree of sequence similarity between proteins reflects their degree of relatedness. This simplification is justified by the observation that the number of amino acid differences between evolutionarily related proteins changes approximately linearly with time (Zuckerkandl et al., 1962). In general, the rate of evolutionary changes of a protein is assumed to be constant over time (Margoliash, 1963; Kimura, 1979). Although several deviations from this assumption have been shown (reviewed in Kumar 2005 and Schwartz and Maresca 2006), it is widely deployed.

A commonly used way to detect orthologs relies on the assumption that a protein exhibits a higher sequence similarity to an ortholog than to any other protein outside its own proteome. A standard procedure to identify these proteins is therefore to measure the sequence similarity between a query protein and each member of a search

proteome. The most similar sequence is then considered as a putative ortholog. An extension of this unidirectional best hit approach is a reciprocal best hit approach (RBH) that consists of 3 steps. We exemplify the procedure by using Blast as sequence similarity search tool (Altschul et al., 1997):

1. **Blast:** Choose a protein from the proteome of species 1 and perform a Blast search in the proteome of species 2.
2. **Re-Blast:** Take the best Blast hit from species 2 as query for the reciprocal similarity search in the proteome of species 1.
3. **Evaluate reciprocity:** If the best Blast hit in step 2 is the query protein from step 1, the two proteins are reciprocal best hits, and we thus consider them as orthologs. Otherwise, we could not identify an ortholog to the protein in species 1.

The reciprocity in the similarity search is crucial for the ortholog identification. Consider the example shown in figure 1.2: Subsequent to a gene duplication in the common ancestor of species 1 and species 2, the A variant in species 1 was lost. In contrast, in species 2 both duplication products (A2 and B2) were retained. Assume we search in species 1 for the ortholog to gene A2. Following the RBH approach, we start a Blast search against species 1. Gene B1 would show up as best Blast hit, since it is closest to A2. However, A1 and B1 were separated by a duplication event and are therefore paralogs. According to the RBH protocol, we now take B1 as query for the Blast search in species 2. B2 will be most similar. Thus, A2 and B1 are not reciprocal best Blast hits and we therefore conclude correctly that species 1 does not have an ortholog to A2. It was shown that RBH based ortholog assignments reduce the false positive rate by a factor of 6 compared to unidirectional best hit based assignments (Chen et al., 2007).

Using the RBH approach we identify pairs of orthologs (one-to-one orthologs). However, as shown in figure 1.1, there exist also one-to-many or many-to-many ortholog relationships. InParanoid (Remm, Storm and Sonnhammer, 2001) is a program that extends the RBH approach to achieve a more exhaustive ortholog inference by adding in-paralogs to a pair of orthologs. InParanoid first identifies a best Blast hit pair as orthologs that additionally fulfills the following two criteria. First, the symmetric Blast score (average of the Blast score (A,B) and (B,A)) needs to be above a certain threshold (50 bits per default) and second, the overlapping region must exceed 50% of the length of the longer protein. These filtering steps are performed in order to

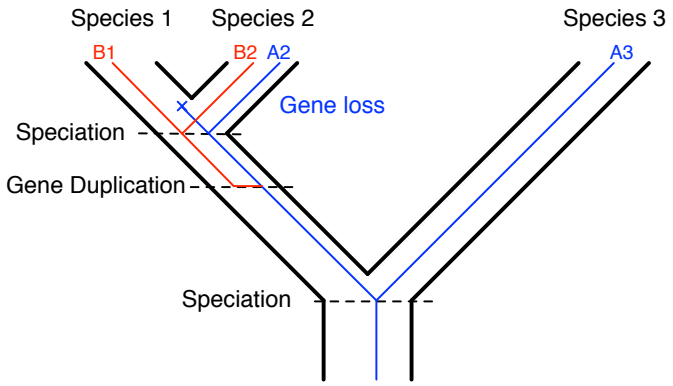


Figure 1.2: Evolutionary relationships of species and their genes. The black tree shows the phylogeny of species 1, 2, and 3. After the speciation of species 1 and species 2, one of the duplicated genes was lost in species 1. Consequently, species 1 comprises one gene (B1), species 2 comprises two genes (A2 and B2), and species 3 comprises one gene (A3). The most similar gene to A2 in species 1 is gene B1. However, it is not an orthologous gene. A Blast search of gene B1 in species 2 reveals this situation by identifying B2 as best Blast hit.

reduce false positives that result from short local similarities. In cases where Blast generates multiple high-scoring segment pairs (HSPs), InParanoid requires that they are in the same relative order on both sequences. Additionally the HSPs must not overlap by more than 5%. For a detailed description of the updated overlap criteria or the low-complexity filters see Ostlund et al. (2010). Afterwards, proteins that are more similar to the orthologous protein in the same species than to any other protein in the other species are added. Again, assuming that all proteins, by and large, evolve at the same rate, these proteins in the same species must have separated after the speciation event and are therefore in-paralogs. Figure 1.3 illustrates the distinction between in- and out-paralogs.

1.2 Three Major Questions

The prevalent method to transfer experimentally verified functional annotations from one protein to yet uncharacterized proteins and to construct phylogenetic profiles is the search for orthologs. However, orthology-based approaches bring along three questions that shall be addressed with the studies and methods presented this thesis.

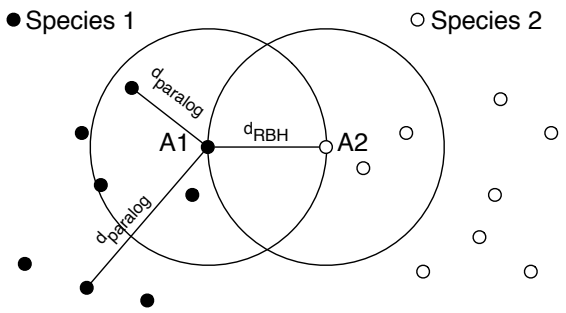


Figure 1.3: Adding in-paralogs to a reciprocal best hit pair (modified from Remm, Storm and Sonnhammer 2001). Filled circles represent proteins from species 1 and empty circles proteins from species 2. Protein A1 from species 1 and A2 from species 2 are a reciprocal best hit pair. The distance between the two proteins in the graph is inversely proportional to their sequence similarity. All other proteins are also placed according to their distances. Any protein that lies within the two big circles with radius d_{RBH} ($d_{paralog} \leq d_{RBH}$) is more similar to the ortholog in the same species that to any other protein in the respective other species. It is, therefore, added as in-paralog.

Do orthologs have the same function? Ohno (1970) was among the first to speculate about the genealogy of genes and their functional similarity. The idea was that one gene copy after a gene duplication preserves the original function and the second copy is more or less free to evolve and has the potential to develop a new function. From this it was concluded that orthologs generally preserve their ancestral function, whereas paralogs are likely to functionally diverge. More than four decades later, this hypothesis has been captured in the so called *ortholog conjecture* - that at a similar degree of sequence divergence, orthologs are generally more conserved in function than genes that result from a duplication. The ortholog conjecture is widely accepted, however some debate is still going on (Dessimoz et al., 2012). Supporting evidence comes, for example, from a study where it was shown that orthologous proteins are significantly more similar in domain structure than paralogs at the same evolutionary distance (Forslund, Pekkari and Sonnhammer, 2011). Furthermore, orthologous gene pairs have a significantly higher degree of intron position conservation compared to non-orthologous pairs (Henricson, Forslund and Sonnhammer, 2010). In contrast, a study using Gene Ontology (GO; Ashburner et al. 2000) terms indicates that paralogs have a higher functional similarity than orthologs and paralogs are therefore often better predictors of function than are orthologs (Nehrt et al., 2011). Thus, in addition to sequence similarity further evidences are needed to infer functional equivalence from

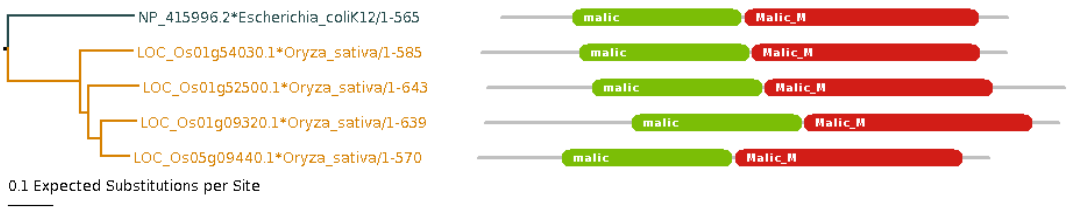


Figure 1.4: Orthologs between *O. sativa* and *E. coli* plus in-paralogs as displayed on the InParanoid webpage. The proteins are related by a tree where the branches leading to sequences of the same species have the same color. Next to the sequence names, the domain annotation of each protein is shown. All proteins consist of a malic (green) and a Malic_M (red) Pfam domain.

orthologs.

Protein domains can be informative about the proteins' function since they take over individual functions within a protein. As a first step, we added a new feature to InParanoid to quickly assess whether orthologous proteins share the same functional domains and thus may have the same function (Ostlund et al., 2010). For an ortholog group the neighbor-joining bootstrap tree is displayed together with the Pfam (Finn et al., 2008) domain annotations for the individual sequences. Figure 1.4 shows four proteins from *O. sativa* that are all orthologous to one protein in *E. coli*. All proteins comprise a malic and a Malic_M Pfam domain. Thus, the conservation of the domain content adds further evidence on the functional equivalence of these orthologs. In contrast, if the domains are different between orthologs a functional annotation transfer between them should be treated with caution. For a more thorough assessment of functional similarity we extended this concept by considering a more comprehensive set of protein features that included functional domains, secondary structure elements, and compositional properties. We implemented a feature dotplot to contrast the feature architectures of two proteins.

What if functional equivalents are not orthologous? There are cases where the same function evolved multiple times independently (Fitch, 1970; Galperin, Walker and Koonin, 1998; Gough, 2005; Forslund and Sonnhammer, 2008). Thus, although being functionally equivalent, the corresponding proteins did not descend from a common ancestor and any sequence similarities are only due to chance. This calls for comprehensive methods to complement sequence similarity-based approaches in the search for functional equivalents. To close this gap in methodology, we introduce

a Feature Architecture Comparison Tool (FACT). FACT is designed to search for proteins of similar feature architecture to a query protein. The similarity in feature architecture between proteins is taken as a proxy for their functional similarity. We investigate the performance of identifying functionally equivalent proteins based on sequences and feature architecture similarity. Our analysis shows that the two approaches complement each other and annotation transfers between proteins are most reliable when feature architecture similarity and sequence similarity are taken into account jointly.

Over what evolutionary distances can orthologs be found? We show that feature architecture similarity based approaches compensate some limitations in the search for functional equivalents. However, feature architecture similarity between two proteins is not sufficient evidence to assign evolutionary relationships. This assignment still relies on sequence similarity. This can pose a problem, because the more orthologous proteins diverge in their sequences, the harder it gets to identify them as orthologs. In cases where orthology inferences fail, the question remains whether an ortholog is truly absent or whether it is present but may be too diverged to exhibit enough sequence similarity, and thus is not detected due to limitations of the algorithm. This issue is especially important, since recent efforts for whole genome sequencing with subsequent downstream analyses target species that are separated from their most closely related well annotated species by billions of years of evolution e.g. early branching eukaryotes such as *T. brucei* (Baurain et al., 2010; Berriman et al., 2005; Embley and Hirt, 1998; Philippe, 2000). The identification of orthologs is also difficult in extremely fast evolving species as, for example, microsporidia (Thomarat, Vivarès and Gouy, 2004). Thus, the integration of these data into the existing network of functionally annotated proteins is challenging. Here, we address the question 'How far back in time can we trace a protein?'. Evolutionarily conserved domains are often the trigger for orthology assignments due to a high local sequence similarity. However, domains are rapidly lost in standard simulations of protein sequence evolution. The first step in this analysis was, therefore, the development of REvolver a program to simulate protein evolution with conserved domain structure. REvolver considers information regarding which sequence sites remain conserved over time and where in a domain insertions or deletions are likely to occur in the simulation of domain evolution. The resulting preservation of domains during simulated evolution is essential for the generation of realistic data that reflects sequences of a protein family sharing functionally important

domains.

With the help of REvolver, we then studied the stability of domains during the course of evolution. We simulate their evolution and assess the number of mutations until a sequence loses its domain specific characteristics and is no longer recognized as an instance of the studied domain. We present a likelihood approach to distinguish between cases where a domain is truly absent in evolutionarily related proteins and cases where it is overlooked in the domain search. Finally, we show how to predict the evolutionary traceability of a protein that serves an estimate whether we expect to find this protein in a specific species or whether the protein is assumed to have accumulated too many substitutions to be identified via sequence similarity based methods.

In summary, we present bioinformatics methods to construct and interpret phylogenetic profiles. The first, and most straightforward approach is the search for orthologs. The functional equivalence of identified orthologs can be validated via the comparisons of their feature architectures. Moreover, for any protein we can estimate over what evolutionary distances we expect to find this protein if it is present. If a protein is expected to be found, however was not found, we conclude that it is indeed missing. In contrast, if a protein is not expected to be found via sequence similarity we complement the construction of phylogenetic profiles via feature architecture similarity based searches for a protein.

To exemplify the application of the introduced approaches and to highlight pitfalls, we investigate the evolutionary ancestry of sex determination and discuss the proposed monophyly of microsporidia and zygomycetes (Lee et al., 2008). This chapter illustrates the versatility and complementarity of ortholog inferences and feature architecture searches, but also their limitations.

Chapter 2

FACT: Functional Annotation Transfer between Proteins with Similar Feature Architectures.

Here, we present the **F**eature **A**rchitecture **C**omparison **T**ool (`http://www.cibiv.at/FACT`) to search for functionally equivalent proteins. FACT uses the similarity between feature architectures of two proteins, i.e., the arrangements of functional domains, secondary structure elements and compositional properties, as a proxy for their functional equivalence. A scoring function measures feature architecture similarities, which enables searching for functional equivalents in entire proteomes. Our evaluation of 9,570 EC classified enzymes reveals that FACT, using the full feature set, outperformed the existing architecture-based approaches by identifying significantly more functional equivalents as highest scoring proteins. We show that FACT can identify functional equivalents that share no significant sequence similarity. However, when the highest scoring protein of FACT is also the protein with the highest local sequence similarity, it is in 99% of the cases functionally equivalent to the query. We demonstrate the versatility of FACT by identifying a missing link in the yeast glutathione metabolism and also by searching for the human GolgA5 equivalent in *Trypanosoma brucei*.

FACT facilitates a quick and sensitive search for functionally equivalent proteins in entire proteomes. FACT is complementary to approaches using sequence similarity to identify proteins with the same function. Thus, it is particularly useful when functional equivalents need to be identified in evolutionarily distant species, or when functional equivalents are not homologous. The most reliable annotation transfers, however, are

achieved when feature architecture similarity and sequence similarity are jointly taken into account.

2.1 Background

The sequencing of entire genomes has become a routine task in molecular biology. To date, about 650 fully sequenced eukaryotic genomes comprising more than 9 Million protein coding sequences are available in the public domain (Hammesfahr et al., 2011). Only a small fraction of these species are model organisms with considerably well characterized protein functions. Most of the remaining species are either of commercial or medical interest, qualify for new model organisms, or hold key positions required for the understanding of organismal evolution. The benefit of a newly sequenced organism essentially depends on the extent to which its data is integrated into existing knowledge about function and evolutionary relationships of genes in other species. A thorough experimental characterization of all proteins is not feasible. Therefore, comprehensive bioinformatics approaches are needed to reliably identify functionally equivalent proteins across species. Two roads are usually followed to accomplish this task.

The first and more common approach searches for proteins with a significant sequence similarity, which is commonly taken as evidence for their common ancestry. For example, a protein with unknown function can be used as query to search for similar sequences in annotated protein databases, e.g., with Blast (Altschul et al., 1997) or, for more sensitive searches, using machine learning algorithms, like PsiBlast (Altschul et al., 1997) or support vector machines (Leslie et al., 2004; Vinayagam et al., 2004; Shah, Oehmen and Webb-Robertson, 2008). The functional annotations of the best hits serve then as tentative annotations for the query (e.g., Quackenbush et al. 2001; Carbon et al. 2009).

Clearly, one limitation is inherent in this approach: Functional equivalence is not tied to a significant sequence similarity. This can have several reasons: First, a query may not obtain a significant hit in a similarity search since the homologous proteins with the same function are too diverged, or are of low complexity. Second, homologs may be identified via sequence similarity but they have assumed different functions (Bartlett, Borkakoti and Thornton, 2003; Kassahn et al., 2009). For example, in the case of enzymes about 60% of sequence identity between homologous proteins is

required to reliably infer functional equivalence (Tian and Skolnick, 2003; Rost, 2002). Thus, a functional annotation transfer between homologs can be wrong. If such an error remains undetected, it can spread through databases (Gilks et al., 2005). Third, it has been shown that proteins with the same function are not always homologous, but rather are a result of convergent evolution (Galperin, Walker and Koonin, 1998). In such instances sequence similarity based searches for functional equivalents produce no results. In summary, functional equivalence is not synonymous with homology.

The second approach to identify functional equivalents does not rely on homology inference by means of pair-wise sequence similarity but rather considers other measures of protein similarity. Amino acid sequences can be annotated with a variety of features, capturing different properties of the protein. Among others, these are functional domains, secondary structure elements and compositional properties. The aggregate of all features in a protein constitutes its feature architecture, and it is supposed that this feature architecture allows conclusions about the function of a protein. A number of studies have shown the applicability of such a feature based approach (e.g., Forslund and Sonnhammer 2008; Hollich and Sonnhammer 2007). The possibility to deduce protein function from the feature architecture suggests that feature architecture similarity can be used to identify proteins sharing a similar function. For example, InParanoid displays the Pfam (Finn et al., 2006) domain annotation of homologous proteins (Ostlund et al., 2010). Thus, we can quickly assess if homologs can be functional equivalents. In the same way, ProteinArchitect (Haimel, Pröll and Rebhan, 2009) finds similar proteins to a query sequence and displays the feature architecture of the hits. However, these tools provide the feature annotation only as an accessory information. The search for similar proteins in the first place is still performed on the amino acid sequence level. The necessity to include information about the feature architecture into the search for functional equivalents was emphasized by Forslund et al. (2008). They showed that roughly 12% of the feature architectures in 96 eukaryotic proteomes evolved more than once independently. Hence, the corresponding proteins are functionally similar although they are not homologous.

Despite its potential for identifying functionally equivalent proteins, only few strategies exploit the feature architecture for similarity searches (Lin, Zhu and Zhang, 2006; Lee et al., 2008; Lee and Lee, 2009). Lin et al. (2006) were the first to measure the similarity between feature architectures using a weighted sum of three indices. The first index measures the ratio of shared features to the total amount of features. The second index assesses the feature duplication similarity, and the third, the Goodman-

Kruskal index, measures to what extent the same feature pairs occur in two proteins. A detailed description of the Lin score is given in the implementation section. Lee and Lee (2009) additionally introduced a weighting scheme that reduces the influence of promiscuous Pfam domains (Basu et al., 2008). Notably, all approaches share the same limitations. Most importantly, feature architectures are constructed only from Pfam domains. Thus, other features such as transmembrane regions or coiled coil domains indicative of protein function are ignored. Furthermore, the positional information of shared features in the compared proteins is not taken into account. Eventually, a systematic evaluation to what extent feature architecture similarity is helpful in detecting functional equivalents is also missing. Lin et al. (2006) and Lee and Lee (2009) evaluated their approaches only for their capability of detecting homologous proteins. Thus, the search for functional equivalents using feature architecture similarity is still in its infancy.

Here, we present FACT a comprehensive method for searching for functionally equivalent proteins using the criterion of feature architecture similarity. FACT considers a broad spectrum of features (functional domains, secondary structure elements, and compositional properties) to determine the feature architecture of a protein. Moreover, the positions of the features in a protein sequence are taken into account. FACT can be used to search for functional equivalents in entire proteomes and the credibility of the best hit is assessed by a p-value. This makes an automated large scale search for functional equivalents possible. A graphical interface, the feature dotplot, complements the automated similarity search and facilitates a visual comparison of two feature architectures. We evaluate the fidelity of FACT using a collection of EC classified enzymes and demonstrate FACT's applicability for identifying functional equivalents. A comparison to the performance of existing approaches to infer functional equivalence from feature architecture similarity, as e.g., described in Lin, Zhu and Zhang (2006) or Lee and Lee (2009), on the same set of enzymes is used to assess the improvement of FACT. Finally, we compare for the first time the usability of two protein similarity measures, sequence similarity and feature architecture similarity, for identifying functional equivalents, and we explore their respective strengths and weaknesses.

2.2 Implementation

As a first step, FACT annotates the query and each protein in the search set with a broad variety of features (figure 2.1A), i.e., functional domains (Pfam domains, SMART domains Letunic, Doerks and Bork (2009), transmembrane regions, signal peptides), secondary structure elements (helix, strand, coiled coils), and compositional properties (low complexity regions, sequence composition). A pipeline of several feature prediction programs serves this purpose. The underlying feature set Φ is, therefore, determined by the collection of prediction programs. The feature architecture of a protein is then the arrangement of instances of features in Φ (figure 2.1B).

2.2.1 Measuring the Similarity of Feature Architectures

To identify proteins with a similar feature architecture to a query protein Q, we measure the pairwise similarity between Q and every protein P in a proteome. We implemented a modified version of the score from Lin et al. (2006) and introduce the $FACT$ score.

Modified Lin Score (MLS)

Lin et al. (2006) score the similarity of two Pfam based feature architectures by combining the Jaccard index, a domain duplication similarity index, and the Goodman-Kruskal index with relative weights 0.36, 0.63, and 0.01, respectively. Calculating the Goodman-Kruskal index requires the order of Pfam domains along the sequence. Our feature set Φ contains a variety of additional features that can overlap in the feature architecture (c.f. figure 2.1B). In such instances, it is unclear how to assess the feature order. However, given its low relative weight of 0.01, the contribution of the Goodman-Kruskal index to the total score is negligible. Thus, we ignored this index in our implementation and adapted the weights of the two other indices accordingly. We calculate the MLS as

$$\begin{aligned} L(P,Q) = 0.365 * \frac{N^{PQ}}{N^P + N^Q - N^{PQ}} + \\ 0.635 * \exp\left(- \sum_{i=1}^{N^P+N^Q} |N_i^P - N_i^Q| / N_{max}\right), \end{aligned} \tag{2.1}$$

where N^{PQ} is the number of shared features between protein P and the query protein Q. N^P and N^Q are the number of different features in P and Q, respectively. N_i^P and N_i^Q count the instances of feature i in P and Q, respectively and

$$N_{max} = \sum_{i=1}^{N^P+N^Q} max(N_i^P, N_i^Q). \tag{2.2}$$

One drawback of the MLS is that it does not include information about the position of individual features in the proteins. Therefore, we introduce a new scoring function.

FACT Score

The *FACT* score computes the feature architecture similarity between proteins as the weighted sum of three scores considering (i) the number of instances for all shared features, (ii) Pfam clan annotations, and (iii) the positions of shared features in the proteins. We describe the three building blocks of the *FACT* score in the following paragraphs.

Feature Multiplicity Similarity (MS): The MS assesses to what extent the numbers of instances for a shared feature agree between two architectures. For each shared feature i, we compute the product of its number of instances in P (N_i^P) and Q (N_i^Q), and normalize this number by the theoretically maximal value $max(N_i^P, N_i^Q)^2$. The MS is then the weighted sum over all shared features.

$$MS(P,Q) = \sum_{i=1}^{N^{PQ}} \omega_i * \frac{N_i^P * N_i^Q}{max(N_i^P, N_i^Q)^2}, \tag{2.3}$$

where $\omega_i > 0$ is the weight for feature i. We use two weighting functions. First, $\omega_i = 1/N^Q$ where $i = 1, ..., N^{PQ}$. This corresponds to an equal weighting of all features in Q. The resulting score is called MS_{uni}. Second, we include the frequency of a feature i from Q in P into the weighting. To this end, N_i^P counts how often feature i from Q is observed in P. The corresponding weight is then

$$\omega_i = \frac{N_i^P}{\sum_{j=1}^{N^Q} N_j^P}, \tag{2.4}$$

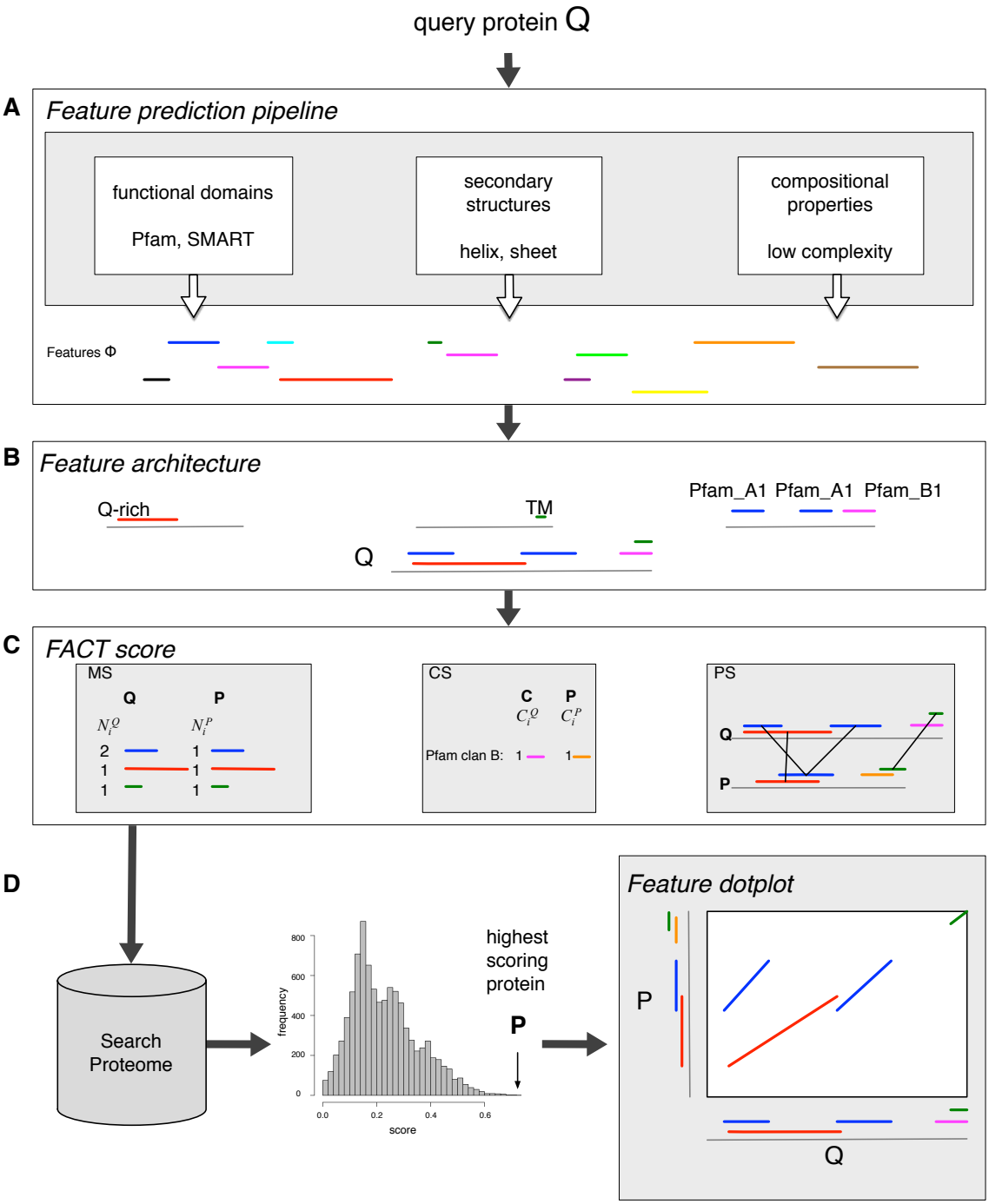


Figure 2.1: Overview of FACT: *(A)* The amino acid sequence of query protein Q serves as input for a collection of prediction programs, which annotate Q with features from Φ. *(B)* The assembly of instances from Φ constitutes the feature architecture of a protein. *(C)* The *FACT* score captures the similarity between two feature architectures by a combination of the *Feature multiplicity similarity* (MS), the *Pfam clan similarity* (CS), and the *Feature positional similarity* (PS). The score is calculated between Q and every protein in a pre-annotated search proteome resulting in a list where the proteins in the search proteome are ranked in decreasing order according to their *FACT* score. *(D)* From the score list any protein P can be extracted and its feature architectures similarity to Q can be visualized in the feature dotplot.

where $i = (1, ..., N^Q)$. This ensures that $\sum_{i=1}^{N^Q} \omega_i = 1$. It is now straightforward to extend this weighting to a set of proteins $\{P_1, P_2, ..., P_l\}$, e.g., a search proteome. We calculate the weight as

$$\omega_i = \frac{\sum_{k=1}^{l} N_i^{P_k}}{\sum_{k=1}^{l} \sum_{j=1}^{N^Q} N_j^{P_k}}. \tag{2.5}$$

We refer to this score as MS_{st}. In the MS_{st}, feature architectures sharing features that are rare in the search proteome receive a higher score than those sharing frequent features. This reflects the intuition that shared rare features are more likely to point towards a similar function than shared frequent features.

Pfam Clan Similarity (CS): Pfam groups similar domains into clans (Finn et al., 2006). For example, the Pfam clan RNase_H (CL0219) consists of 25 domains with a tertiary structure similar to that of Ribonuclease H. This structural similarity implies similarity in the function of the clan members. The CS score takes into account the co-occurrence of different Pfam domains in a clan. It is calculated analogously to the MS_{uni} score.

$$CS(P, Q) = \frac{1}{C^Q} \sum_{i=1}^{C^{PQ}} \frac{C_i^P * C_i^Q}{max(C_i^P, C_i^Q)^2}, \tag{2.6}$$

where C^Q is the number of different Pfam clans in Q, C^{PQ} is the number of shared Pfam clans between P and Q, and C_i^P and C_i^Q are the numbers of instances of clan i in P and Q, respectively.

Feature Positional Similarity (PS): The PS measures the distance between the relative positions a shared feature occupies in the compared proteins. For every instance of a shared feature in P and Q, we first determine the relative position within P and Q. Subsequently, we identify for each instance in Q the instance in protein P having minimal distance. One minus the minimal distance between two feature instances yields a similarity. We calculate PS as following

$$PS(P, Q) = \sum_{i=1}^{N^{PQ}} \frac{\omega_i}{N_i^Q} \sum_{j=1}^{N_i^Q} (1 - \min_{1 \le l \le N_i^P} |q_j - p_l|), \tag{2.7}$$

where the relative position q_j of the j^{th} instance of feature i in protein Q is the center position of this instance divided by the sequence length. The positions p_l of the feature

instances in protein P are calculated accordingly. The use of relative positions ensures that shared features located at the C-terminus in both proteins have a small distance even if the protein lengths are different. The weights ω_i of the individual features correspond to those of the MS_{st}.

The *FACT* score is a weighted linear combination of the *Feature multiplicity similarity* (MS_{st}), the *Pfam clan similarity* (CS), and the *Feature position similarity* (PS) (figure 2.1C).

$$FACT = \alpha * MS_{st} + \beta * CS + \gamma * PS \in [0, 1], \tag{2.8}$$

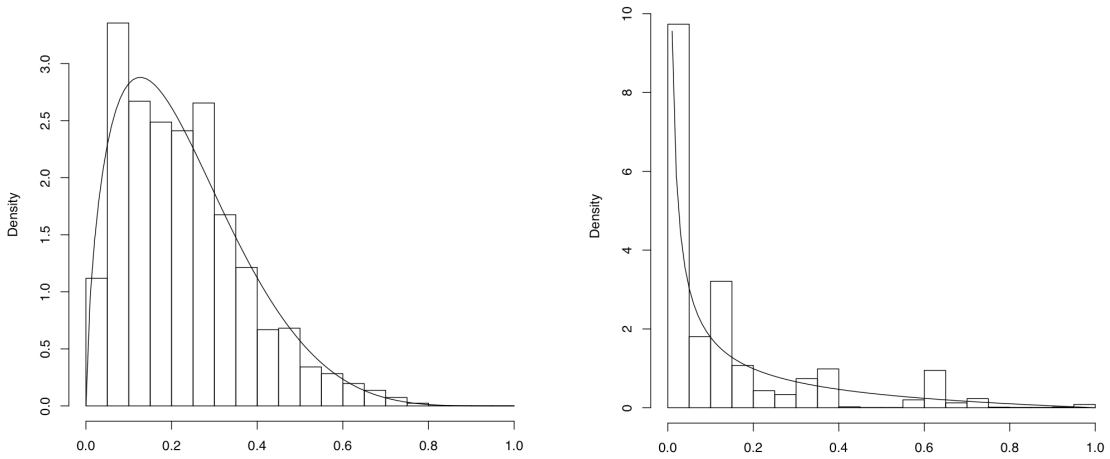
where $\alpha + \beta + \gamma = 1$, and $\alpha, \beta, \gamma \geq 0$.

2.2.2 Score Statistics

Using the scoring functions introduced in the previous section, we calculate the feature architecture similarity scores between a query protein and every protein in a search proteome. From the resulting distribution of scores, we assess the extent to which the top scoring protein stands out from the lower ranking proteins (figure 2.1D). For this purpose, we fit a beta distribution (Kotz, Johnson and Balakrishnan, 2000) to the score histogram. We have chosen the beta distribution for two reasons. First, it can assume different shapes. This fits well with histograms, even when different scoring functions are used (figure 2.2). Second, it is defined in the range from 0 to 1, which is the exact range of the scores. We estimate the two shape parameters of the beta distribution from the mean, and the variance from all scores by the method-of-moments (Kotz, Johnson and Balakrishnan, 2000). The p-value for a score x is then calculated as one minus the cumulative distribution function of the beta distribution of x. The smaller the p-value is, the more pronounced is the feature architecture similarity between the query and the highest scoring protein compared to that of the lower ranking proteins.

2.2.3 Feature Dotplot

For a visual inspection of individual query hit pairs, we have developed the *feature dotplot* (FDP, figure 2.1D), which extends the idea of a classical dotplot to the feature level. The FDP projects the features of two proteins along the x- and y-axis, respectively. A feature occurring in both proteins is represented by a diagonal line in the dotplot, where the slope of the line indicates the length ratio of the features in the



(a) MS_{uni} **(b)** MS_{st}

Figure 2.2: Fit of the beta distribution to the score histograms: Shown are typical score histograms from two FACT searches in the *T. brucei* proteome using the scoring function (a) MS_{uni}, and (b) MS_{st}. Despite the different shapes of the histograms, the beta distribution displays in both cases a good fit to the data.

proteins. Different features are represented by different colors. The standard amino acid dotplot is embedded into the FDP as well.

2.2.4 FACT Webpage

Version 1: FACT is provided online on the webpage `http://www.cibiv.at/FACT_V1/`. The user can search for functional equivalents to a query protein in entire proteomes. The collection in version 1 consists of 26 eukaryotic species (13 animals, 7 fungi, 3 plants, and 3 protists). For every query protein FACT determines the feature architecture. Then the *FACT*, MS_{uni}, MS_{st} and the MLS scores between the query protein and all proteins in a search proteome are computed. For each scoring function the 100 highest scoring proteins are listed and a histogram of all scores is displayed. Additionally, the p-values for the highest scoring protein are shown. The FDP between the query protein and any protein from the score list can be viewed. The FDP links Pfam and SMART domains to the corresponding web pages. Furthermore, possibilities for displaying or hiding specific features, changing the word size for the amino acid dotplot, and for exporting the feature dotplot are provided. Finally, a Blast search

against the search proteome is performed and the best three hits are listed.

As an alternative to the proteome wide search, the FDP can be used to compare two user-defined proteins. The features of both sequences are annotated automatically and displayed in the dotplot.

Version 2: We recently updated the FACT webpage (`www.cibiv.at/FACT`). All pHMM based protein features are now annotated with HMMER Version 3 (`http://hmmer.janelia.org/`). This significantly decreases the time to annotated proteins with features. Furthermore, batch mode searches are enabled, where the user can paste in more than one sequence or upload a multi-fasta file and can also choose more than one search species. All query proteins will then be searched against all chosen species. The new version consists of 37 eukaryotic species (21 animals, 7 fungi, 5 plants, and 4 protists) and a set of experimentally characterized proteins[1]. However, the FACT webpage version 1 is still online and can be reached via the new version.

2.3 Results

FACT has been developed for identifying functionally equivalent proteins. To assess the applicability of our program we require that the tested proteins have their exact function assigned. To the best of our knowledge, only the proteins annotated by the Enzyme Commission (EC) satisfy this condition. The EC provides a hierarchical classification of the reaction catalyzed by an enzyme. The code consists of four numbers separated by dots. The first number determines the main catalyzed reaction (1=Oxidoreductases, 2=Transferases, 3=Hydrolases, 4=Lyases, 5=Isomerases, 6=Ligases), while the last number provides the most specific information about the catalyzed reaction. If two enzymes share the same EC number, they catalyze the same reaction and are therefore functional equivalents. Thus, the EC classifies enzymes according to their function and not according to their evolutionary relationships (c.f., Galperin, Walker and Koonin 1998). We collected EC annotated proteins from human, fly, worm, yeast, and arabidopsis and filtered the dataset such that each EC number is represented at least twice. The final test set is comprised of 9,570 proteins. The average and median numbers of proteins with the same EC number are 10 and 4, respectively.

[1] `http://www.jcvi.org/charprotein-ext/char.cgi/home`

scoring function	# prot (%)
MLS (Eq. 2.1)	7,685 (80.30)
MS_{uni} (Eq. 2.3)	7,712 (80.59)
MS_{st} (Eq. 2.3, 2.5)	7,908 (82.63)
FACT (Eq. 2.8)	8,017 (83.77)

Table 2.1: Fidelity of FACT using different scoring functions. '# prot (%)' denotes the number and percentage of correctly identified functional equivalents in the test set of 9,570 proteins.

2.3.1 Comparison of Different Scoring Functions

For our evaluation, each protein from the test set served as a query for FACT. The similarity scores between the query protein and the remaining 9,569 proteins from the test set were then calculated. Subsequently, we compared the EC number of the highest scoring protein(s) to that of the query. If one highest scoring protein has the same EC number as the query, the proteins are functional equivalents. The fidelity of a scoring function is then the percentage of searches where a functional equivalent gets the best score. Table 2.1 shows the fidelities for the different scoring functions. For the *FACT* score we chose α, β and γ in the ratio 3:1:1 (cf. equation 2.8). The MLS and MS_{uni} display fidelities of around 80%, thus in 20% of the 9,570 searches a protein that is not functionally equivalent to the query obtains the highest score. Weighting the individual features according to their frequency in the test set (MS_{st}) increases the fidelity to 83%. The best result was obtained with the *FACT* score which also takes clan similarity and positional information into account. In 8,017 out of 9,570 cases (84%), a functional equivalent to the query obtained the highest score.

When we analyzed the fidelity with respect to the main reaction catalyzed (first digit of the EC number), a functional equivalent was identified for 9,018 query proteins (94%; *FACT* score).

2.3.2 Relevance of Features

In addition to Pfam and SMART domains, the underlying feature set Φ of FACT includes a variety of other protein features, e.g., secondary structure elements and compositional properties. We next assessed the relevance of including these features. We compared the fidelity of the functional equivalent search using Pfam domains only to the fidelity based on the full feature set. The median number of proteins having

scoring function	Pfam domains # prot (%)	all features # prot (%)
MLS	891 (9.31)	5,618 (59.70)
MS_{uni}	572 (5.98)	5,592 (58.43)
MS_{st}	594 (6.21)	5,792 (60.52)
FACT	-	7,091 (74.10)

Table 2.2: Fidelity of FACT depending on the feature set. '# prot (%)' denotes the number of top ranked functional equivalents with a unique highest score. Since the *FACT* scoring considers clan information it was not used for the calculation with only Pfam domains.

the same best score is 9-13 (depending on the scoring function) for the Pfam only set. This number decreases to 1 for the full feature set. Thus, considering a broad variety of features leads to a better discrimination in the assessment of feature architecture similarity. In contrast, searches using only Pfam domains frequently end up with many equally best scoring proteins representing different EC numbers. For further evaluation, we consider a hit protein only then as an identified functional equivalent when its EC number matches that of the query, and additionally when it is uniquely top ranked in the score list. Table 2.2 shows the results of this analysis. The fidelities for the Pfam only set range, depending on the scoring function, between 6 and 9%. Using the full feature set leads to a drastic increase of the fidelity to values between 58 and 74%.

2.3.3 p-value for FACT Hits

For each highest scoring protein a p-value is calculated. We determined the relation between p-value and the fidelity of FACT using the *FACT* score. With a decreasing p-value, the fidelity increases to a maximum of 98% at a p-value smaller than 10^{-11} (Appendix figure A.1). Considering only those functional equivalents as identified that are uniquely top ranked, the fidelity increases up to 85% at a p-value below 10^{-9}. However, as expected, the increased fidelity comes at the cost of the coverage. For example, of the 9,570 searches only 1,558 have a highest scoring protein with a p-value smaller than 10^{-9} (Appendix table A.1). Our analysis shows that an annotation transfer between the query hit pair becomes more reliable when the p-value is small. Thus, we conclude that the choice of the beta distribution leads to sensible results.

2.3.4 Feature Architecture Similarity vs. Sequence Similarity as a Proxy for Functional Equivalence

With FACT we provide a comprehensive tool to search for functional equivalents using feature architecture similarity. We now compare FACT to the alternative approach that identifies functional equivalents via a significant sequence similarity, e.g., using Blast (Quackenbush et al., 2001; Carbon et al., 2009). Therefore, we run both methods on the test set. To ease the comparison between the two results, we again required a correctly identified functional equivalent to be uniquely top ranked. Figure 2.3 breaks down the results from Blast and FACT (*FACT* score). Blast identified 6,935 (72.5%) functional equivalents compared to 7,091 (74.1%) for FACT. In 5,805 (60.7%) searches both approaches obtained a functional equivalent as highest scoring protein. Moreover, in 4,017 (42%) searches the highest scoring proteins were even identical. 1,286 (13.4%) functional equivalents were detected exclusively by FACT, whereas 1,130 (11.8%) were detected only by Blast. Although FACT performs slightly better than Blast, the large number of functional equivalents found only by Blast indicates that both approaches are complementary. This conjecture is further corroborated by the following observation: When FACT and Blast detect the same protein as best hit, the query hit pair is in 99% of the cases functionally equivalent.

FACT outperforms Blast in situations where sequence similarity between functional equivalents is low. When the E-value exceeds one, the best Blast hit is only in 1% a functional equivalent. For the same query proteins FACT still achieves a fidelity of 31% (Appendix figure A.2). For E-values $\geq 10^{-20}$ the fidelity of Blast increases to 39%, but it is still higher for FACT (46%).

To further explore the complementarity of both approaches we conducted a more detailed analysis. For any query protein in our test set, Blast and FACT each identified a top scoring protein with an associated E-value and p-value, respectively. First we showed that E-value and p-value are not correlated (Pearson correlation coefficient: 0.09). Thus, a query obtaining a Blast hit with a small E-value does not imply a FACT hit with a small p-value, and vice versa. Second, we binned the query proteins according to their E-value/p-value combination. For each combination, we counted the number of query proteins that fall into the bin. Then for each bin we counted how often Blast and FACT identified a functional equivalent. These numbers are represented in the matrix shown in figure 2.4. This matrix gives a guideline under which E-value/p-value combination either Blast or FACT is more likely to find a functional equivalent.

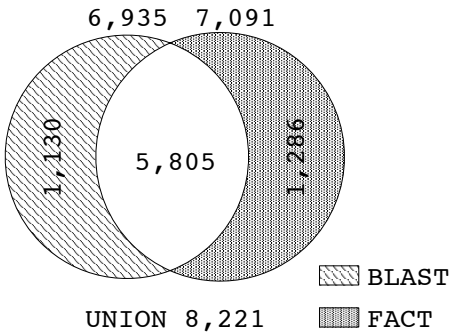


Figure 2.3: Venn diagram contrasting the performance of FACT (*FACT* score) and Blast on the test set: Given are the numbers of uniquely top ranking proteins having the same EC number as the query. For 1,286 (FACT) and 1,130 (Blast) queries, respectively, only one program identified a functional equivalent.

For query proteins obtaining poor Blast hits (E-value > 0.1), the FACT predictions are more credible. A similar picture emerges for queries having a Blast hit with a reported E-value of zero. Once the p-value exceeds 10^{-3}, FACT always identifies more functional equivalents than Blast. Finally, we note that PsiBlast is more sensitive than Blast in detecting even weak sequence similarities that may be indicative of a similar function. We therefore compared FACT also to PsiBlast. This confirmed our findings from the FACT–Blast comparison (Appendix figures A.3 and A.4).

2.3.5 Example Applications of FACT

To illustrate the versatility of FACT in searching for functional equivalents we discuss two examples. The general procedure of a FACT search is summarized in figure 2.1.

Missing Link in the Glutathione Metabolic Pathway

A common task in comparative genomics is the identification of proteins that are involved in known metabolic pathways in different species. As of today, the evolutionary relationships between proteins are usually used for this purpose, e.g., Kanehisa et al. (2008). In some cases however, orthologs to individual proteins cannot be identified. Consequently, the question is raised of whether the corresponding functional equivalents are not present in the respective species or whether sequence similarity based searches cannot detect them. The glutathione metabolic pathway in the KEGG database (Kanehisa et al., 2008) constitutes one illustrative example. It is one of the

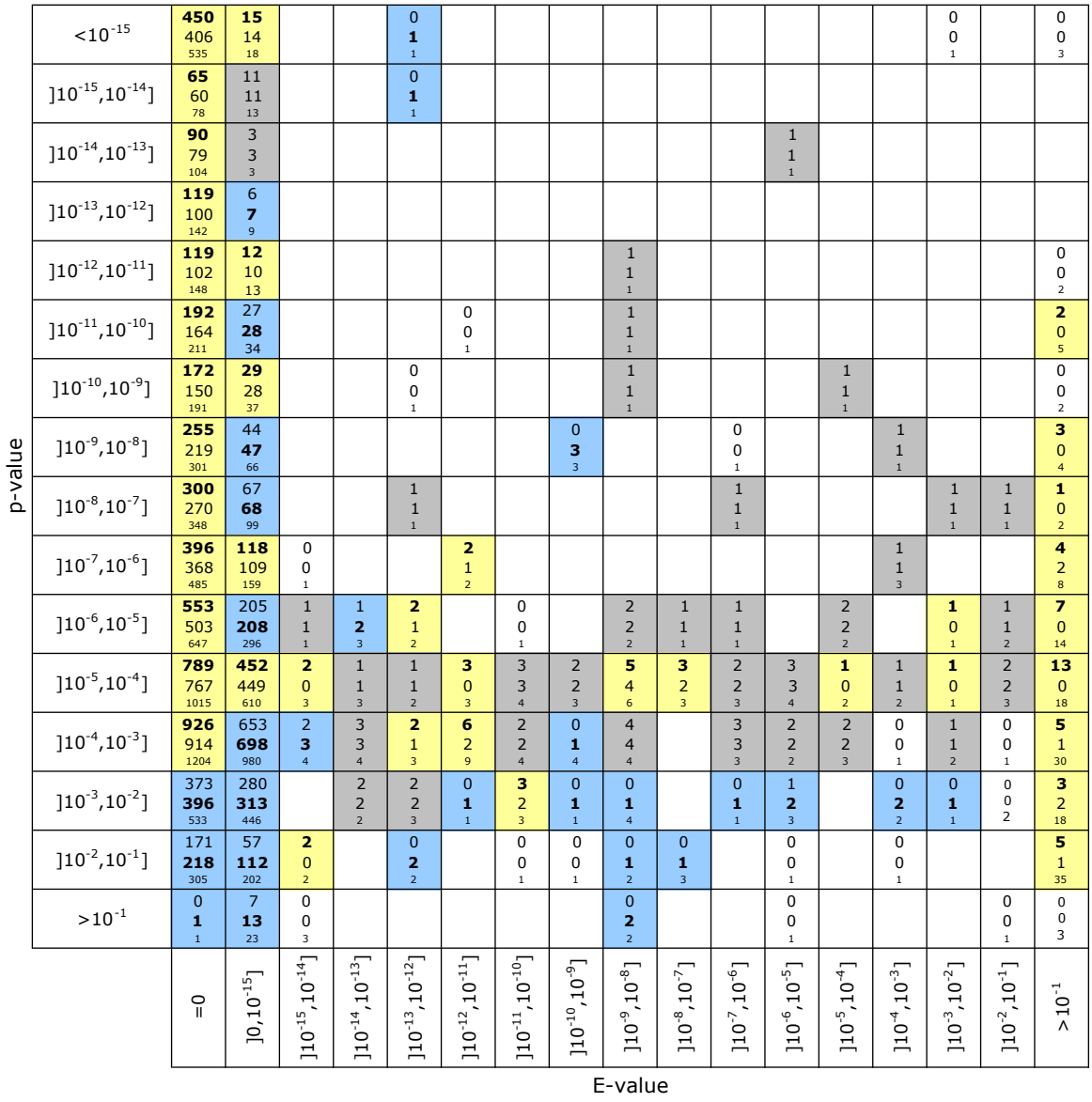


Figure 2.4: Contrast of Blast and FACT (*FACT* score) for different E-value/p-value combinations: The matrix bins the 9,570 proteins according to the E-value and the p-value of the best hit when used as query for Blast and FACT, respectively. The total number of proteins for a E-value/p-value combination is given by the bottom number in the corresponding cell. The two further numbers in a cell give the number of searches FACT (top) and Blast (middle) had a functional equivalent as top scoring protein. The number for the better performing tool is given in bold face. Yellow cells show E-value/p-value combinations where FACT identified more functional equivalents than Blast, whereas the blue cells indicate a higher fidelity of Blast. Grey cells mark ties.

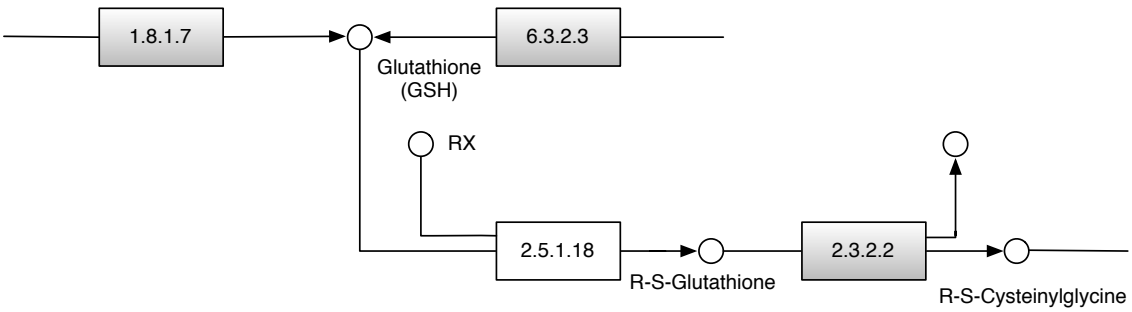


Figure 2.5: Section of the KEGG glutathione metabolic pathway (ko00480): Grey filled boxes represent proteins of the human pathway for which KEGG orthologs exist in *S. cerevisiae*. An ortholog to the human glutathione S-transferase (EC 2.5.1.18), a central component of this pathway, could not be identified in yeast.

central detoxification pathways in animals and fungi. An ortholog to the human glutathione S-transferase (EC number 2.5.1.18), a central enzyme in this pathway, is not annotated in the yeast genome. However, orthologs to the human proteins flanking the glutathione S-transferase in the pathway are present (figure 2.5).

A Blast search using the human glutathione S-transferase protein as query revealed no significant hits in the yeast proteome. The best Blast hit (YNL286W; E-value = 0.51) has no feature in common with the human query protein except α-helices and β-sheets. Instead, it contains two RNA recognition motifs (RRM_1). Similarly, the best PsiBlast hit (YCL009C; E-value = 1.3) has no feature in common with the human query protein except α-helices and β-sheets. Next, we performed a FACT search in the yeast proteome, again with the human enzyme as query. This revealed the same best hit for all scoring functions (YLL060C; *FACT* score: p-value= $3 * 10^{-6}$). Thus, from the corresponding E-value/p-value entry in figure 2.4 ($> 10^{-1}$ / $]10^{-6}, 10^{-5}]$), there is a 50% chance of having detected a functional equivalent. We next used the FDP of the FACT hit and the query protein to validate the candidate (figure 2.6). Both proteins have the N-terminal and the C-terminal glutathione S-transferase (GST) domains and share a predicted transmembrane region. Therefore, we conclude that the two proteins are functional equivalents. This has indeed been confirmed, since both proteins have been annotated with the same EC number (Choi, Lou and Vancura, 1998). Thus, FACT helps to identify candidate proteins that may close gaps in biochemical pathways.

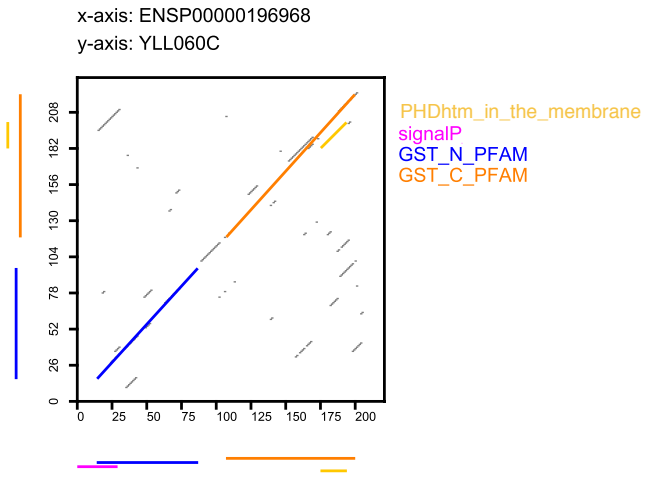


Figure 2.6: Feature dotplot of the human glutathione S-transferase and the best FACT hit in yeast: The two proteins share the Pfam domains GST_N (PF02798) and GST_C (PF00043), as well as a transmembrane domain. A signal peptide (signalP) is present only in the human protein. For better readability α helix and β sheet annotations are not shown.

Functional Equivalents for GolgA5

In our second example we focused on a structural protein, GolgA5, which is important for assembling and maintaining the structure of the human golgi apparatus (Diao et al., 2003; Satoh et al., 2003). Almost the entire protein is made up of coiled coils. This structure is formed by low complexity repeat units consisting of hydrophobic and polar residues. Consequently, many different sequences can assume a coiled coil structure. Thus, a sequence similarity based search for functional equivalents in very distantly related species is likely to be not successful. We performed a FACT search with the human GolgA5 in *Trypanosoma brucei*. The highest scoring protein agrees only between the MS_{st} and the *FACT* score (Tb11.02.5040), while two different proteins were identified by the MS_{uni} (Tb11.02.4670) and the MLS (Tb927.5.1900). For that reason, all top ranked hits, the best Blast hit (Tb11.52.0008), and best the Psi-Blast hit (Tb927.7.3330) were analyzed with the FDP (see Appendix figure A.5-A.8. The FDP of the PsiBlast hit is not shown since this protein is 4,334 aa in length.) Since the function of GolgA5 requires its anchoring in the plasma membrane, we curated the results according to the presence of a transmembrane region. Among all hits, Tb11.02.4670 (MS_{uni}) is the only protein that shares a C-terminal transmembrane region with the human GolgA5. Thus, we consider it to be the most promising candidate

for the GolgA5 functional equivalent in *T. brucei*. Notably, it was recently shown that this protein exerts the same function in *T. brucei* as GolgA5 does in humans (Ramirez et al., 2008).

2.4 Discussion

Here we present FACT, a tool for searching for functionally equivalent proteins. FACT computes the pairwise similarity between feature architectures and identifies for a query protein the highest scoring hit in an entire proteome. Evaluating the performance of FACT on EC classified enzymes reveals a fidelity of 84%.

How to measure the similarity between feature architectures still remains an open question. So far, all suggestions are ad-hoc solutions to the scoring problem. For example, the Lin score (Lin, Zhu and Zhang, 2006) assesses the similarity between two proteins from their features in common and also considers the set difference. Thus, features that are not shared between two proteins reduce the score. This scoring appears reasonable when feature architectures consist only of functional domains, e.g., Pfam domains. In such cases, the presence of an extra feature in one protein is likely to also reflect a functional difference between the compared proteins. However, in our study we used a comprehensive feature set, where some features lack an obvious connection to function. Therefore, we introduce a new score that considers only shared features. Our evaluation on a set of EC classified enzymes reveals that the fidelity in identifying functional equivalents does not heavily depend on whether or not the feature set difference between two proteins is taken into account. Both scoring functions, the MLS and MS_{uni} perform equally well. Their conceptual difference, however, becomes relevant in individual cases as shown by our GolgA5 example application. The best scoring protein according to the MLS shares 4 features with the query and has one extra feature (Appendix figure A.5). In contrast, MS_{uni} identifies a highest scoring protein that shares 5 features with the query but has 4 extra features (Appendix figure A.6).

The idea of giving individual features different weights has been presented before. Lee and Lee (2009) weight a Pfam domain depending on its frequency in the RefSeq database (Pruitt, Tatusova and Maglott, 2007) and on the diversity of its flanking Pfam domains. Note that the latter criterion is not straightforward to implement when features can fully overlap, and hence, feature order cannot be determined. In the MS_{st},

we weight a feature according to its inverse frequency in the search proteome. This weighting scheme can be applied to any feature, and takes into account that feature frequencies can vary between search proteomes. The comparison of MS_{uni} and MS_{st} reveals that the introduction of weighting increases the fidelity by 2%. Unfortunately, comparing the effect of our weighting to that of Lee and Lee (2009) is impossible, since in their evaluation the scoring functions differed not only in the weighting but also in the way shared domains are scored.

Among all scoring functions, the *FACT* score performs best (table 2.1 and 2.2). This is the consequence of including clan similarity and positional similarity. We compute the *FACT* score by combining the scoring functions MS_{st}, CS, and PS in a ratio of 3:1:1. We consider the number of shared features and their number of instances to be the most important parameters in determining the similarity between feature architectures. The clan annotation as well as the position of features are supplementary information that only have a moderate influence on the final score. Note that we deliberately did not optimize the weight parameters α, β, and γ with respect to the fidelity on the EC based functional annotation. Enzymes cover only a fraction of the diversity of protein functions. We wanted to avoid a bias towards this particular class of proteins, which could interfere with the general applicability of FACT (Boulesteix, 2010).

In contrast to existing tools that use Pfam domains for identifying functionally similar proteins (Lin, Zhu and Zhang, 2006; Lee and Lee, 2009), FACT recruits a diverse set of features for building the feature architectures. Our evaluation highlights the significance of using a comprehensive feature set. When considering only Pfam domains, the median number of equally best scoring proteins is 9-13, depending on the scoring function. The most extreme case comprises the 589 enzymes lacking any Pfam domain. When these proteins are used as query, all proteins in the search proteome obtain the same score. However, the median number of enzymes with the same EC number as the query is only 3. Consequently, in the vast majority of searches more than one EC number is represented by the top ranked proteins. The search result is therefore ambiguous. To facilitate a meaningful assessment of the fidelity, we required a correctly identified functional equivalent to be uniquely top ranked. This reveals a maximal fidelity of 9% (table 2.2). In contrast, when we use the FACT feature set, the median number of equally best scoring proteins reduces to one. This shows that the similarity score becomes more discriminative when more features are considered. As a consequence, the fidelity raises to 74% (*FACT* score). Notably, for the proteins

without Pfam domains a correct functional equivalent was still identified in 158 cases.

There is still room for improvement regarding the search for functional equivalents. So far, all approaches are based on ad-hoc solutions for measuring the similarity between feature architectures since modeling their evolution is still an open problem. Moreover, the function of a protein essentially depends on its tertiary structure. However, tertiary structure elements are not yet part of the feature set. Both the integration of evolutionary models and of complex features is likely to result in more sensible similarity measures.

Feature architecture similarity based approaches identify functional equivalents. This supposedly complements sequence similarity based approaches represented, e.g., by Blast or PsiBlast. Here we have compared the fidelity of FACT to that of Blast. A substantial fraction of functional equivalents were top ranked exclusively by FACT. This includes the cases where sequence similarity was too low to result in a significant Blast hit, but FACT still detected functional equivalents. Finally, we observed no linear correlation between the E-value of the best Blast hit and the p-value of the best FACT hit for a given query. In summary, these results confirm the complementarity of feature architecture similarity based approaches and sequence similarity based approaches in the search for functional equivalents. This finding is independent of whether we used Blast or PsiBlast. The complementarity is further corroborated by those searches where FACT and Blast identify the same best hit. In such instances, the fidelity increases to 99%. Thus, a joint application of a feature architecture measure and a sequence similarity measure allows for highly reliable automated functional annotation transfers. However, this increase of the fidelity comes at the cost of detecting only 42% of the present functional equivalents in our test data.

In cases where the two approaches disagree, we need to decide which of the hits is more likely to be a functional equivalent. To facilitate this decision, we have compared the fidelities of Blast/PsiBlast and FACT depending on the E-value and p-value of the highest scoring protein for a given query (c.f. figure 2.4, Appendix figure A.4). Notably, for searches where both methods obtained a good hit, i.e., small E-value and small p-value, respectively, FACT finds a functional equivalent more often than the other program. However, in most cases, a decision of whether a FACT hit that is not confirmed by Blast, or vice versa, is a functional equivalent will require manual curation. We have presented two examples where we searched for functional equivalents to the human glutathione S-transferase in yeast, and to the human GolgA5 in *T. brucei*.

These examples showed that the feature dotplot is a versatile tool to curate results from FACT searches. The feature dotplot facilitates an educated judgment of how similar two feature architectures are, and how likely it is that the corresponding proteins are functionally equivalent. Together with the implementation of four different scoring functions and the Blast search, the feature dotplot complements the toolbox for a comprehensive search for functional equivalents.

2.5 Conclusion

FACT uses the similarity of feature architectures between two proteins to search for functional equivalents in entire proteomes. FACT has a high fidelity and outperforms existing approaches that identify functional equivalents based on the presence of PFAM domains. This increase in fidelity is mainly accomplished by using a diverse set of features that are recruited for building the feature architectures. The different weighting of individual features and the relative position of shared features in the compared proteins provide additional information. FACT complements sequence similarity based approaches, such as Blast, in the search for proteins with an equivalent function. It is, thus, particularly useful when distantly related species with highly diverged sequences are analyzed, or in cases where functional equivalents are not homologous. Both aspects will become increasingly relevant the more genome data from 'exotic' species becomes available. However, there exists no globally optimal solution to the problem of identifying functionally equivalent proteins. It is therefore necessary to compare the results from different scoring functions measuring feature architecture similarity and from sequence similarity based searches to select the most promising functional equivalent candidates. The feature dotplot to visually inspect the feature architectures of two proteins facilitates this manual curation. We have demonstrated the joint use of FACT, Blast, and the feature dotplot in a comprehensive search for functional equivalents in two example applications. They serve as a guideline of how to use these tools to propagate existing knowledge about the function of proteins from one species to another.

2.6 Methods

FACT

FACT annotates functional domains, secondary structure elements and compositional properties in protein sequences using the tools in the Sfinx package (Sonnhammer and Wootton, 2001). Low complexity regions are identified with the program *seq*, helices and strands with the program *PHDseq*, coiled coils with the program *COILS2*, and signal peptides with the program *SignalP*. Transmembrane regions are predicted both with *TMHMM* and *PHDhtm*. Pfam (version 23; Finn et al. 2008) and SMART (smart_16_04_2008; Letunic, Doerks and Bork 2009) domains are identified with HMMER2 and with HMMER3 in the new version (`http://hmmer.janelia.org/`). Regions enriched for a particular amino acid are annotated with CAST (Promponas et al., 2000). Pfam clan information was downloaded from `http://pfam.sanger.ac.uk/`. All annotation results are transformed into the SFS format (Sonnhammer and Wootton, 2001). This data structure allows for an easy extension of the feature set with features currently not considered by FACT. For sequence similarity searches Blast version 2.2.13 and PsiBlast version 2.2.23 was used. PsiBlast searches were run with default parameter settings using 5 iterations. The FDP is implemented as a Java applet requiring Java 1.5 or higher. It can be accessed with a web browser with Java and with Java script enabled.

Test Set

We compiled the test set for the FACT evaluation using an initial collection of 9,897 EC annotated enzymes from *Homo sapiens* (6,339), *Arabidopsis thaliana* (1,156), *Saccharomyces cerevisiae* (1,099), *Drosophila melanogaster* (896) and *Caenorhabditis elegans* (407). Protein sequences were downloaded from Ensembl 52 (*D. melanogaster, C. elegans, S. cerevisiae*), Ensembl 51 (*H. sapiens*) and UniProt 1.0 (*A. thaliana*). The associated EC annotations were retrieved from Ensembl and UniProt. From this set we removed all proteins that were annotated with more than one EC number or with partial EC numbers. Subsequently, we discarded all EC numbers and associated proteins which were present only once in the protein collection. The final test set consists of 9,570 proteins representing 1,016 different EC numbers.

Data

Proteome data for *Trypanosoma brucei* was obtained from the Sanger Center (`http://www.sanger.ac.uk`). The human glutathione S-transferase was identified in the glutathione metabolic pathway in the KEGG database at `http://www.genome.jp/kegg/pathway/map/map00480.html`. The human protein ENSP00000196968 (Ensembl 51) was used as query for the FACT search in the yeast proteome. For the GolgA5 search, the human protein ENSP00000163416 (Ensembl 51) was used as query for the FACT search in the *T. brucei* proteome.

2.7 Availability and Requirements

Project name: FACT

Project home page: http://www.cibiv.at/FACT

Operating system: Platform independent

Programming language: Java

Other requirements: Java 1.5 or higher, java script enabled

Chapter 3

REvolver: Modeling Sequence Evolution under Domain Constraints

Simulating the change of protein sequences over time in a biologically realistic way is fundamental for a broad range of studies with a focus on evolution. It is, thus, problematic that typically simulators evolve individual sites of a sequence identically and independently. More realistic simulations are possible, however they are often prohibited by limited knowledge concerning site-specific evolutionary constraints or functional dependencies between amino acids. As a consequence, a protein's functional and structural characteristics are rapidly lost in the course of simulated evolution.

Here we present REvolver (`www.cibiv.at/software/revolver`), a program that simulates protein sequence alteration such that evolutionarily stable sequence characteristics, like functional domains, are maintained. For this purpose, REvolver recruits profile hidden Markov models (pHMMs) for parameterizing site-specific models of sequence evolution in an automated fashion. pHMMs derived from alignments of homologous proteins or protein domains capture information regarding which sequence sites remained conserved over time and where in a sequence insertions or deletions are more likely to occur. Thus, they describe constraints on the evolutionary process acting on these sequences. To demonstrate the performance of REvolver as well as its applicability in large-scale simulation studies, we evolved the entire human proteome up to 1.5 expected substitutions per site. Simultaneously, we analyzed the preservation of Pfam and SMART domains in the simulated sequences over time. REvolver preserved 92% of the Pfam domains originally present in the human sequences. This value drops to 15% when traditional models of amino acid sequence evolution are used. Thus, REvolver represents a significant advance towards a realistic simulation

of protein sequence evolution on a proteome-wide scale. Further, REvolver facilitates the simulation of a protein family with a user-defined domain architecture at the root.

3.1 Introduction

Molecular sequences change over time and their rate and pattern of sequence change are influenced by a variety of different parameters, such as mutation rate or functional and structural constraints. Simulating the evolution of biological sequences is therefore a trade-off between simplifying assumptions to reduce complexity of the problem and biological reality. Several programs exist to simulate the evolution of proteins along a phylogenetic tree (Rambaut and Grassly, 1997; Stoye, Evers and Meyer, 1998; Fletcher and Yang, 2009; Strope et al., 2009). All either start with a user-provided sequence or generate a random sequence at the root. Seq-Gen (Rambaut and Grassly, 1997) simulates the evolution of the root sequence only by substitutions and does not consider insertions and deletions. ROSE (Stoye, Evers and Meyer, 1998) was the first program to close this gap by also modeling the insertion and deletion process. By default, both programs assume that sites evolve independently and identically. While this is a fair assumption for sequences not assuming any structure or exerting any function, it is an obvious oversimplification when it comes to the simulation of sequence change in functional sequences, such as genes or gene products. As a result, relevant sites that remain unchanged over considerable evolutionary distances in real sequences may be altered by a simulator after only a few simulation steps. To cope with this problem, both programs consider substitution rate heterogeneity by randomly assigning rate scaling factors to individual sequence positions. While this is valid for random root sequences, it is not when the evolution of real protein sequences should be simulated. In such cases it cannot be avoided that a functionally relevant site, which is unlikely to change over time, is assigned a high substitution rate by chance. Even more problematic is the modeling of insertion and deletion events (indels) that are typically placed randomly in a sequence by the simulator. Moreover, indel lengths are often drawn from a single distribution. However, a biologically meaningful simulation requires that the placement of indels be guided by information, about where in a sequence insertions or deletions can be tolerated and where they likely interfere with the protein's function. Since the spacing between interacting amino acids in the native structure of a protein is important, the length distribution of indels may also vary between individual positions of a sequence (see Laity, Lee and Wright 2001). INDELible (Fletcher and Yang,

2009), SIMPROT (Pang et al., 2005), and indel-Seq-Gen (iSG; Strope et al., 2009) represents major steps towards more realistic simulation. These programs facilitate the manual assignment of different evolutionary parameters to specific segments of the sequence. This enables explicit differentiation between evolutionary constraints acting for example on functional protein domains, and those acting on intervening linker regions. Despite this progress, two major limitations remain. First, it is not feasible to use these programs in large scale studies where the evolution of hundreds or thousands of protein sequences is simulated, as there is no automatized procedure to extract meaningful constraints. Second, there is no standard operating procedure for inferring evolutionary constraints. This opens the door for ad-hoc decisions that may later be hard to justify or reproduce. Considering sequence structure is an obvious solution for the second problem. The emergence of fast algorithms capable of evaluating the effects of mutations on the structure of the protein facilitated the development of programs integrating structural consequences of individual mutations into the simulation (e.g. Parisi and Echave, 2001; Rastogi, Reuter and Liberles, 2006; Lakner et al., 2011; Grahnen et al., 2011; Grahnen, Kubelka and Liberles, 2011). Unfortunately, for the vast majority of sequences the relevant information for deriving evolutionary constraints, i.e. the structure, is not available. Moreover, predicting the exact effect of individual mutations on structure and function of a protein, and extrapolating this to the evolutionary behavior of individual sites of a protein is still hard. This limits a wide use of structure-informed constraints in simulated sequence evolution.

Here, we suggest a pragmatic approach to achieve a biologically meaningful simulation of sequence evolution. Homologous sequences have been evolving independently since they last shared a common ancestor. The comparison of such sequences reveals sites that remain entirely conserved over time, sites displaying only a subset of the amino acid alphabet, as well as sites that appear to be free to change. Moreover, it indicates the preferred positions of insertions and deletions, as well as their respective length distributions. This pattern of sequence conservation and alteration represents the footprint of a constrained evolutionary process acting on these sequences. In principle, databases such as PFAM (Finn et al., 2010) or SMART (Letunic, Doerks and Bork, 2009) provide exactly this information. They have been specialized in the collection and alignment of homologous protein sequences or protein domains and describe the characteristics of the resulting alignments by a profile hidden Markov model (pHMM; figure 3.1). In these models, site-specific emission probability vectors reflect the frequencies of the 20 amino acids at the corresponding positions in real instances of

the modeled domain. Similarly, the models provide site-specific insertion and deletion probabilities. Unfortunately, it is not straightforward to exploit this information for the simulation of sequence evolution. Traditionally, pHMMs are defined as generative models that produce instances of a domain or protein family rather than modeling its change. Consequently, time is not considered in the pHMM formalism. Our new simulator, REvolver, solves this problem by implementing the following key features:

- Emission probabilities of the pHMM are used as site-specific amino acid equilibrium frequencies in the substitution model.
- Insertions and deletions are placed preferentially at positions where they have been already observed in real instances.
- REvolver corrects for the formation of artificially large insertions due to repeated nested insertions.
- Evolution acts on the amino acid sequence AND on the relationship between the amino acids sequence and the constraints. Hence, the information about site-specific evolutionary constraints is maintained throughout the simulation.
- A mechanism counterbalancing the erosion of characteristic sites prevents a simulated sequence from losing its identity as a domain instance.

3.2 The Simulator

In the following sections we describe the general procedure to simulate the evolution of a domain along a phylogenetic tree. The ancestral evolutionary instance consists of the amino acid sequence together with its state path through a pHMM. A typical pHMM is depicted in figure 3.1. It consists of match states (M), insertion states (I), deletion states (D), and a *Begin* state, and an *End* state. States are connected via transitions, where each transition has its individual transition probability (P). Match states and insertion states emit amino acids according to an emission probability vector $E = (e_1, ..., e_{20})$ for the 20 amino acids. A random path through a pHMM starts at the *Begin* state, passes through match, insertion and deletion states, and terminates at the *End* state. By that, an instance of the modeled domain/protein is generated. The resulting state path represents the relationship between the constraints and the specific amino acid positions.

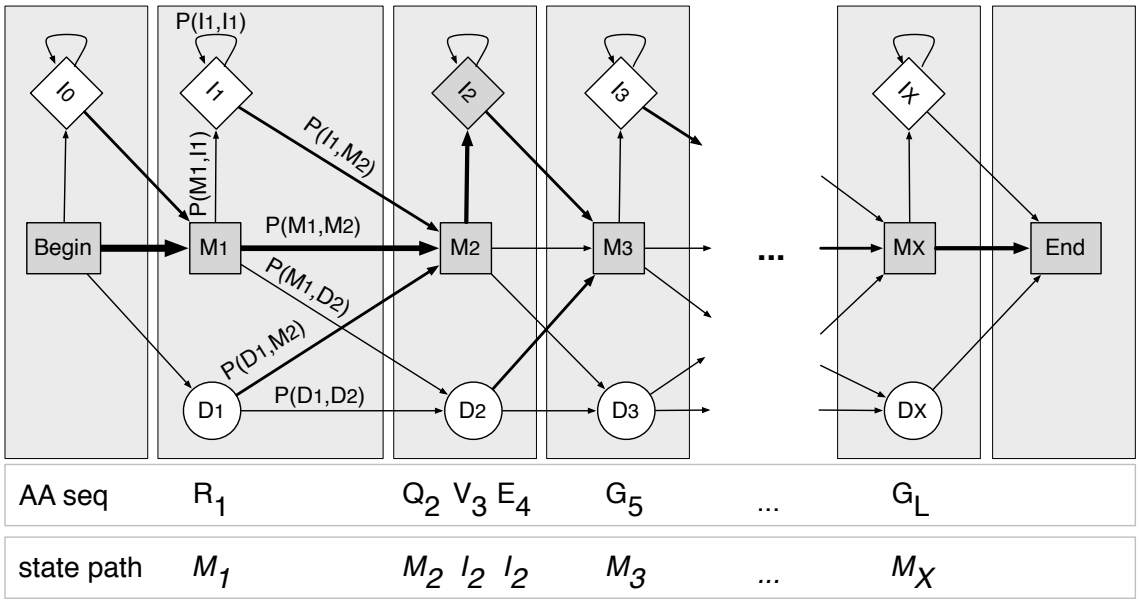


Figure 3.1: Structure of a pHMM: The pHMM comprises match states (M_x), insertion states (I_x), deletion states (D_x), a *Begin* state, and an *End* state. The index x ranges from 0 to X, where X is the length of the pHMM. Since states (M_x, I_x, D_x) of the same position x (except for position 0 and X) have together 7 transitions to states $x+1$, the model is called Plan7. Arrows indicate transitions between individual states, where the line weight is proportional to the transition probability $P(State_x, State_y)$. The amino acid sequence is indexed from 1 to the sequence length L. The respective states in the corresponding state path are shaded in grey in the pHMM. The sequence `RQVEG...G` are amino acids emitted from match (`RQGG`) and insertion (`VE`) states.

Starting at a node in a phylogeny, the parent instance evolves along a branch leading to a child instance. Mutations result in changes in the amino acid sequence, but can also alter the state path. Thus, the state path must evolve with the sequence. The procedure for the simulated evolution on one branch is repeated for each branch in the tree, resulting in protein sequences on the leaf nodes sharing a common ancestry, together with their state paths. Next, we explain the realization of the individual mutations (substitutions, insertions, and deletions) with and without domain constraints. Note, that in the context of this manuscript, we partition a protein sequence into domains and linker regions. We refer to a domain as a segment of a protein that is modeled by a pHMM and refer to the remainder of the protein as linker sequences. In our simulations, domain regions evolve under constraints inferred from the pHMM, whereas linker regions evolve free of constraints. If a protein contains more than one segment, we perform the simulation on each segment separately. REvolver is based on Plan7 pHMMs produced by the program hmmscan from the HMMER3 software package (`http://hmmer.janelia.org/`; cf. figure 3.1).

3.2.1 Simulation Procedure

To simulate substitutions, insertions, and deletions, we apply the Gillespie algorithm (Gillespie, 1977) as outlined in Algorithm 1. Substitutions are described by a continuous-time Markov chain that is characterized by a matrix Q of instantaneous rates q_{ij}, where q_{ij} is the product of the relative rate of substitution ρ_{ij} from amino acid i to amino acid j, and the amino acid frequency π_j. Currently, 14 standard amino acid substitution models are implemented in REvolver (table 3.1). In addition, REvolver can use any user-defined substitution model composed of the relative rate matrix $R = \{\rho_{ij}\}$ and the equilibrium frequencies π_j. The substitution rate for any amino acid i is given by $q_i = \sum_{j \neq i} q_{ij}$. Finally, the total substitution rate $\Lambda_S = \sum_{l=1}^{L} q_{i_l}$, where L is the sequence length and i_l the amino acid at position l of the sequence. In addition to substitutions, we simulate insertions and deletions. The rates for insertions λ_I and deletions λ_D are independent from each other. Since a sequence of length L has L possible positions to start a deletion, the total deletion rate is $\Lambda_D = L\lambda_D$. Insertions can occur before the first amino acid and after every amino acid. Consequently, the total insertion rate is $\Lambda_I = (L+1)\lambda_I$. The insertion position at the very beginning of a sequence is considered to be an immortal link (Thorne, Kishino and Felsenstein, 1991). Thus, an insertion can occur even when all amino

Substitution model	Reference
JTT	Jones, Taylor and Thornton 1992
JTT_dcmut	Kosiol and Goldman 2005
Dayhoff	Dayhoff, Schwartz and Orcutt 1978
Dayhoff_dcmut	Kosiol and Goldman 2005
WAG	Whelan and Goldman 2001
mtMAM	Yang, Nielsen and Hasegawa 1998
mtART	Abascal, Posada and Zardoya 2007
mtREV	Adachi and Hasegawa 1996
cpREV	Adachi et al. 2000
Vt	Müller and Vingron 2000
Blosum62	Henikoff and Henikoff 1992
LG	Le and Gascuel 2008
HIVb	Nickle et al. 2007
HIVw	Nickle et al. 2007

Table 3.1: Standard protein evolution models implemented in REvolver.

acids were deleted in a previous step. Note that, $\Lambda_I = (L_1 + 1)\lambda_I$ applies only to the first segment. The total insertion rate for the remaining segments is $L_n\lambda_I$, where L_n is the length of the n^{th} segment, $n > 1$. Eventually, we set the total event rate $\Lambda = \Lambda_S + \Lambda_I + \Lambda_D$.

To simulate the evolutionary process along a branch of a tree (cf. Algorithm 1), we divide the branch into a number of time steps that are exponentially distributed. To this end, we draw a 'waiting' time t_w from an exponential distribution with mean $1/\Lambda$ during which exactly one mutation occurs (von Haeseler and Schöniger, 1998). t_{rem} is the remaining time, initialized with the branch length t. If t_w is smaller than or equal to t_{rem}, a mutation occurs. We next choose according to Λ_I, Λ_D, and Λ_S whether an insertion, deletion, or a substitution should occur. The sequence and the state are then changed, and we update Λ as follows: If the event was an insertion or deletion, we adjust the sequence length L by adding or subtracting the length of the insertion or deletion, respectively, and recalculate Λ_I and Λ_D accordingly. An important property of REvolver is that once a sequence has been inserted it undergoes the same evolutionary process as the root sequence, i.e. it can be substituted, deleted, and the insertion can be extended. If a substitution occurred in which amino acid j replaced amino acid i, we exchange q_i by q_j to update Λ_S. Finally, we set $t_{rem} = t_{rem} - t_w$, draw a new t_w from the exponential distribution with the updated parameter Λ, and repeat until $t_w > t_{rem}$.

This general procedure is used for all sequences. The specific details of simulating unconstrained and constrained sequences are described in the next sections.

Algorithm 1 Outline of the simulation procedure

$\Lambda \leftarrow \Lambda_S + \Lambda_I + \Lambda_D$
$t_{rem} = t$
$t_w \sim Exp(\Lambda)$
while $t_w \leqslant t_{rem}$ **do**
  $randomVariable \sim Uniform()$
  if $randomVariable \leqslant \Lambda_I/\Lambda$ **then**
    doInsertion()
  else if $randomVariable \leqslant (\Lambda_I + \Lambda_D)/\Lambda$ **then**
    doDeletion()
  else
    doSubstitution()
  end if
  Λ = updateEventRate()
  $t_{rem} \leftarrow t_{rem} - t_w$
  $t_w \sim Exp(\Lambda)$
end while

3.2.2 Evolutionary Events for Unconstrained Segments (Linker)

In the following we describe the simulation of substitutions, insertions, and deletions for unconstrained segments, where the evolutionary instance is simply the amino acid sequence.

Substitutions REvolver simulates the substitution process in unconstrained segments based on a substitution model Q plus a parameter r that encodes variation in rate across sites (RAS). The substitution rate at a given site l is, thus, calculated as $q_i r_l$, where r_l is a rate scaling factor and i is the current amino acid at site l. We provide three types of RAS models, where r_l is always independently and identically distributed among sites: the scaling factor is (i) the same at all sites (default), (ii) drawn from a continuous gamma distribution, and (iii) drawn from a discrete gamma distribution. Both gamma distributions have a mean of 1 and shape parameter α. In the case of rate heterogeneity ((ii) and (iii)), rate scaling factors are assigned to each position l in the root sequence. Child nodes inherit the scaling factors from their

parent node. Newly inserted positions receive a scaling factor from this gamma distribution. The sequence site l where the substitution occurs is chosen proportional to its substitution rate $q_i r_l$. The probability that amino acid i is substituted with amino acid j is proportional to q_{ij}/q_i for $i \neq j$ (Karlin and Taylor, 1975).

Insertions and Deletions Insertion and deletion positions are distributed uniformly along the unconstrained segments. To determine the length of an individual insertion or deletion, we draw a value from a probability distribution. Currently, we have implemented the geometric distribution and the Zipfian distribution (Benner, Cohen and Gonnet, 1993; Chang and Benner, 2004). The parameters for the distributions are user-defined. Once position and length of an insertion are determined, we sample the amino acids from the equilibrium frequency of the selected substitution model Q.

3.2.3 Evolutionary Events for Constrained Segments (Domains)

Next, we describe the simulation of substitutions, insertions, and deletions for a constrained segment. The evolutionary instance is now the amino acid sequence together with its state path through the pHMM (cf. figure 3.1).

Substitutions For each site l, the emission probabilities of the associated pHMM state are taken as the stationary amino acid frequencies of the user-selected substitution model Q. Thus, each site l in the domain gets assigned its own model Q_l. The substitution rate q_{i_l} at site l is therefore $\sum_{j \neq i} \rho_{ij} e_{j_{M_x}}$ for a match state or $\sum_{j \neq i} \rho_{ij} e_{j_{I_x}}$ for an insertion state, where $e_{j_{M_x}}$ and $e_{j_{I_x}}$ are the state specific emission probabilities for amino acid j of the pHMM. The sequence site l where the substitution occurs is chosen proportional to the substitution rate q_{i_l}. The probability that amino acid i is substituted with amino acid j is proportional to q_{ij}/q_i for $i \neq j$ (Karlin and Taylor, 1975).

Insertions The probability of placing an insertion after position l in the amino acid sequence is $P(M_x, I_x)$ if l is associated with M_x, or $P(I_x, I_x)$ otherwise. The probability of placing an insertion before the first amino acid is $P(Begin, I_0)$. We apply a geometric distribution with parameter $1 - P(I_x, I_x)$ to determine the insertion length n. Simply adding the insertion to the sequence, however, poses one problem. Subsequent nested insertion events would allow insertions to grow to total lengths that

do not adhere to the model. To counterbalance this effect we have implemented the following procedure. If there are already k insertion states I_x in the state path, we only add the number of insertion states required to achieve length n, rather than adding all n insert states. Thus, at one insertion event only $n - k$ amino acids are inserted. Finally, we sample the amino acids proportional to the emission probabilities of I_x and insert them to the right of any amino acids that are already associated with state I_x.

Deletions The site l where a deletion occurs is either associated with state M_x or state I_x. In the case of M_x, we enter D_x from the respective previous state $x - 1$ to realize the deletion. Recall, that D_x can be reached either via the transition $M_{x-1} \rightarrow D_x$ or via the transition $D_{x-1} \rightarrow D_x$. The deletion probability is, therefore, either $P(M_{x-1}, D_x)$ or $P(D_{x-1}, D_x)$. We replace M_x with D_x in the state path and remove the corresponding amino acid l from the sequence. Next, we determine the length of the deletion. Note, that the pHMM does not provide an explicit deletion length distribution. Instead, it gives two choices to leave D_x: either we can move to M_{x+1} and terminate the deletion or we can move to D_{x+1} and extend the deletion. Thus, amino acids get deleted one by one, where in each step we have the choice to terminate the deletion. If D_{x+1} is already present in the state path, we move to the last deletion state in a row D_{x+z}, where z is the number of successive deletion states, and consider $P(D_{x+z}, D_{x+z+1})$ for a deletion extension. Alternatively, if the amino acid l marked for deletion is associated with I_x we proceed as follows: Transitions from I states to D states are not allowed in Plan7 pHMMs. Therefore, we first assign the same deletion probability to each I state, namely, the mean deletion probability of all match states. Then we choose the deletion length either from a geometric or a Zipfian distribution with the same parameters as for unconstrained sequence parts. Note that the deletion length is limited by the number of consecutive I states. Finally, we remove the I states from the state path and the corresponding amino acids from the sequence. This, in principle, completes the simulation schema. However, we take into account one more detail.

Resurrection of M States Deletions remove amino acids that are associated with I or M states. Insertions, on the other hand, only create I states. Consequently, on the long run the state path gets depleted of M states until only I states remain. To compensate for this erosion, we allow the insertion of amino acids that are associated with lost M states. More formally, if I_x emits an amino acid and I_x is followed by

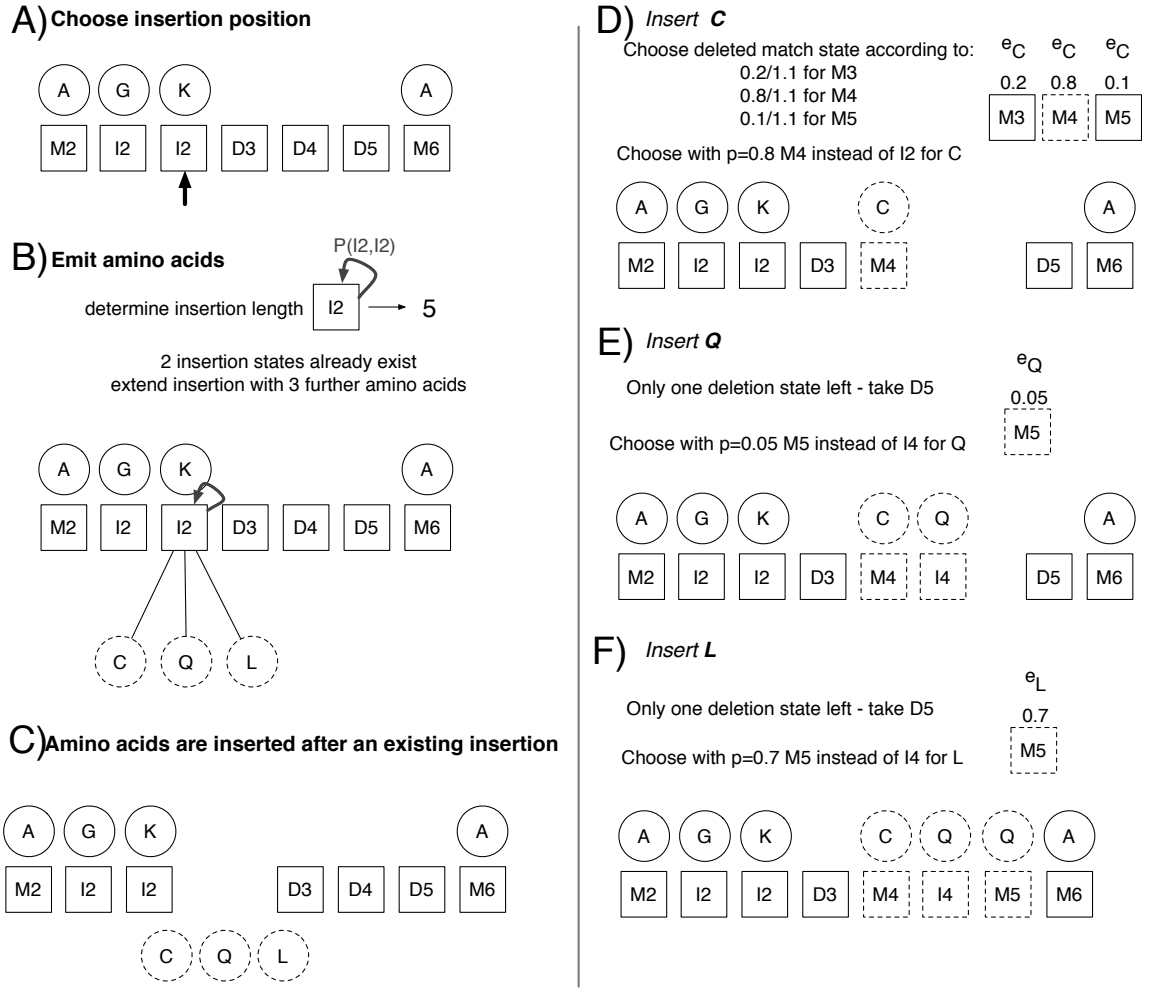


Figure 3.2: A generic insertion scenario: circles represent the amino acid sequence, the corresponding state path is shown as squares. Dashed circles and dashed squares represent newly inserted amino acids and the corresponding states, respectively. The insertion position is chosen at amino acid K with the corresponding state I_2 (A). The geometric distribution with the transition probability $1 - P(I_2, I_2)$ as parameter determines the length of the insertion (B). Amino acids are randomly emitted according to their emission probabilities E_{I_2} at state I_2. The stepwise insertion of amino acids considers the emission probabilities e_i for individual amino acids i at deleted match states (M_3, M_4, M_5) (C-F).

D_{x+1}, we facilitate the resurrection of M_{x+1}. Thus, the new amino acid emitted by the I_x state can be assigned to the M_{x+1} state.

Let us illustrate this by the example in figure 3.2: Suppose the amino acid sequence is associated with a state path as follows:

sequence :	A	G	K				A
state path :	M_2	I_2	I_2	D_3	D_4	D_5	M_6

Furthermore, suppose that an insertion length of 5 was drawn from the geometric distribution with parameter $1 - P(I_2, I_2)$ to extend I_2. Since I_2 already appears two times in the state path, we extend this insertion by additional three amino acids. We emit amino acids (CQL) proportional to the emission probabilities in vector E_{I_2}, and insert them stepwise, starting with the C, after amino acid K (cf. figure 3.2). The deletion states D_3, D_4, D_5 follow directly after I_2, and thus the C is now given the chance to resurrect one of the corresponding match states: M_3, M_4, or M_5. We first choose the candidate for resurrection with probabilities proportional to the match state emission probability for C. Assume M_4 was selected, then we decide whether or not M_4 will be resurrected. The emission probability for C at M_4 is 0.8. Consequently C will be associated with M_4 with probability 0.8 and with probability 0.2 it will stay with I_2. We then continue with the next amino acid in the insertion string, Q. Since we associated C with M_4, Q can either be associated with I_4 or M_5. In our example, we selected I_4. Finally, we insert L. With probability 0.7 (e_L at M_5), we associate L with M_5. The resulting sequence with the associated state path after the insertion is then:

sequence :	A	G	K		C	Q	L	A
state path :	M_2	I_2	I_2	D_3	M_4	I_4	M_5	M_6

and M_4 and M_5 are the newly populated match states.

3.2.4 Additional Features

Input REvolver takes a user-defined phylogenetic tree in Newick format and a root sequence as input. If the root sequence is also user-specified, a protein sequence together with its protein domain annotation via hmmscan (HMMER software package) is required. If the same amino acid in a protein is assigned to more than one domain, REvolver considers only the domain with the smallest E-value. Alternatively, the root sequence can be randomly generated. In this case, the user defines a domain architecture, i.e. a linear order of domains from the pHMM database (e.g. Pfam or SMART)

together with the lengths of any linker regions. The root protein can consist of any combination of domains and linkers. REvolver extracts the corresponding pHMMs from the database and generates a random instance for each domain. For unconstrained segments, the sequence is sampled proportional to the equilibrium frequency of the substitution model Q. Then the root sequence evolves along the tree.

When REvolver is invoked without any input, REvolver guides the user interactively through the setting of all required parameters and input files suggesting reasonable default values. Upon execution, the program generates a configuration file encoding these input parameters in xml format, which can be re-used, e.g. when integrating REvolver into an automated workflow.

Output After the simulation, REvolver outputs a multiple alignment of the simulated leaf node sequences with the options to include the root sequence or inner node sequences. Simulated sequences can be annotated with models from a pHMM database, e.g. Pfam or SMART automatically. Moreover we provide the option to present the domain architectures of the sequences visually.

Lineage specific evolution REvolver allows the specification of the substitution model and the insertion and deletion parameters individually for each branch in the tree. The model and the insertion and deletion rates will then apply to all domains and linkers.

Running time The simulation of evolution of constrained segments is obviously computationally more expensive than of unconstrained segments. Nevertheless, REvolver runs in a reasonable time. For example, the simulated evolution of a root protein of 500 amino acids with two domains along a tree with 30 leaf nodes (total tree length: 24.17 expected substitutions per site) with equal insertion and deletion rates of 0.01 took 4.9 seconds (user + sys time) on an Intel quad core i5 PC (3.30 GHz). The simulation with the same setup, but without domain constraints ran for 1.2 seconds.

3.2.5 Availability

REvolver, the manual, and example files are available for download at `www.cibiv.at/software/revolver`. The source code is available upon request. The software is

written in java, and thus runs on any platform where java 6 is installed. REvolver requires the HMMER3 software package, which is freely available at `http://hmmer.janelia.org`. pHMM databases for the REvolver simulations (e.g. Pfam, or SMART) have to be downloaded from the appropriate sources. Alternatively, custom pHMM collections may be used.

3.3 Verification of the Implementation

REvolver is the first simulator of protein sequence evolution that uses profile hidden Markov models for automatically customizing evolutionary models. In the following, we evaluated the effect of pHMM informed constraints on simulated sequence change.

3.3.1 Simulation of Substitutions

The equilibrium frequencies of REvolver's site-specific substitution models are derived from the emission probabilities of the corresponding states in the pHMM. Consequently, if related sequences evolve long enough and are then aligned, the amino acid frequencies at individual positions should again reflect the emission probabilities of the corresponding states in the original pHMM. To demonstrate this property, we used the Pfam domain A1_Propeptide whose pHMM was trained on a gap free seed alignment of 85 sequences. We evolved a single domain instance along a star tree with 85 branches to obtain a corresponding simulated seed alignment. Every sequence position on each branch was substituted on average 30 times. From the simulated sequences we then constructed a pHMM and computed a similarity score to the original A1_Propeptide pHMM with hhalign (Söding, 2005). The similarity score between the original pHMM and the pHMM inferred from the simulated data is 75.13, only slightly smaller than the score that is obtained when the original pHMM is compared to itself (80.83). In contrast, when we repeated the simulation, this time without domain constraints, the similarity score between the original pHMM and the pHMM based on the simulated sequences dropped to only 0.22. This demonstrates that REvolver's domain constraint maintains site-specific compositional properties of protein sequences.

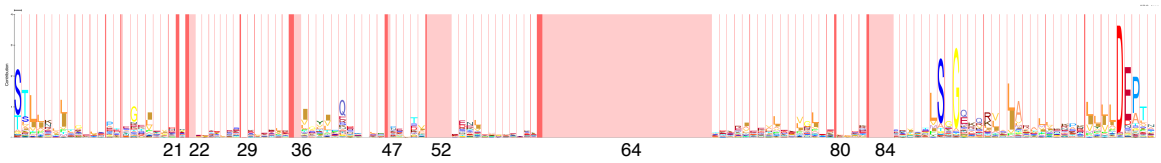
3.3.2 Simulation of Insertions

The placement of insertions within a domain, as well as their individual length distributions, are guided by the transition probabilities in the domain pHMM. Insertions are placed preferentially at such positions where the probability for reaching an insert state is high. In the same way, the transition probability to leave the insert state is used as parameter for the insertion length distribution. To verify the implementation of this procedure, we tracked insertions in the simulated evolution of the ABC transporter domain (ABC_tran; PF00005). In 10,000 simulations, we each started with an instance of the ABC_tran domain at the root that consisted of only match states ($M_1 M_2 M_3 ... M_{118}$). This sequence was then evolved under the WAG substitution model (Whelan and Goldman, 2001) up to 0.5 expected substitutions per site with $\Lambda_I = \Lambda_D = 0.1$. Then we tracked the positions of insertions as well as their respective lengths on the state path level and compared the results with the expected positions and lengths given the pHMM (figure 3.3a). We observed insertion hot-spots in our simulations at match states 21, 22, 36, 47, 64, 80, and 84 (figure 3.3b). The same match states are flagged as the most prominent insertion positions in the pHMM logo (cf. fig. 3.3a). However, note that insertions in our simulation were not restricted to these positions. They also occurred after other match states, but with considerably lower frequency. Similar to the position of the insertions, their respective length distributions also meet the expectations given the pHMM. We observed the longest insertions (mean length of 31.78 aa) at M_{64} (figure 3.3c). Similarly, insertions at M_{52} and M_{84} tend to be longer than those at other states. In summary, our results indicate that REvolver models insertions in such a way that both their distribution along the sequence and their lengths agree with what is seen in real sequences.

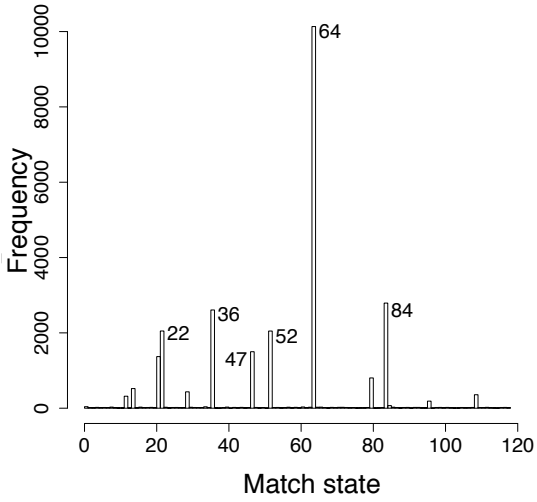
3.4 Benchmarking and Example Applications

3.4.1 Comparing REvolver to other Simulation Programs

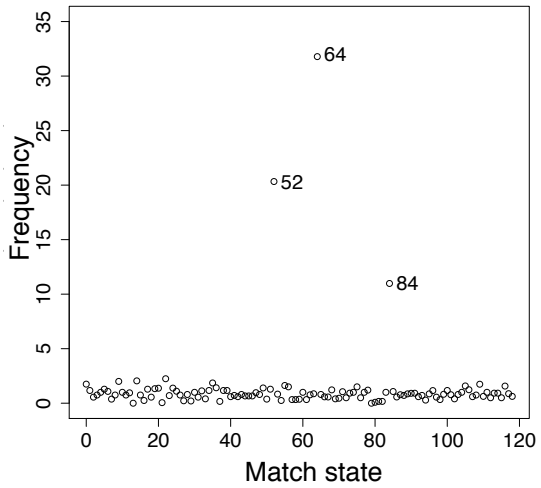
For the benchmarking of REvolver we utilized the framework introduced by Strope, Scott, and Moriyama (2007), which is based on the simulated evolution of G protein-coupled receptors (GPCR). The GPCR superfamily includes a vertebrate olfactory receptor protein family, characterized by, on average, 7 transmembrane (tm) regions and an extracellular N-terminus. Strope, Scott, and Moriyama (2007) collected 29



(a)



(b)



(c)

Figure 3.3: Positions and lengths of insertions in the ABC_tran domain. (A) The pHMM logo (Schuster-Bockler, Schultz and Rahmann, 2004) of the ABC_tran domain (`http://pfam.sanger.ac.uk/family/PF00005`) summarizes for each pHMM position information about emission probabilities, transition probability to enter an insertion state and the probability to stay in an insertion state. The relative height of an amino acid at a certain match state reflects its emission probability. The thickness of dark pink bars represent how likely an insertion occurs at a given position whereas the thickness of light pink bars represent the expected length of an insertion. (B) The histogram shows how often a pHMM position was chosen for an insertion event in 10,000 REvolver simulations starting from an ABC_tran root sequence. (C) The plot displays for each of the 118 positions of the ABC_tran pHMM the mean insertion length in the 10,000 simulations.

olfactory receptors, constructed an alignment, and inferred a maximum parsimony (MP) tree. The consensus sequence of the 29 proteins was then evolved on this MP tree. For the simulations, Strope and colleagues manually defined a variety of individual parameter settings, including the assignment of site-specific rates, invariant sites, individual rates and length distributions for insertions and deletions, and tree scaling factors for different protein segments. With these optimized settings, they compared iSG (Strope, Scott and Moriyama, 2007), ROSE (Stoye, Evers and Meyer, 1998), Seq-Gen (Rambaut and Grassly, 1997), and SIMPROT (Pang et al., 2005) with respect to the following properties of the simulated sequences: (i) the preservation of transmembrane regions, (ii) the preservation of Pfam domains, and (iii) the maintenance of a significant sequence similarity to the GPCR superfamily.

We simulated the evolution of GPCRs with REvolver adhering as closely as possible to the procedure described by Strope et al. (2007). To this end, we took the published MP tree topology and the alignment, and estimated the number of substitutions on each branch with PAUP (Wilgenbusch and Swofford, 2003). The number of substitutions per site was obtained by dividing the number of substitutions per branch inferred from the MP tree by the alignment length. We constructed a consensus sequence from the alignment of 29 olfactory receptors with iSG, annotated this sequence with Pfam, and performed 1,000 independent REvolver simulations starting from this sequence. The simulations were performed using the JTT substitution model (Jones, Taylor and Thornton, 1992). The insertion and deletion rates of 0.018 were chosen as the mean of 15 different insertion and deletion rates that were assigned to individual segments of the root protein in the analysis by Strope et al. (2007). For the benchmark test, we analyzed the simulated sequences by (i) counting the number of transmembrane regions with a transmembrane prediction program (hmmtop v2.1; Tusnády and Simon 2001), (ii) determining the presence of Pfam domains (Finn et al., 2010) with hmmscan, and (iii) assessing their similarity to GPCRs as represented in Uniprot (The UniProt Consortium, 2010) with BlastP (Altschul et al., 1997).

Table 3.2 displays the results of the benchmark test. (i) REvolver preserves on average 6.89 transmembrane regions, which is close to 7, the expected number for GPCRs. The mean observed number of transmembrane regions for sequences simulated with the other simulators are as follows: ROSE: 5.94, SIMPROT: 0.20, Seq-Gen: 6.84, and iSG: 7.03. Considering the standard deviations for the individual experiments, the differences between Seq-Gen, iSG, and REvolver are negligible. (ii) The average bit score between REvolver simulated sequences and the 7tm_1 Pfam domain (PF00001)

	REvolver	iSG	ROSE	SIMPROT	Seq-Gen
tm regions	$\mathbf{6.89 \pm 0.60}$	7.03 ± 0.30	5.94 ± 1.25	0.20 ± 0.37	6.84 ± 0.91
Pfam bit score	**102.75**	-5.09	-31.47	-	-7.18
Top n BlastP hits					
25	**152.0**	174.0	141.1	-	196.7
100	**143.6**	164.7	132.7	-	183.3
250	**135.5**	155.9	124.4	-	177.8

Table 3.2: Comparison of REvolver to other simulators. Results for the analysis of GPCR proteins. Values for iSG, ROSE, SIMPROT, and Seq-Gen were taken from Strope et al. (2007). 'tm regions' denotes the number of transmembrane regions. 'Pfam bit score' shows the mean bit scores between simulated sequences and the Pfam domain 7tm_1. No score is given for SIMPROT, because of missing 7tm_1 hits. The mean bit scores of the first 25, 100, and 250 BlastP hits are shown under 'Top n BlastP hits'. Values for SIMPROT are missing because the top scoring BlastP hits did contain non-GPCR proteins.

is 102.8 (cf. table 3.2). In contrast, sequences simulated with the other programs achieve a mean bit score of no more than -5.1 (iSG). (iii) In the third part of the analysis we show that REvolver simulated sequences have a higher sequence similarity to members of the GPCR protein family than to any other protein in the Uniprot database. For each simulated sequence, the top 250 BlastP hits were only comprised of GPCRs. The mean bit scores lie in the range between those of ROSE and iSG (table 3.2).

In summary, REvolver performs comparable or even outperforms existing protein simulators in the maintenance of functional characteristics in the chosen benchmark dataset. The major improvement however is that the parameterization to achieve this performance was done automatically and did not require any manual interaction. Thus, REvolver is able to deal with large scale data as demonstrated next.

3.4.2 Proteome wide Evaluation of Domain Content Preservation

We simulated the evolution of human proteins on a proteome-wide scale. For this purpose, we annotated 21,971 human proteins (Ensembl 51) with Pfam (Finn et al., 2010) and with SMART (Letunic, Doerks and Bork, 2009) using hmmscan with default settings. This procedure identified 45,738 Pfam and 32,289 SMART domains. Then

	insertion and deletion rates	abbreviation
unconstrained	0	I 0
	0.05	I 0.05
	0.1	I 0.1
constrained	0	F + I 0
	0.05	F + I 0.05
	0.1	F + I 0.1

Table 3.3: Parameter settings for the simulations of human protein evolution. All simulations were performed for 0.1, 0.5, 1.0, and 1.5 expected substitutions per site under the WAG substitution model (Whelan and Goldman, 2001). The geometric distribution ($p = 0.25$) was used to model indel lengths. The last column shows the abbreviations for the parameter setting used in figure 3.4, where F labels simulations under domain constraints and I denotes the parameter for the insertion and deletion rates. The analysis was performed once using the Pfam database and once using the SMART database for protein domain annotation.

we took each human protein as root sequence, simulated its evolution over different evolutionary times T (scaled in expected substitutions per site) and annotated the resulting sequences again with hmmscan. Finally, we compared the domain content for each simulated sequence with that of the root sequence. We considered a domain to be preserved if it was present both in the root sequence and in the respective simulated sequence. The fractions of preserved domains for T, ranging from 0.1 to 1.5, are shown in figures 3.4A (Pfam) and 3.4B (SMART). The parameter settings for the individual rounds of simulations are summarized in table 3.3. In the first round we set the insertion and deletion rates to 0 ($\lambda_I = \lambda_D = 0$). When simulating in the traditional way, i.e. without domain constraints, only 15% of the Pfam and 9% of the SMART domains were preserved at $T = 1.5$. This figure changes substantially, when we impose domain constraints. In this case, more than 90% of the Pfam and SMART domains were detected in the simulated sequences at $T = 1.5$. Subsequently, we assessed the effect of insertions and deletions. For the evolution without domain constraints, the percentages of retained domains decreased rapidly with increasing evolutionary time. At $T = 1.5$ only 1%/2% (Pfam/SMART) of the original domains were maintained with insertion and deletion rates of 0.05, and only 0.5%/1% with insertion and deletion rates of 0.1.

Here, the effect of domain constraints on the preservation of domains over time was even more pronounced. At $T = 1.5$ still 79%/74% (insertion and deletion rates

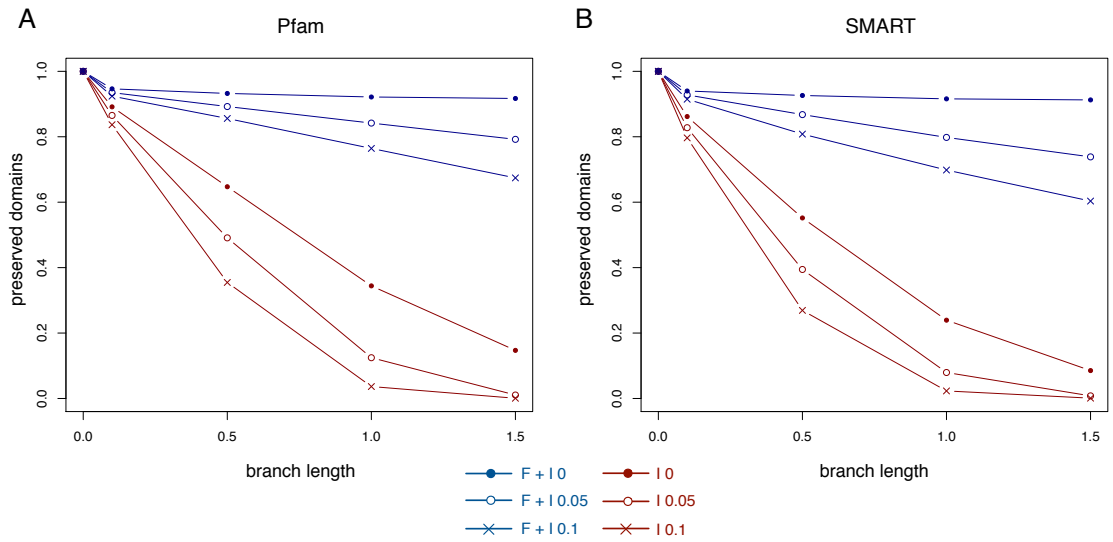


Figure 3.4: Fraction of preserved Pfam (A) and SMART (B) domains. All human proteins were taken as root sequences and evolved with 0.1, 0.5, 1.0 and 1.5 expected substitutions per site. F denotes simulations with domain constraints (blue lines). Simulations without domain constraints are colored in red. I 0 stands for simulations without indels, I 0.05 for insertion and deletion rates of 0.05, and I 0.1 for insertion and deletion rates of 0.1 (cf. table 3.3).

of 0.05) and 67%/60% (insertion and deletion rates of 0.1) of the domains were preserved. Simulations under a rate across sites (RAS) model are often used to account for sites under different evolutionary constraints in a protein. We therefore repeated our simulation procedure for the unconstrained case using two different values for the shape parameter of the gamma distribution ($\alpha = 1$ and $\alpha = 0.5$). Despite the case of the RAS model, the increase in the number of retained domains was only marginal, when insertions and deletions were included in the model (Appendix tables B.1 and B.2). Without indels, simulations under domain constraints still outperformed the RAS model by a factor of 2-3.

3.4.3 Preservation of Structure

So far we have shown that REvolver substantially increases the evolutionary stability of protein domains in the course of simulated sequence change. Although structural constraints are not explicitly captured in pHMMs (but see Eddy 1998), we next assessed whether sequences simulated under domain constraints are also structure-wise more similar to the native protein than sequences simulated without constraints. For our analysis we used the human SAP SH2 protein (Poy et al., 1999) and evolved it with and without domain constraints ($\Lambda_D = \Lambda_I = 0$). Then we assessed the rooted mean square distance (RMSD) between the structure of the native protein (1d4tA; Velankar et al. 2011) and the inferred structure of the simulated sequences. SARA (Grahnen, Kubelka and Liberles, 2011) was used for analyzing the RMSD between corresponding side chains in the two structures. Next, we used MODELLER (Eswar et al., 2006) to analyze the RMSD between the peptide backbones of two structures. This analysis was performed with 3 different insertion and deletion rates ($\Lambda_D = \Lambda_I = 0/0.05/0.1$). In all comparisons, the RMSD between the native structure and the inferred structure of the simulated sequence was smaller for the constrained simulation than for the unconstrained simulation (Appendix figure. S1 and S2). A one-sided t-test ($\alpha = 0.05$) revealed that, except for a single case, the differences are significant.

3.4.4 Simulation of Proteins with user-defined Domain Architectures

REvolver is the first program that offers the possibility to simulate protein evolution with user-defined domain architectures. To exemplify this feature, we used REvolver

to generate a random root sequence consisting of instances of an RLI domain (Possible metal-binding domain in RNase L inhibitor; PF04068), a Fer4 domain (4Fe-4S binding domain; PF00037), and two ABC_tran domains (ABC transporter; PF00005). The domains are separated by unconstrained segments of different lengths. The root sequence was then evolved under the WAG model along an arbitrary phylogeny displayed in figure 3.5, using insertion and deletion rates of 0.1. Figure 3.5 shows the input tree together with the resulting domain architectures of the simulated sequences at the leaves. Not all domains are preserved in all of the simulated sequences. For example, at leaves G, F, and E the Fer4 domain was lost. Domains also diverged in length due to insertions and deletions. Hence, REvolver produces sequences of similar, but not identical, domain architectures. The resulting pattern of presence and absence of protein domains resembles what can be observed in real protein families.

3.5 Discussion

In recent years, a number of approaches were developed to simulate evolutionary protein sequence change (e.g. Rambaut and Grassly 1997; Stoye, Evers and Meyer 1998; Fletcher and Yang 2009; Pang et al. 2005; Strope et al. 2009; Rastogi, Reuter and Liberles 2006; Lakner et al. 2011; Grahnen et al. 2011). With REvolver we present a new, versatile simulator that stands out from existing programs in two relevant aspects: The maintenance of protein domains in the course of evolution, and the large-scale applicability due to the automatic inference of sequence specific evolutionary constraints. We have shown that the pattern of sequence differences between homologous sequences, as captured in pHMMs, can be used to describe adequately the constrained evolutionary process to which a protein domain is subjected. REvolver is the first tool that integrates this information about protein sequence evolution in an automated fashion. To facilitate the use of pHMMs in sequence evolution simulations we implemented several essential features. The first aspect is concerned with the modeling of insertions. We have derived the parameter for the geometric distribution used to model insertion lengths from the transition probability $P(I_x, I_x)$ of an insertion state. This transition probability was trained on an alignment of contemporary sequences. Consequently, sampling from the resulting geometric distribution results in insertions lengths that are observed in extant sequences. However, they do not necessarily represent the lengths of individual insertion events. Multiple nested insertions in the simulation would therefore result in much longer insertions than they

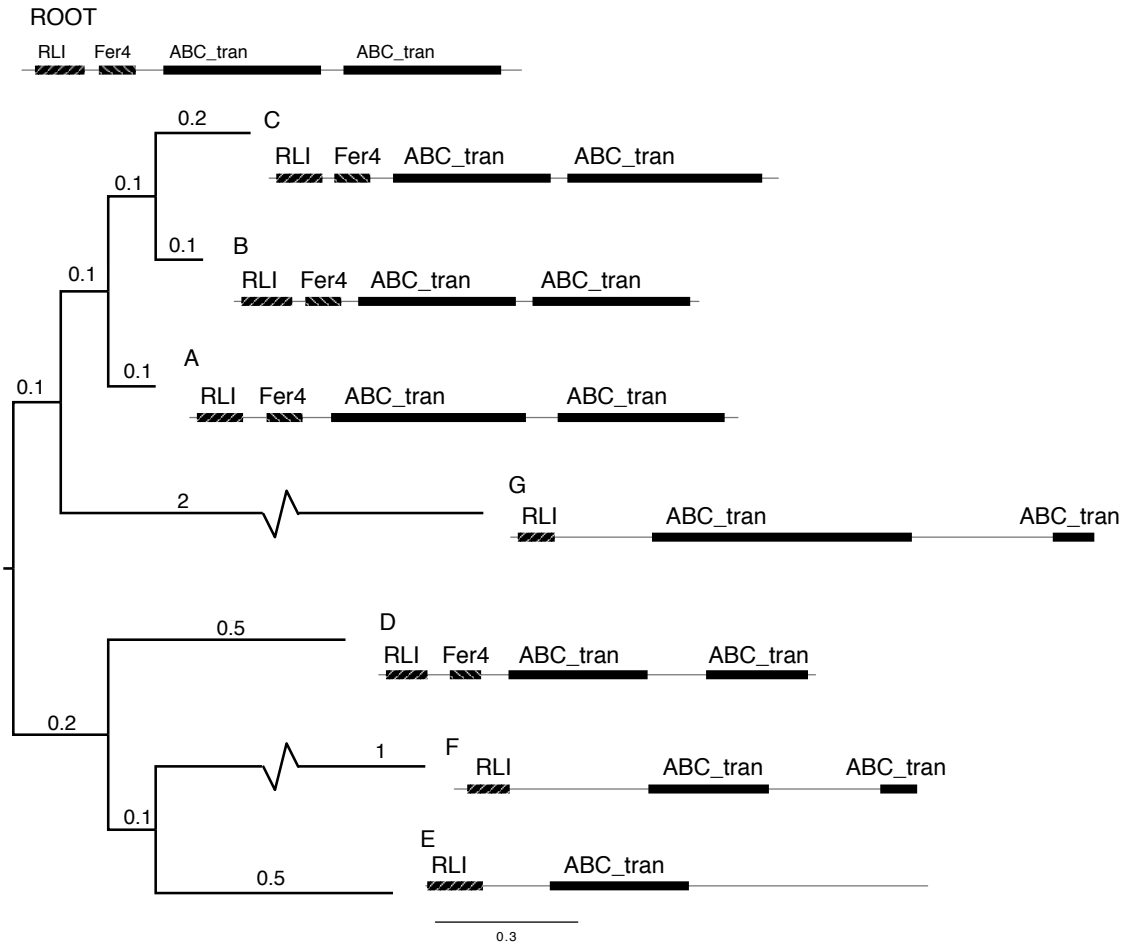


Figure 3.5: Domain architectures of sequences evolved with REvolver. A root sequence with the specified domain architecture was evolved on the shown tree. The root sequence consists of one RLI (PF04068), one Fer4 (PF00037), and two ABC_tran (PF00005) domains (Finn et al., 2010) separated by linker regions. Domains were evolved under domain constraints, linker regions were evolved without domain constraints. Branch lengths are given in expected substitutions per site, but are not drawn to scale for lengths ≥ 1.

were observed in the sequences used to train the model. To prevent the formation of such unrealistically long insertions, REvolver only extends insertions to the actually drawn random variable from the geometric distribution. Thus, the total length of an insertion in the sequence is always a value from the geometric distribution. The second aspect is concerned with the gradual erosion of M states due to the deletion process. We counterbalance this effect by facilitating the resurrection of M states via the insertion process. This is important to maintain the identity of the domains; otherwise, it would just be a matter of time until all match states have been lost and amino acids are all associated with insertion states. From the biological point of view, our procedure is also reasonable: Suppose, for example, that at one point during evolution, a functional site is deleted. This deletion may not abolish the functionality of the protein or domain but modify it. If at some point later in time, an amino acid is inserted at the previously deleted position that, by chance, has similar or the same properties as the original amino acid, the protein's function would be fully restored. In the current version of REvolver we assess the probability that an inserted amino acid revives a previously lost M state using the probability that this M state emits exactly this amino acid. We can think of alternative ways of realizing the resurrection. One possibility would be to consider the inserted segment as a single entity rather than individual amino acids. The goal would then be to find the state path that most likely emitted that amino acid segment (Viterbi 1967). Insertion states and deleted match states would be valid states for the path, deletion states would be forbidden. However, for now we decided to implement the step-wise insertion procedure, since it is simpler and computationally less expensive.

Our comparison of REvolver to other simulators of protein sequence evolution has shown that REvolver solves two tasks in the benchmarking optimally, i.e. the maintenance of 7 tm domains, and maintaining a significant similarity of the simulated sequences to the GPCR protein family. However, in contrast to the other programs, for which 7 transmembrane regions were explicitly defined and parameters had to be tweaked manually to obtain optimal performance, REvolver performed the parameterization automatically. The difference between the compared simulators becomes even more obvious in the third task, namely the maintenance of the similarity of the simulated sequences to the 7tm_1 pHMM (PF00001). This pHMM models a 7 transmembrane receptor domain, which is characteristic for the GPCR protein family (Palczewski et al., 2000). While the similarity between the sequences generated with the existing simulators and the 7tm_1 pHMM is poor, sequences simulated with our

program achieve average bit scores (102.8) that are only slightly lower than what is achieved on average when comparing real GPCRs to the pHMM (124.4). Thus, REvolver not only preserves the correct number of tm domains, but also the intervening regions required for placing them in a functional context of a 7 transmembrane receptor. This result suggests that REvolver may also conserve structural properties of protein domains, although they are not embedded in the pHMMs (but see Eddy 1998). To follow this issue up further, we simulated the evolution of the human SAP SH2 protein both with and without domain constraints and determined the RMSD between the structure of the native protein and the inferred structure of the simulated sequences. The results confirmed that, indeed, the simulation of sequence evolution under domain constraints not only maintains domain sequences, but also has a positive influence on the preservation of their structure.

So far we have demonstrated the use of REvolver only in the combination with pHMMs derived from public databases. However, REvolver simulations under domain constraints are applicable to all proteins even if they show no significant sequence similarity to any of the domains for which public pHMMs are available. Alternatively, it may be desired to use pHMMs more specific than those available in the public databases, e.g. when a particular protein sub-family is analyzed. In such instances, the protocol is straightforward: For any given root sequence, homologous sequences can first be identified, e.g. via a Blast search. The root together with a set of homologous sequences can then be aligned and used to construct and train a pHMM. REvolver then uses this custom pHMM to infer the evolutionary constraints for the root sequence. We have exemplified this procedure with the GPCR dataset. To this end, we constructed a pHMM from the alignment of the 29 GPCRs. Next, we simulated the evolution of the GPCR protein family using this custom pHMM. The simulated sequences still retain most of the transmembrane regions, show a significant sequence similarity to the 7tm_1 domains, and find only other GPCRs among the top BlastP hits (Appendix table B.3). This shows that even in the case of missing explicit information about protein specific features, REvolver still preserves most of them.

In summary, REvolver is a versatile tool for simulating evolutionary sequence change and improves in many aspects over existing simulators. Although not limited to it, one obvious application of REvolver is the generation of benchmark datasets for programs designed to trace and interpret the evolutionary signal in molecular sequences, e.g. programs for sequence alignment, orthology prediction, or tree reconstruction (e.g. Felsenstein 2004; Notredame 2007; Remm, Storm and Sonnhammer 2001). Testing

the accuracy of these tools with real data is obviously problematic, since the evolutionary history is frequently not known (cf. Chen et al. 2007). Benchmarking on sequences that have been evolved *in silico*, in principle, overcomes this problem. Still, the results are of little relevance if the scheme used for simulating sequence evolution is unrealistic (Kim and Sinha, 2010). From this perspective, we expect that REvolver is a significant contribution to this field. We envision an even stronger impact when it comes to the benchmarking of programs that search for proteins with similar feature architecture (Koestler, von Haeseler and Ebersberger, 2010) or that infer the function of a protein based on its domain content (Forslund and Sonnhammer, 2008). The simulated evolution of a domain architecture along a tree is still in its infancy, as REvolver does not consider evolutionary events like domain shuffling and domain stealing. However, an integration of such mutation events will be a logical extension to REvolver's simulation scheme.

Chapter 4

Evolutionary Stability of Protein Domains

In chapter 2, we have shown that domains are important for the functional annotation of proteins. Profile Hidden Markov Models (pHMMs) trained on a set of known domain instances summarize the characteristics of a domain and are commonly used to search for new domain instances in uncharacterized proteins. The corresponding program in the software package HMMER (`http://hmmer.janelia.org`), for example, measures the similarity between a sequence and a pHMM. If the similarity is considered significant - by default an E-value < 0.01 is used - the program reports the domain as present. Otherwise it is typically concluded that the domain is absent. As a matter of fact, the result from HMMER does not tell us whether a domain is truly absent or whether it just escaped detection. The differentiation between the two possibilities is, however, sometimes important. For instance, orthologous proteins that have preserved the ancestral function are also expected to have the same homologous domains. However, the domain content of orthologs is in many cases not identical. A decision whether or not such orthologs can still be functionally equivalent depends on an assessment of the sensitivity of the domain search, i.e. how sure we are that we have not missed a domain in our search. In this study, we simulate the evolution of domains with REvolver (chapter 3) to assess the extent of evolutionary change a domain can bear before it is no longer detected by HMMER. We characterized 11,912 Pfam domains by computing their half-lifes i.e. the number of mutations per site it takes until 50% of instances are no longer recognized as domain. First, domains with long half-lifes can be detected even after many mutations, thus, over large evolutionary distances (measured in substitutions per site). Second, domains with short half-lifes

become undetectable already after few sequence changes. We consider the first domains as evolutionarily stable and the second as unstable. An interesting third type are *zombie* domains that repeatedly disappear and reappear in the course of simulated evolution. This temporary disappearance can again be interpreted as a lack of sensitivity in the domain search. We use this information to parameterize an evolutionary model of domain loss and gain. This serves as a null model to distinguish between cases where the domain is truly absent (true negatives) in a phylogenetic tree and cases where it is overlooked in the domain search (false negatives). In summary, the knowledge about evolutionary domain stability enables a more meaningful interpretation of the presence-absence patterns of a domain in evolutionarily related proteins.

4.1 Introduction

Profile hidden Markov models (pHMMs) are probabilistic models describing a multiple sequence alignment and are commonly used as position-specific scoring systems (Eddy, 1998). Pfam (Finn et al., 2010), SMART (Finn et al., 2010), Superfamily (Gough et al., 2001), and prosite (Hulo, 2006) are examples of databases dedicated to collect domain sequences, align them, and finally construct and train pHMMs. These pHMMs have become a standard bioinformatics tool for analyzing sequences. For example, they are used for remote homolog detection or to annotate a protein with domains (see chapter 2). Protein domain annotations are important for a variety of biological and evolutionary questions. For instance, the function of a protein can be inferred based on its domain content (Forslund and Sonnhammer, 2008). The similarity between domain architectures (i.e. the linear order of domains) helps to identify functionally equivalent proteins even if they miss a significant sequence similarity (chapter 2; Koestler, von Haeseler and Ebersberger 2010; Song, Sedgewick and Durand 2007; Lee and Lee 2009). Apart from inferring a protein's function, the annotation of whole proteomes with pHMMs facilitates studies on the evolution of domain architectures (Forslund et al., 2008; Buljan and Bateman, 2009; Gough, 2005). It has been shown that domain losses and gains are enriched at protein termini compared to losses and gains at more central regions of the protein (Weiner, Beaussart and Bornberg-Bauer, 2006) and that between 0.4% and 12.4% of the domain architectures evolved at least two times independently during evolution (Gough, 2005; Forslund et al., 2008). Furthermore, some domains occur always in the same domain context whereas others contribute to many different domain architectures (cf. promiscuous domains; Marcotte et al. 1999).

Inferences of domain loss and gain rates, domain architecture evolution, or assessment of domain promiscuity heavily depend on the accurate recognition of domains. However, examples were reported where a domain is physically present in a protein, but does not show a significant similarity to the corresponding pHMM (Weiner, Beaussart and Bornberg-Bauer, 2006). Especially, since pHMMs are models and, thus, simplifications of the reality, they do not necessarily reflect the full sequence space of a domain. For example, the underlying multiple sequence alignment used to train the pHMM may miss instances of the domain that deviate in sequence from the known instances. The resulting biased training of the model parameter can be one source of erroneous domain annotation. Moreover, it is unclear how many substitutions can happen until a sequence is no longer detected as domain instance. Some domains are characterized by a strict motif. Hence, already a single substitution might destroy the domain. Other domains show a much higher sequence variability. Here, the domain instance matches the domain specific characteristics even after several substitutions. Apparently, also insertions and deletions, that are modeled by a pHMM, can destroy the domain characteristics. Some domains require a fixed sequence lengths, whereas others tolerate length deviations and are, thus, detected as domain instance even after several insertions and deletions. Thus, the sensitivity in the domain annotation is domain specific.

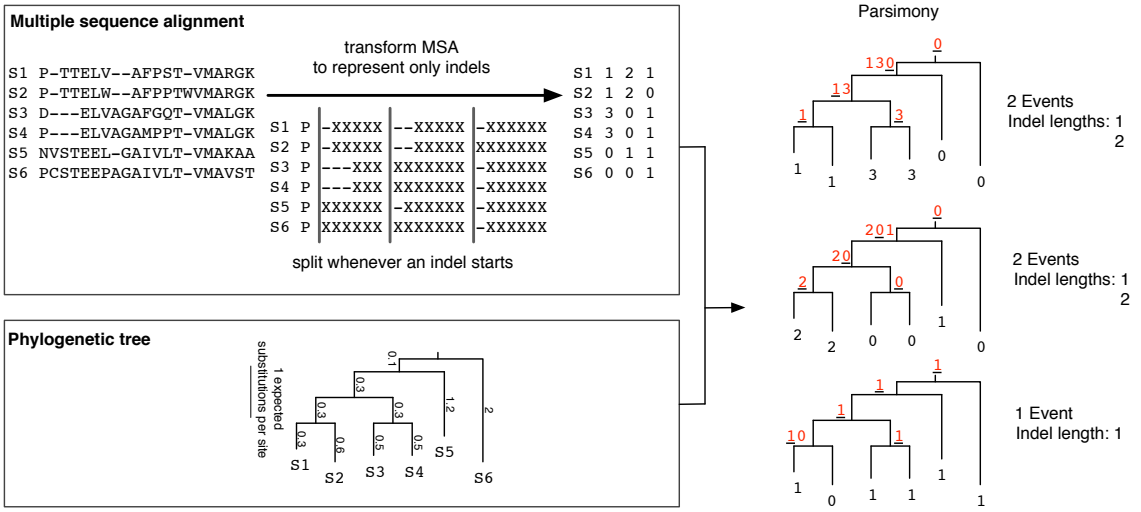
Here, we present a study on the sensitivity of domain annotations. For this purpose, we asses over what evolutionary distances a sequence can be detected as domain instance that we then call the evolutionary stability of the domain. In particular we are focusing on Pfam (Finn et al., 2010) one of the most widely used protein family databases.

4.2 Methods

Insertion and deletion rate estimation To overcome ad-hoc decisions in the insertion and deletion rate parameterization during simulation of domain evolution, we estimated the insertion and deletion rates for each domain using the following parsimony approach (summarized in figure 4.1): We downloaded the pHMMs and the corresponding multiple sequence alignments (seed alignments) for all domains from the Pfam database. For each seed alignment, we calculated a bifurcating tree with FastTree (Price, Dehal and Arkin, 2010). Subsequently, we split the seed alignment at

every column where a gap started. We then transformed the alignment by replacing each block of consecutive gaps in a sequence by the number of consecutive gaps. The transformed alignment and the phylogeny served as input to calculate the minimum number of events (insertions and deletions) to explain the observed pattern (Fitch, 1971). Next, we calculated the event rate by dividing the parsimony score of the transformed alignment by the length of the seed alignment and the tree length. Thus, the unit of the event rate is per substitution per site. The insertion rate and the deletion rate were then taken as half of the calculated event rate. Additionally, we chose one most parsimonious solution of the insertion and deletion history for which we stored the individual lengths for each node in the tree and calculated the mean indel length. We estimated the parameter p for a geometric indel length distribution to be 1 divided by the mean indel length. This distribution is taken by REvolver to simulate the evolution of sections that are not covered by a pHMM (linker regions; see chapter 3).

Evolutionary stability estimation We simulated the evolution of all 11,912 Pfam domains (Finn et al. 2010; Version 24) where time ranges from 0.5 to 11 expected substitutions per site using REvolver (figure 4.2). The root domain instance was generated with REvolver by passing through all match states in the pHMM of a domain and emitting an amino acid according to the state specific emission probabilities. We estimated the insertion and deletion rates from the Pfam seed alignment as described in the previous paragraph. In cases where the seed alignment comprised only 1 or 2 sequences, or if all sequences in the seed alignment were identical, we used the default insertion and deletion rates of 0.01 insertions and deletions per substitution, respectively (subsection 4.3.1). Finally, we started REvolver for each Pfam domain. Every 0.5 substitutions per site a copy of the simulated sequence was made and stored for subsequent domain analyses. All simulated sequences were analyzed with HMMER. If a sequence or parts of a sequence showed a significant similarity (E-value ≤ 0.01) to the domain pHMM, we considered a domain to be recognized. We repeated the procedure 100 times for each domain. Eventually, we calculated the half-life of a domain as the number of substitutions per site it takes until 50% of the simulated sequences were no longer recognized as domain.



Insertion rate = 1/2 Events/(Alignment length x tree length) = 1/2 5/(21x6.1) = 0.0195

Deletion rate = 1/2 Events/(Alignment length x tree length) = 1/2 5/(21x6.1) = 0.0195

p = 1/mean indel length = 1/ (7/5) = 0.7143

Figure 4.1: Parsimony approach to estimate indel rates and length distribution parameters. We split a multiple sequence alignment (MSA) whenever an indel starts. We construct a transformed alignment by writing down for each alignment part the number of consecutive gaps for each sequence. Next, we calculate the maximum parsimony score for each column of the transformed alignment given the tree that was inferred form the original alignment. The maximum parsimony score is the minimum number of insertions and deletions required to generate the transformed alignment. The parsimony score is equally distributed between insertions and deletions. The insertion and deletion lengths of one most parsimonious solution are used for calculating p, the parameter for the geometric indel length distribution.

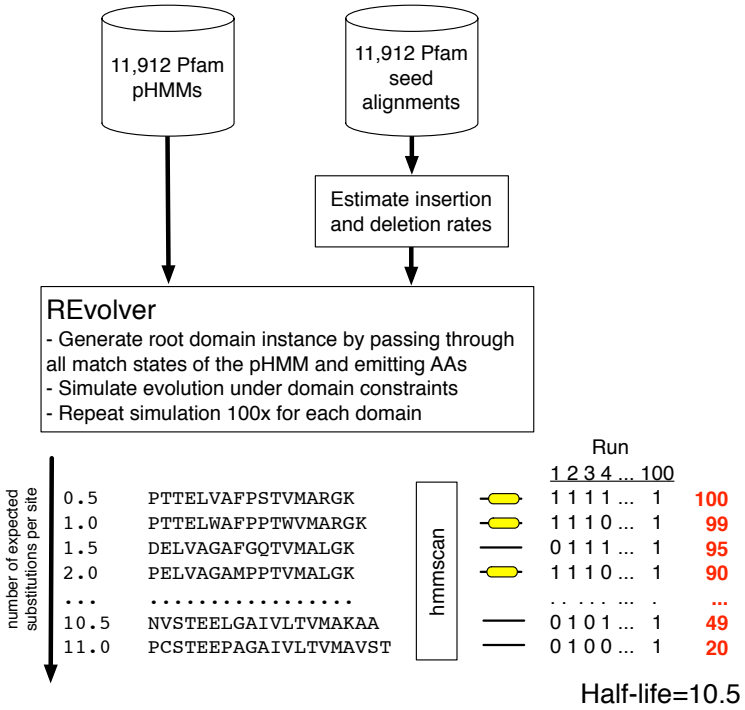


Figure 4.2: Workflow to estimate the evolutionary stability of a Pfam domain. The first step is the estimation of domain specific insertion and deletion rates based on its seed alignment. A domain instance generated by REvolver serves as the root sequence and its evolution is simulated up to 11 substitutions per site considering insertions and deletions. The simulation is performed in steps of 0.5 substitutions per site where a copy of the simulated sequence is made before the simulation continues. All simulated sequences are analyzed with the corresponding program in the HMMER software package (hmmscan). The table next to the hmmscan box summarizes the detection (1) or non-detection (0) of the domain. The procedure is repeated 100 times for each domain. The row sums, given in red, represent in percent how often a sequence was recognized as a domain instance. In this example, after 10.5 substitutions per site only 49% of the sequences show a significant similarity to the corresponding pHMM. The half-life for this domain is therefore 10.5.

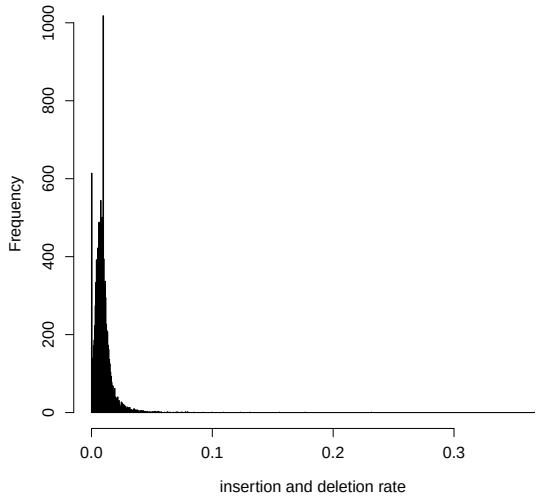


Figure 4.3: Histogram of estimated insertion and deletion rates for all Pfam domains. Most rates are close to the average of 0.01 insertions and deletions per substitution. Insertion and deletion rates of 0.05 or higher are very rare.

4.3 Results

4.3.1 Insertion and Deletion Rates in Domains

Figure 4.3 shows the histogram of estimated insertion and deletion rates. Most rates are close to the average of 0.01 insertions and deletions, respectively, per substitution (standard deviation: 0.009). A substantial fraction (5%) of the Pfam domains had no insertion or deletion in the corresponding seed alignment. Consequently, the estimated indel rate was zero. On the other hand, insertion and deletion rates above 0.05 occurred very rarely (0.5%). Obviously, a maximum parsimony approach is a very conservative way to estimate insertion and deletion rates and likely underestimates the true number of events. To assess the impact of higher insertion and deletion rates, we performed the simulated evolution first with the estimated indel rate, second with twice the estimated indel rate, and third with a fixed indel rate of 0.1.

4.3.2 Half-lifes of Domains

Figure 4.4a shows the histogram of half-lifes with a mean of 11.2±1.23 substitutions per site using the indel rates estimated with the parsimony approach. For 10,986 domains

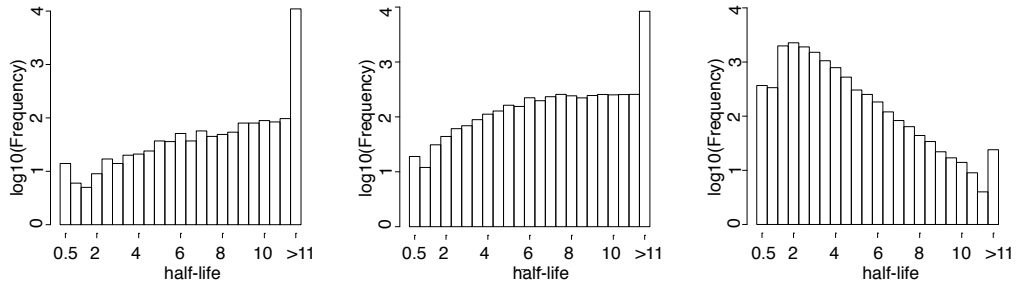


Figure 4.4: Log scaled histograms of half-lifes (in substitutions per site) of all Pfam domains resulting from simulations with (a) estimate indel rates, (b) doubled estimated indel rates, and (c) fixed indel rates of 0.1.

(92%) the half-life exceeds 11 substitutions per site. However, some domains are no longer detected already after a few substitutions. A doubling of the estimated indel rates slightly shortens the half-lifes of domains to a mean of 10.3 ± 2.3 substitutions per site (cf. figure 4.4a and b). However, the overall shape of the histogram does not change. This shows that the half-life estimations are fairly robust against small variations of the insertion and deletion rates. Only when we increase the insertion and deletion rates to an extreme value of 0.1, the shape of the half-life histogram changes substantially (mean of 2.8 ± 1.54; figure 4.4c). Now, only 24 domains have an half-life ≥ 11 substitutions per site. Since the half-lifes between simulations with twice the estimated indel rates are similar and since we consider an indel rate of 0.1 as unrealistically high, we focus on the parsimony indel rates.

In summary, the vast majority of domains display a very long half-life and are, therefore, considered evolutionarily stable. We, therefore, expect that such domains can be detected even in highly diverged sequences. Oppositely, some domains are no longer recognized already after very few mutations. These domains have short half-lifes and are, thus, hard to detect among distantly related sequences.

Why are some domains more stable than others?

The question remains why some domains are not recognized as domain already after a few mutations, while others are recognized even after 11 substitutions per site? The histograms in figure 4.4 show that the insertion and deletion rates might influence the domains' half-life. To follow this issue up further, we contrast the indel rates for

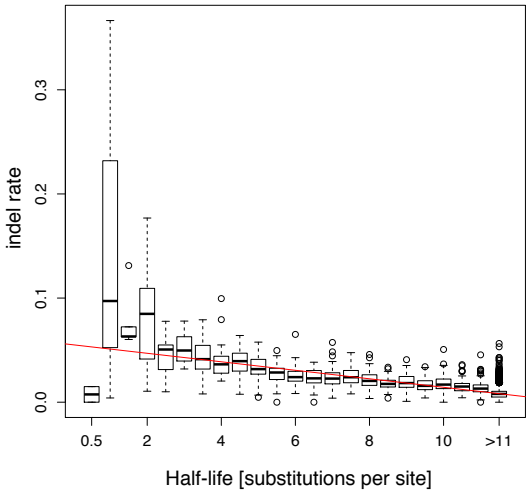


Figure 4.5: Boxplots of the parsimony insertion and deletion rates for domains of the same half-life. The red line shows the linear correlation between half-life and insertion and deletion rate.

the domains with their half-lifes (figure 4.5). We observe a negative linear correlation (Pearson correlation coefficient = -0.6) between the half-lifes and the indel rates. In contrast to the general trend, domains with a half life of 0.5 substitutions show very small indel rates. The coefficient of determination between the rates and the half-lifes is 0.35. In other words, the insertion and deletion rate explains only a part (35%) of the differences in half-lifes. Moreover, we also observe differences in half-lifes in the simulation where the insertion and deletion rate was the same for all domains (see figure 4.4c). We, therefore, analyzed the following additional factors for their influence on the half-lifes:

sequence diversity We calculated the mean pairwise sequence identity (%) of the seed alignments using `alistat` and compute the sequence diversity = (100 - percent identity).

nseq[1] Number of sequences in the seed alignment.

eff_nseq[1] Effective number of sequences in the seed alignment.

M[1] Number of match states in the pHMM.

compKL[1] The Kullback-Leiber distance between the average amino acid composition of the pHMM and the background frequency distribution.

[1]Factor is extracted from the pHMMs with hmmstat (HMMER software package).

	r	R^2
indel rate	-0.59	0.35
compKL	-0.19	0.04
relent	-0.10	0.01
info	-0.09	0.01
p_relE	-0.08	0.01
nseq	-0.07	0.01
eff_nseq	0.04	0
M	0.12	0.01
sequence div.	0.15	0.02

Table 4.1: Pearson correlations coefficient r and the coefficient of determination R^2 between individual factors (characteristics of a pHMM and the underlying trainings data) and the half-lifes of domains. r is commonly interpreted as follows: 0 to 0.09 = none; 0.1 to 0.3 = small; 0.3 to 0.5 = medium; 0.5 to 1.0 = strong; the same applies to negative correlations by adding a minus sign.

relent[1] Mean relative entropy per match state.

info[1] Mean information content per match state.

p_relE[1] Mean positive relative entropy.

Table 4.1 summarizes the Pearson correlation coefficient and the coefficient of determination between the individual factors and the half-life. The indel rate influences the half-lifes to the greatest extent. Intuitively, one would expect that pHMMs trained on many sequences from a diverse set of species capture the sequence space more comprehensively than those trained on only a few sequences and thus should tend to have longer half-lifes. However, this trend is not prominent in the data. Both, the absolute and the effective number of sequences do not correlate with the domain half-life. The correlation coefficients are close to zero. The sequence diversity shows a small positive correlation with the half-life. Hence, pHMMs trained on seed alignments with diverse sequences indeed tend to be recognized longer than models trained on seed alignments with very similar sequences. This is however independent of the number of sequences. The factors describing the distinctiveness of a pHMM (compKL, p_relE, info, and relent) have none or a very small negative correlation with the half-life. In contrast, the number of match states and the half-lifes are positively correlated. In summary, long domains trained on a diverse set of sequences with a small indel rate are in general evolutionarily stable. Notice, that only linear correlations between the half-life and individual factors were analyzed.

4.3.3 Zombie Domains

In the previous sub-section we evaluated the half-lifes of domains. However, the half-life is a summary statistic of the simulated evolution of 100 instances per domain. In the following, we took a closer look at the individual domain instances. Lets assume that after some substitutions, insertions, and deletions a protein sequence is no longer recognized as domain. It may now happen that after an additional substitution the sequence is again recognized as domain. Two simulation runs in figure 4.2 (run 1 and 4) show such a situation. In simulation run 4, for example, the sequence was recognized as domain after 0.5 substitutions per site, not recognized after 1 substitution per site, and then again recognized after 1.5 substitutions per site. Apparently, the domain was overlooked in the domain search due to a lack of sensitivity (false negatives). In the following, we investigate this behavior in our simulations.

Of all domains for which at least one instance was not recognized at one point in the simulated evolution (10,797) 99% recurred later in the simulation. Thus, an apparent loss was followed by an apparent gain. We measured two values for each domain. First, the percent of recurrences: For each domain we generated 100 instances and simulated their evolution independently. 100 percent recurrence means that all 100 instances were apparently lost and subsequently re-gained. Second, the mean number of recurrences: We counted how many times a single domain instance was apparently lost followed by an apparent gain. From the 100 simulations we determined the mean number of recurrences for the domain. Figure 4.6 shows the percent of recurrences versus the mean number of recurrences for all Pfam domains. Interestingly, the phenomenon of apparent losses with subsequent gains is widely spread in Pfam domains. Extreme cases are domains where almost all instances underwent several rounds of apparent loss and gain (top right dots in figure 4.6). For example, the bacterial transferase hexapeptide (Hexapep; PF00132), is such a *zombie* domain where each simulated instance disappeared and recurred. Other extreme examples with a 99% recurrence are the following seven domains: Involucrin repeat (Involucrin; PF00904), KID repeat (KID; PF02524), Insect kinin peptide (Kinin; PF08260), Seven Residue repeat (SRR, PF07709), Coagulation Factor V LSPD repeat (LSPR; PF06049), Copper binding octapeptide repeat (Prion_octapep; PF03991), and RII binding domain (RII_binding_1; PF10522). Five of the seven domains are repeats and all pHMMs consist of very few match states. The RII binding domain is with 18 match states the longest of these domains. In general, we observe a negative correlation between

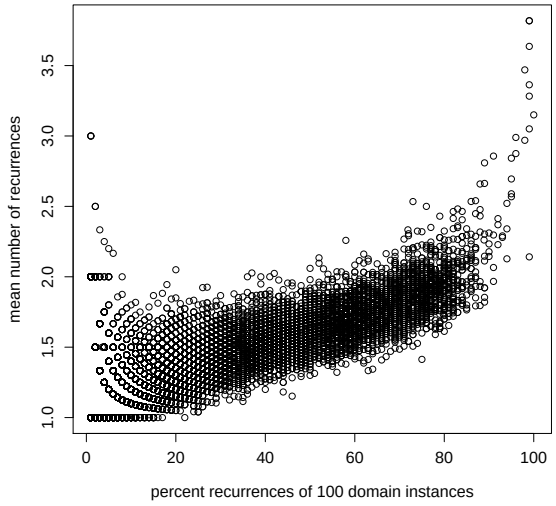


Figure 4.6: Comparison of the percentage of instances per domain that recurred after they disappeared (x-axis) and the mean number of recurrences per domain (y-axis).

the number of match states in the pHMM and its zombie behavior. The Pearson correlation between the number of match states and the percent of recurrences is -0.48. Similarly, the correlation between the number of match states and the mean number of recurrences per domain is -0.55. Thus, the less match states a pHMM contains the more it tends to behave like a zombie.

The temporary disappearance of zombie domains can be viewed as a lack of sensitivity of the domain detection method. This raises the question, how to distinguish between erroneous non-detection of a domain (false negatives) and a real loss of the domain (true negatives) in a phylogenetic tree? We, therefore, suggest a general time reversible model Q for domain losses and gains with one free parameter and normalize it to one event per time unit.

$$Q = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} -\frac{1}{2\pi_0} & \frac{1}{2\pi_0} \\ \frac{1}{2(1-\pi_0)} & -\frac{1}{2(1-\pi_0)} \end{pmatrix} \end{array} \tag{4.1}$$

where, π_0 is the equilibrium frequency of 'domain not detected' (0). The equilibrium frequency for 'domain detected' $\pi_1 = 1 - \pi_0$ (1). Based on the simulations, we parameterize the Q matrix for a specific domain. This matrix then serves as a null model to

evaluate the possibility of a real domain loss in a phylogenetic tree. In the following, we exemplify this approach on the RII binding domain.

To estimate π_0 and π_1, we determine the number of substitution where the proportion of detected domains per simulation stays approximately constant. This is the case for the RII_binding_1 domain, where from 1 substitution per site onwards $58 \pm 6\%$ of the instances are recognized as domain. We count how often a domain (i) remains recognized, (ii) remains not recognized, (iii) flips from recognized to not recognized (disappears), and (iv) flips from not recognized to recognized (reappears) after the period of 1 substitution per site. The following matrix $P(t)$ is the resulting probability matrix for the RII_binding_1 (RII) domain and time $t = 1$:

$$P_{RII}(1) = \begin{pmatrix} 0.60 & 0.40 \\ 0.70 & 0.30 \end{pmatrix} \tag{4.2}$$

From $P = e^{Qt}$, we can approximate Q_{RII} for the RII_binding_1 using a least square approach:

$$Q_{RII} = \begin{pmatrix} -0.73 & 0.73 \\ 1.61 & -1.61 \end{pmatrix} \tag{4.3}$$

According to matrix 4.1, the approximate $\pi_0 = 0.69$ and $\pi_1 = 0.31$ for the RII_binding_1 domain. Next, we use this evolutionary model to calculate a likelihood for observing a presence-absence pattern of the RII binding domain in real data. To this end, we search for orthologs to the 14 proteins present in the seed alignment of the RII_binding_1 domain in the OMA database of orthologs (Altenhoff et al., 2010). We find 11 groups of orthologs that include proteins from the seed alignment (OMA22923, OMA22764, OMA23705, OMA26667, OMA29639, OMA36688, OMA445565, OMA445570, OMA-50272, OMA61969, OMA85984) summing up to a total of 263 sequences. A Pfam annotation of these proteins reveals that 119 sequences contain the RII_binding_1 domain. We align each OMA group individually with mafft (Katoh et al., 2005) and construct the trees with FastTree (Price, Dehal and Arkin, 2010). Figure 4.7 shows the phylogeny of orthologous sequences from OMA22923. 12 of the 26 sequences are annotated with the RII_binding_1 domain. In the remaining 14 proteins this domain is not detected. This allows now different possible interpretations concerning the evolution of this domain. Either it was present in the common ancestor protein and was lost several times, or it was not present and, thus, gained independently on some branches, or a mixture of gain and loss events. Any of these scenarios requires a considerable

number of events. Given that this domain acts like a zombie in our simulations we have to consider the possibility of false negatives. We, therefore, calculate the likelihood for the tree and the corresponding presence-absence pattern. As null model we use Q_{RII} (matrix 4.3) and as alternative model we optimize the free parameter π_0 along the tree to obtain two maximum likelihood estimations of the presence-absence pattern. The log-likelihood for OMA22923 (figure 4.7) using the null model is -24.16 and it is -30.05 when we use the alternative model ($\pi_0 = 0.91$). A likelihood ratio test shows, that we reject the null model (p-value=0.0006). From this it can be concluded that the presence-absence pattern of the RII_binding_1 domain represents real losses and gains of this domain. In support of this conclusion, the protein sequences of *Myotis lucifugus, Pteropus vampyrus, Oryctolagus cuniculus,* and *Sorex araneus* show gaps in the alignment regions where the RII binding domain is located in the other proteins. Next, we remove these four species with an obvious loss of the domain (gaps in the domain region) and redo the analysis. This time, the log-likelihood for the null model is -25.54 and it is -24.82 for the alternative model ($\pi_0 = 0.09$). The likelihood ratio test now shows, that we do not reject the null model (p-value=0.23). Thus, the remaining presence-absence pattern can also be explained by the zombie behavior of the RII binding domain.

4.4 Discussion and Conclusion

Given that a domain is physically present in a protein and that pHMM based domain identifications are without errors, we would always observe a significant similarity between the sequence and the pHMM. However, here we have shown, that annotations of proteins with domains are not always reliable and we, therefore, miss some domains. Our analysis of domains' half-lifes provides an estimation of the evolutionary stability of a domain that reveals over what evolutionary distances we expect to detect a domain. For instance, for domains with a half-life above 11 substitutions per site the situation is considerably simple. If such a domain is present in a protein one can be certain that the sequence shows a significant similarity to the pHMM. On the other hand, if this domain is not found it is highly likely that it is indeed not present. The situation is different for domains with a very short half-life. In such cases it can happen that a domain is physically present, however it does not exhibit a significant similarity to the corresponding pHMM. Our simulations show, that for the vast majority of Pfam domains pHMM based annotations have a low number of false negatives.

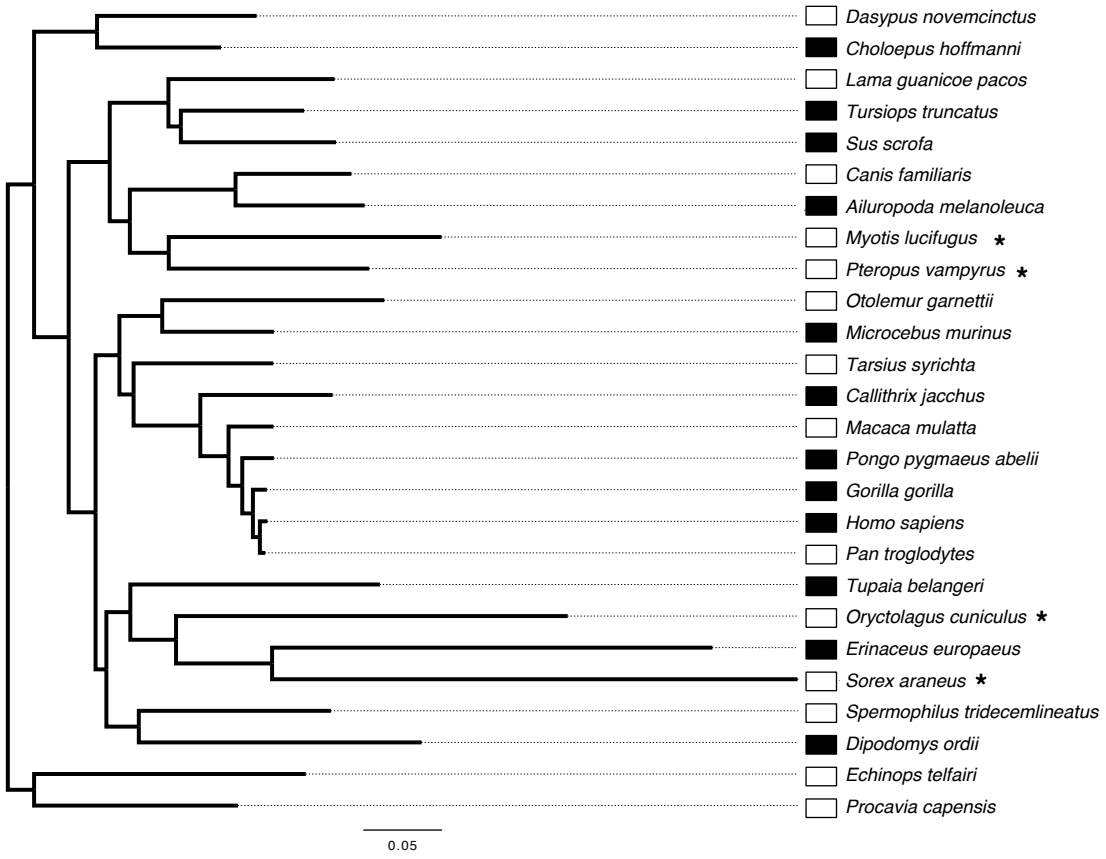


Figure 4.7: Phylogenetic tree for the ortholog group OMA22923. The boxes indicate presence (black filled box) or absence (white filled box) of the RII_binding_1 Pfam domain. * indicates gaps in the alignment regions where the RII_binding_1 domain is located in the other proteins.

However, for almost 8% of the domains we have to take care when annotating very distantly related sequences. The most extreme cases are 14 Pfam domains that are recognized in less than 50% of their instances already after 0.5 substitutions per site. These domains might be physically present in some proteins, but are not detected by a pHMM scan. Here, only experiments can give a final answer. We identified factors that influence whether a domain is stable. The most influential factor is the rate of insertions and deletions. As part of this study, we developed a parsimony approach to estimate parameters for insertions and deletions. We could show that for most protein domains insertion and deletion rates are in the range of one insertion and one deletion per 100 substitutions. Rates of 0.05 or higher are very rare. Our parsimony approach presumable underestimates the rates. However, we have shown that although the indel rates influence the stability of a domain, the half-life did generally not change with doubled rates. Obviously, extreme rates of insertion and deletion (e.g. 0.1) reduce the half-lifes substantially. However, the insertion and deletion rates and other analyzed factors explain the variability in half-lifes only partially. The question therefore remains what is the driving factor of domain stability?

Furthermore, another interesting observation resulted from this study. Some domains seem to be lost, but re-appear after additional mutations. An increase of the E-value threshold in the pHMM search would presumably solve mostly this problem of false negatives. However, this comes at the cost of an increase of false positives. Thus, a more permissive approach is not advisable in general. However, in the rare cases of zombie domains an explorative strategy is encouraged.

The half-life and the zombie behavior of domains influence the domain annotation. Hence, the observed presence-absence pattern of a domain in homologous proteins is a result of true evolutionary events like domain losses and gains plus errors in the domain detection. This needs to be considered in estimations of domain loss and gain rates, horizontal transfers, or convergent evolution. A first step is the likelihood ratio test of the evolutionary domain loss and gain models to distinguish between cases where the domain is truly absent in a phylogenetic tree and cases where it is overlooked in the domain search. In summary, this study on domain evolution reveals new insights and highlights the limitations of pHMM based domain annotations.

Chapter 5

Zygomycetes, Microsporidia, and the Evolutionary Ancestry of Sex Determination

Zygomycetes and their alleged sister taxon, the microsporidia, exclusively share the presence of a cluster of three genes encoding a sugar transporter, a High Mobility Group (HMG) type transcription factor, and an RNA helicase. In zygomycetes the HMG type transcription factor acts as the sole sex determinant. This intimately ties the evolutionary history of this gene cluster to the evolution of sex determination. Here we have unraveled the relationships of the two gene clusters by vicariously analyzing the sugar transporters and the RNA helicases. We show that if the two gene clusters share a common ancestry it dates back to the early days of eukaryotic evolution. As a consequence, the zygomycete MAT locus would be old enough to represent the archetype of fungal and animal sex determination. However, the evolutionary scenario that has to be invoked is complex. An independent assembly of the two clusters deserves therefore consideration. In either case, shared ancestry or convergent evolution, the presence of the gene cluster in microsporidia and in zygomycetes represents at best a plesiomorphy. Hence, it is not phylogenetically informative. A further genome-wide re-analysis of gene order conservation reveals that gene order is not significantly more similar between microsporidia and zygomycetes than between microsporidia and any other fungal taxon or even humans. Consequently, the phylogenetic placement of microsporidia as sister to the zygomycetes needs to be re-considered.

5.1 Introduction

In zygomycetes, an early branching fungal lineage, the mating type is determined by a single gene (Idnurm et al., 2008) encoding a High Mobility Group (HMG) type transcription factor (Thomas and Travers, 2001). The sex determining locus is flanked by two genes, one encoding a triosephosphate transporter (TPT) and the other encoding an RNA helicase. Initial analyses suggested that the arrangement of these three genes is unique to the zygomycetes. However, recently, it was reported that a similar cluster encompassing also genes for a TPT, an HMG protein, and an RNA helicase is present in microsporidia (figure 5.1). It was concluded that the two gene clusters have a common evolutionary origin, i.e. they are shared syntenic (Lee et al., 2008). The consequences of this conclusion are of relevance for two open questions.

First, is sex determination via HMG type transcription factors evolutionarily ancient? Fungal sex is determined by mating-type specific genes organized in so-called MAT loci. A number of MAT loci have been described in ascomycetes and basidiomycetes (e.g. Butler et al. 2004; Haber 1992; Lengeler et al. 2002, reviewed in Lee et al. 2010b). Based on the transcription factors present, the MAT loci are classified into three major groups: i) HMG type, ii) homeodomain type, and iii) alpha-domain type. The identification of the zygomycete MAT locus (Idnurm et al., 2008) revealed that in the earliest branching fungal lineage characterized so far, an HMG type transcription factor determines sex. This laid the odds on an HMG type MAT locus having determined sex in the last common ancestor of all fungi. The simplicity of the zygomycete MAT locus further suggested that it could resemble the archetype of fungal sex determination (Dyer, 2008b). Interestingly, also in mammals sex is determined by a single HMG type transcription factor (Haqq et al., 1993). This coincidence was taken as an indication that both fungal and mammalian sex determining systems descended from the same HMG type MAT locus in the last common ancestor of fungi and animals (Dyer, 2008b; Idnurm et al., 2008). However, this scenario is speculative. The high evolutionary rate of sex determining genes (Swanson and Vacquier, 2002) prevents a reconstruction of their evolutionary relationships already within the fungi (e.g. Lee et al. 2010a). Thus, protein sequence data provide no information about whether sex determining HMG type transcription factors in fungi and animals are derived from a single ancestral gene, or whether they are a product of convergent evolution. To still establish homology for highly diverged genes, gene order has proven helpful (Dietrich et al., 2004). The sex determining gene in zygomycetes and

the gene for the microsporidian HMG protein identified by Lee et al. (2008) are both flanked by genes encoding a TPT and an RNA helicase. They concluded that this results from shared synteny, ergo the microsporidian HMG type transcription factor is the first homolog to the zygomycete sex determinant identified in a non-zygomycete taxon. Consequently, we can now investigate the evolutionary history of HMG-driven sex determination by reconstructing the evolutionary history of the two gene clusters in zygomycetes and microsporidia.

Second, what is the exact position of microsporidia in the eukaryotic tree of life (reviewed in Corradi and Keeling 2009)? Initially, microsporidia were considered an early branching eukaryotic lineage (e.g. Cavalier-Smith 1986; Vossbrinck et al. 1987). Later findings, however, were not consistent with this view (e.g. Thomarat, Vivarès and Gouy 2004; Brinkmann et al. 2005; Gill and Fast 2006; James et al. 2006; Keeling 2009). After several taxonomic revisions it is now widely accepted that microsporidia are associated with the fungi (Corradi and Keeling, 2009). Alas, so far protein phylogenies failed to resolve whether microsporidia are sister to the fungi, or whether they fall within the true fungi. Some studies suggested a grouping of microsporidia with various fungal lineages, such as the ascomycetes, the basidiomycetes, the zygomycetes, or *Rozella* a chytridiomycete (e.g. Gill and Fast 2006; James et al. 2006; Keeling 2003; Keeling, Luker and Palmer 2000; Thomarat, Vivarès and Gouy 2004). However, their position as a sister taxon to all fungi could not be rejected (James et al., 2006). Only recently comparative genome structure analyses provided complementary information about the phylogenetic position of the microsporidia. Among all tested fungal and non-fungal species only the zygomycetes are reported to have a gene order that is more similar to that of the microsporidia than it is expected by chance (Lee et al., 2008). The microsporidian gene cluster that is shared syntenic to the zygomycete sex related locus was the most prominent example of conserved gene order. Its presence, together with the finding that microsporidia contain several genes required for meiosis, implies that microsporidia actually may have sex (Lee et al., 2010b). In summary, the analysis of gene order indicated that microsporidia share an exclusive common ancestry with the zygomycetes, and it was concluded that microsporidia evolved from ancient sexual fungi (Lee et al. 2008; Dyer 2008a, reviewed in Corradi and Keeling 2009).

Studies of both the evolutionary origins of HMG driven sex determination and the phylogenetic position of the microsporidia hinge on the microsporidian gene cluster. It is, therefore, unfortunate that the evolutionary history of this gene cluster itself is

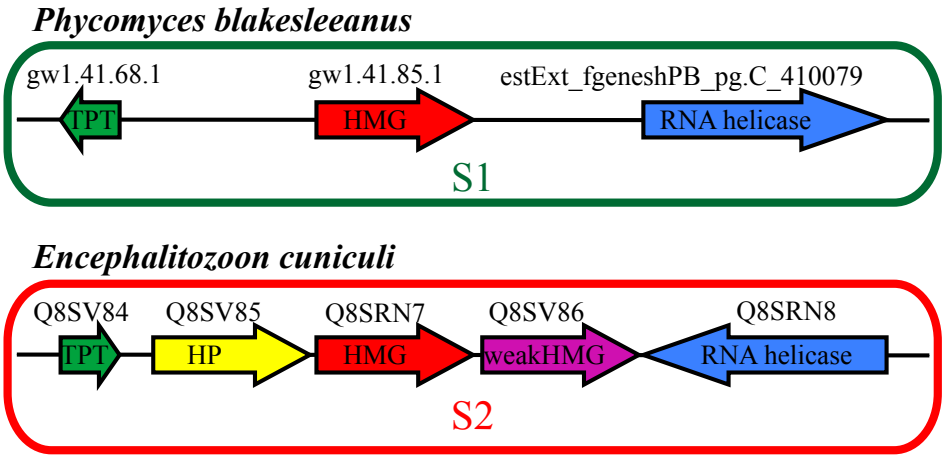


Figure 5.1: Gene arrangements in the MAT locus of *P. blakesleeanus* and in the corresponding gene cluster of *E. cuniculi*. In the sex related region of *P. blakesleeanus* (S1) the gene encoding the sex determining transcription factor (HMG) is flanked by two genes coding for a sugar transporter (TPT) and an RNA helicase, respectively. The corresponding cluster in *E. cuniculi* (S2) contains also genes for a TPT, an HMG type transcription factor, and an RNA helicase. The two additional genes in S2 encode a hypothetical protein (HP) and a protein with a weak similarity to a HMG domain protein.

not clear (Lee et al., 2010a). Here we perform a comprehensive analysis to unravel the phylogenetic relationships of the genes linked to the sex determining transcription factor in zygomycetes and their counterparts in the microsporidia. Based on the results we discuss the evolutionary history of the microsporidian gene clusters, as well as the implications for both the evolutionary ancestry of sex determination and the phylogenetic placement of the microsporidia. In a subsequent genome-wide analysis of gene order conservation we carefully re-address the proposed sister-group relationship of microsporidia and zygomycetes.

5.2 Materials and Methods

5.2.1 Ortholog Search and Phylogeny Reconstruction

We predicted orthologs to the RNA helicases and the TPTs using InParanoid v. 3.0 (Berglund et al., 2008). For a less stringent ortholog search a standard reciprocal Blast (Altschul et al., 1997) search using NCBI Blast v. 2.2.13 was performed. The gene IDs for the identified orthologs together with the corresponding data sources are summarized in Appendix table C.1. RNA helicase and TPT alignments were generated with MAFFT v. 6.833b (Katoh et al., 2005). Throughout all analyses MAFFT was used with the options –maxiterate 1000 –localpair. The resulting multiple sequence alignments were then each used for tree reconstruction. Maximum likelihood tree reconstruction was performed with RAxML v. 7.2.2 (Stamatakis, 2006) and branch support was assessed with 100 bootstrap replicates. Bayesian tree reconstruction was performed with Phylobayes v. 2.3 (Lartillot and Philippe, 2004) running two independent chains per dataset. The chains were stopped after 84,000 generations (TPT) and 120,000 generations (RNA helicase), respectively and we discarded the first 10,000 generations as burn-in. Convergence was confirmed with bpcomp from the Phylobayes package sampling every 10th tree (maxdiff: TPT: 0.02; RNA helicase: 0.08).

5.2.2 Analysis of Characteristic Sites in the Multiple Sequence Alignments

For the analysis of characteristic sites we pursued the following strategy: We aligned the sequences individually for the four ortholog groups, TPT-S1 and TPT-S2, and RNA helicase-S1 and RNA helicase-S2. The corresponding S1 and S2 alignments were then combined with MAFFT using the option –addprofile. We then called a site characteristic if in the combined S1-S2 alignment the majority of sequences from one group share an amino acid or an insertion/deletion that is not seen in the respective other group. To assess whether the microsporidian sequences share more characteristic sites with the S1 or the S2 sequences we added them individually to the appropriate S1-S2 alignment using MAFFT and the option –add. For the analysis of characteristic sites in the DEXDc domain (SM00487), we downloaded the alignment for this domain from the SMART database (Letunic, Doerks and Bork, 2009).

The alignment was converted into a profile Hidden Markov Model with hmmbuild from the HMMER package v.3 (`http://hmmer.janelia.org/`). The Logo (Schuster-Bockler, Schultz and Rahmann, 2004) for the pHMM was generated with the tool provided at `http://www.sanger.ac.uk/cgi-bin/software/analysis/logomat-m.cgi` (Appendix figure C.1). The sub-sequences in the RNA helicases corresponding to the DEXDc domain were extracted and aligned with hmmalign using the option –trim. In the resulting pHMM alignment the analysis of characteristic sites were performed as described above.

5.2.3 Identification of RNA helicases, TPTs and HMG box Proteins

We identified putative RNA helicases, TPTs, and HMG type transcription factors in *B. dendrobatidis*, *P. blakesleeanus*, and *S. cerevisiae* by searching for proteins harboring the characteristic conserved functional domains. For the RNA helicase we used the DEXDc SMART domain (SM00487), for the HMG type proteins the HMG_box PFAM domain (PF00505), and for the TPTs the TPT PFAM domain (PF03151). All three domains are present both in the proteins encoded in the sex related cluster of *P. blakesleeanus* and in the microsporidian counterparts. Domain annotations of the proteins were performed as described in section 2.6. Feature dotplots were generated with FACT (see chapter 2) and in the case of overlapping PFAM domains or overlapping SMART domain only the domains with the smallest E-value are shown.

5.2.4 Analysis of Gene Order Conservation

The extent of gene order conservation to *E. cuniculi* was determined in two zygomycete taxa *Phycomyces blakesleeanus* and *Rhizopus oryzae*, as well as in the following species: *Batrachochytrium dendrobatidis* (Fungi; Chytridiomycota; Chytridiomycetes), *Sporobolomyces roseus* (Fungi; Dikarya; Basidiomycota; Pucciniomycotina; Microbotryomycetes), *Laccaria bicolor* (Dikarya; Basidiomycota; Agaricomycotina; Agaricomycetes), *Aspergillus niger* (Dikarya; Ascomycota; Pezizomycotina; Eurotiomycetes), *H. sapiens* (Metazoa). The non-zygomycete species were chosen to complement the fungal lineages whose gene order conservation with respect to *E. cuniculi* were already found to be not conserved (Lee et al., 2008), i.e. *Saccharomyces cerevisiae* (Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes), *Ashbya gossypii*

(Dikarya; Ascomycota; Saccharomycotina; Saccharomycetes), *Schizosaccharomyces cerevisiae* (Dikarya; Ascomycota; Taphrinomycotina; Schizosaccharomycetes), *Neurospora crassa* (Fungi, Dikarya, Ascomycota, Pezizomycotina; Sordariomycetes), *Cryptococcus neoformans* (Fungi; Dikarya; Basidiomycota; Agaricomycotina; Tremellomycetes). Genome assemblies and annotated gene sets were downloaded from: Joint Genome Institute (`http://genome.jgi-psf.org/euk_cur1.html`): *A. niger*, *B. dendrobatidis*, *L. bicolor*, *P. blakesleeanus*, *S. roseus*; Broad Institute (http://www.broad.mit.edu/): *R. oryzae*; EBI (`http://www.ebi.ac.uk/integr8`): *E. cuniculi*; ENSEMBL (http://www.ensembl.org): *Homo sapiens*. For taxa for which the order of the annotated genes was not readily provided for download, we mapped the coding sequences for the predicted genes to the genome assembly using BLAT (Kent, 2002). The position of the best BLAT hit was taken as the gene position.

5.3 Results and Discussion

5.3.1 The Evolutionary History of the TPTs and the RNA helicases

In shared syntenic gene clusters, each gene shares the evolutionary history of the entire gene cluster. Thus, the split between the sex related region in zygomycetes and its counterpart in the microsporidia can be vicariously dated by analyzing the evolutionary relationships of the TPTs and RNA helicases, respectively. The HMG type transcription factors were omitted from this analysis, since they lack any phylogenetic information (Lee et al., 2010a). In the following, we refer to the zygomycete sex related gene cluster as syntenic region 1 (S1) and to the microsporidia gene cluster as syntenic region 2 (S2). Correspondingly, we refer to the respective genes as TPT-S1/RNA helicase-S1 and as TPT-S2/RNA helicase-S2 (figure 5.1).

To start our analyses we used the *Phycomyces blakesleeanus* (zygomycetes) S1 proteins to identify orthologs in *Encephalitozoon cuniculi* (microsporidia). Similarly, we searched for orthologs to the *E. cuniculi* S2 proteins in *P. blakesleeanus*. We chose the two species in which the S1 and S2 gene cluster had been initially described (Idnurm et al., 2008; Lee et al., 2008). InParanoid (Remm, Storm and Sonnhammer, 2001), one of the most reliable orthology prediction programs (Chen et al., 2007), was used for this purpose. No orthologs to the S1 proteins were found in *E. cuniculi*. In contrast,

	P. blakesleeanus[$] (Genomic location)	*E. cuniculi* (Genomic location)
TPT S1	*11516* (Scaffold 41)*	
RNA helicase S1	*80075** (Scaffold 41)*	
TPT S2	4053 (Scaffold 4)	*Q8SV84 (Chr VI)*
	19565 (Scaffold 8)	
RNA helicase S2	14395 (Scaffold 1)	*Q8SRN8 (Chr VI)*

[$] JGI Gene ID
* Accession-No. ABX27908.1
** Accession-No. ABX27910.1

Table 5.1: Ortholog pairings for the *P. blakesleeanus* and *E. cuniculi* S1 and S2 genes. Genes located in the sex related gene cluster of zygomycetes and in its microsporidian counterpart are emphasized.

both S2 proteins have orthologs in *P. blakesleeanus*. The corresponding genes are, however, not located in the sex determining region but resided on different scaffolds in the *P. blakesleeanus* genome assembly (table 5.1). Note that the results did not change when we reduced the stringency of the ortholog search by performing only a reciprocal best Blast hit search and omitted the additional filtering steps invoked by InParanoid (Remm, Storm and Sonnhammer, 2001). Thus, neither TPT-S1/TPT-S2 nor RNA helicase-S1/RNA helicase-S2 were identified as ortholog pairs.

We assessed next when during evolution the corresponding genes in the S1 and S2 clusters of zygomycetes and microsporidia have separated. A screen in 15 plant, animal, and fungal species for orthologs to each of the four genes resulted in four disjoint ortholog groups (c.f. Appendix table C.1). We combined all RNA helicases and all TPTs, respectively, and conducted maximum likelihood (ML) tree reconstructions for both data sets. The resulting trees are shown in figure 5.2 (RNA helicases) and Appendix figure C.2 (TPTs). In both trees the S1 orthologs and the S2 orthologs are placed into two well-supported clades (RNA helicase: BS=100, TPT: BS=100). A complementary Bayesian analysis corroborated the results (RNA helicase: BPP=1; TPT: BPP=1; trees not shown). All four clades, corresponding to the four ortholog groups, contain sequences from animals, fungi, and plants. This indicates that the genes in the zygomycete sex related region have separated from their microsporidian homologs already before the three eukaryotic kingdoms emerged.

To further substantiate the hypothesis that the zygomycete S1 genes and the microsporidian S2 genes are evolutionarily only very distantly related, we analyzed the

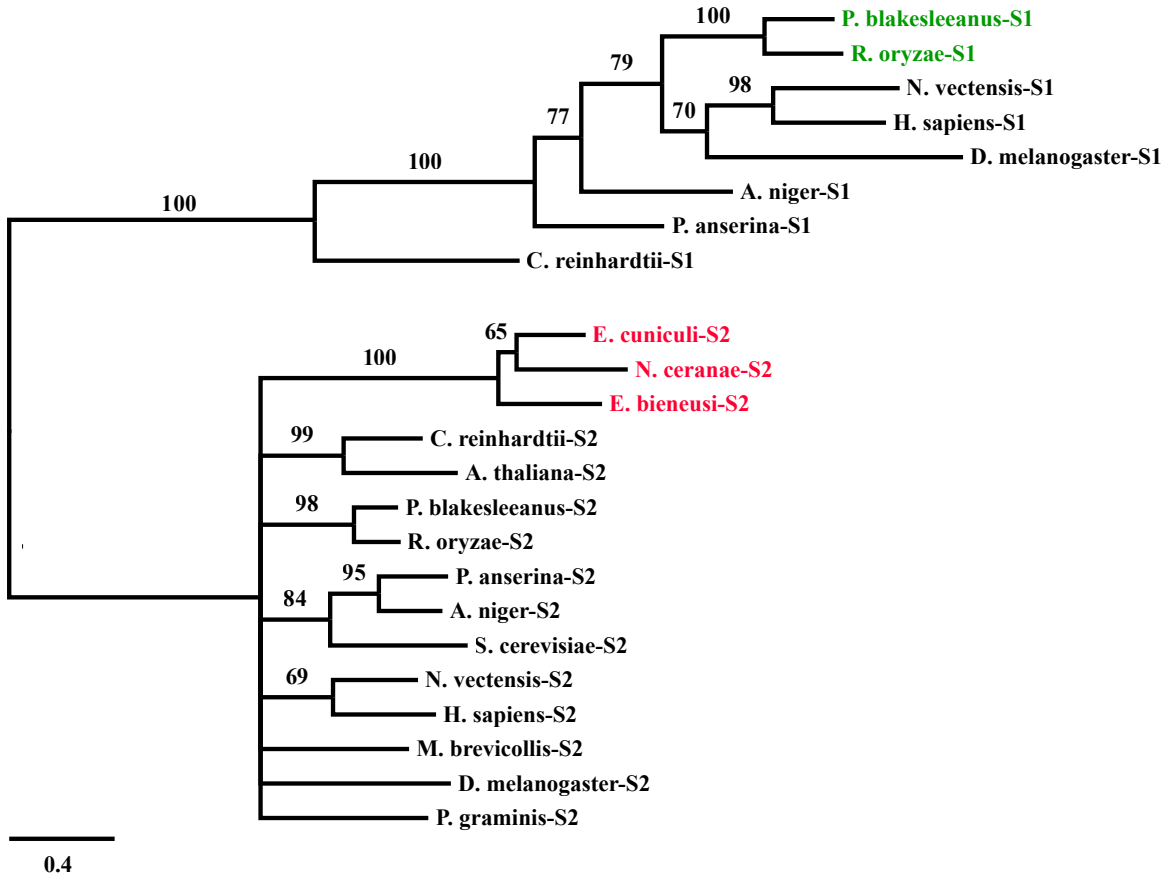


Figure 5.2: Maximum likelihood tree of the S1 and S2 RNA helicases. Sequences in the zygomycete sex related region are labeled in red, sequences in the corresponding region of the microsporidia are labeled in green. Branch labels denote bootstrap support values.

	#char. sites*	Ortholog group	*E. cuniculi*	*A. locustae*	*N. ceranae*	*E. bieneusi*
RNA helicase	390	S1	24	-	22	17
		S2	197	-	194	206
TPT	58	S1	0	0	1	3
		S2	22	23	21	15

* total number of characteristic sites distinguishing the S1 from the S2 sequences.

Table 5.2: Number of characteristic sites conserved in the microsporidian RNA helicases and TPTs.

protein sequence alignments. We first removed all microsporidian RNA helicases. We then aligned the S1 RNA helicases and the S2 RNA helicases separately and subsequently combined them using a profile-to-profile alignment. In the resulting alignment we searched for evolutionarily conserved sites that characterize the S1 and the S2 ortholog groups. We called a site characteristic if in the combined alignment of two ortholog groups the majority of sequences from one group share an amino acid or an insertion/deletion that is not seen in the respective other group. 390 characteristic sites distinguish the S1 RNA helicases from the S2 RNA helicases. We applied the same procedure to the TPTs and identified 58 characteristic sites. To assess whether the microsporidian sequences display any marked similarity with either the S1 or the S2 sequences, we aligned each of them to the corresponding combined S1-S2 alignment. This revealed that the *E. cuniculi* RNA helicase shares 197 characteristic sites with the S2 RNA helicases and only 24 with the S1 RNA helicases. Similarly, the *E. cuniculi* TPT shares 22 of the 58 characteristic sites with the S2 TPTs and 0 with the S1 TPTs. The same results were obtained with the other microsporidian sequences (table 5.2). Thus, the proteins encoded in the microsporidian gene cluster share a substantial extent of sequence conservation with the other S2 sequences. In contrast they have virtually nothing in common with the S1 sequences.

We pursued the analysis of characteristic sites in greater depth exemplarily for the RNA helicases. We extracted the subsequences matching to the DEXDc SMART domain (Letunic, Doerks and Bork, 2009), the functional domain of DEAD and DEAH box helicases, and performed a pHMM alignment. A section from this alignment is shown in figure 5.3. Figure 5.3 shows clearly that the marked sequence conservation between the microsporidian RNA helicase and the S1 RNA helicases is present also in the functional domain of the proteins. The helicase domain of the RNA helicase encoded in the zygomycete sex related locus contains two short sequence motifs IQGPPGT-

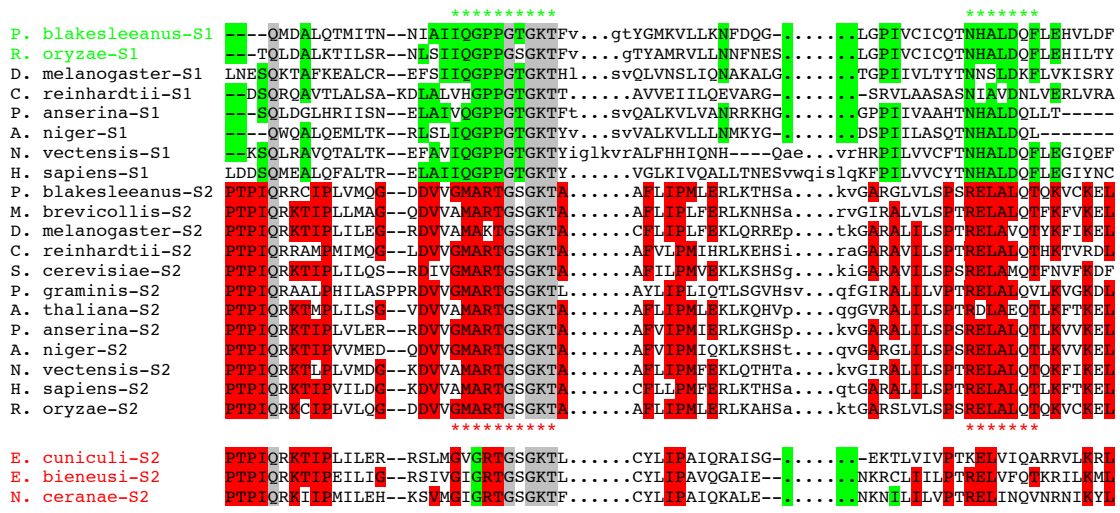


Figure 5.3: Section of the pHMM guided multiple sequence alignment of the DEXDc domain in the S1 and S2 RNA helicases. Characteristic sites for the S1 RNA helicases are labeled in green and for the S2 RNA helicases are labeled in red. The microsporidian sequences display almost exclusively characteristic sites of the S2-type RNA helicases. Amino acids in the grey shaded columns are conserved in all sequences and are specific for the DEXDc domain (c.f. Appendix figure C.1). Dashes denote deletion states in the pHMM alignment, and lower case letters opposed to dots denote insertion states. Green and red sequence labels denote the sequences in the zygomycete sex related cluster and its microsporidian counterpart, respectively. The green stars denote two evolutionarily conserved motifs of the RNA helicase in the sex related region of *P. blakesleeanus*. The red stars denote the corresponding motif in the S2 RNA helicases.

GKT and NHALDQF that are almost perfectly conserved among all sequences in the RNA helicase-S1 group (green stars in figure 5.3). A parsimony argument implies that these motifs were already present in the most recent common ancestor of these sequences. At the same alignment positions two evolutionarily highly conserved motifs are also seen in the helicase domains from the S2 group, which however are distinct from the S1 motifs (*P. blakesleeanus*: GMARTGSGKT and RELALQT; red stars in figure 5.3). The microsporidian sequences display slight variants of the S2 motifs.

Next, we contrasted the feature architecture of the *E. cuniculi* RNA helicase with the feature architecture of the *P. blakesleeanus* RNA helicase S1 and S2, respectively, with FACT (chapter 2). Figure 5.4 shows the feature dotplot of both comparisons. The *E. cuniculi* RNA helicase consists of a DEXDc SMART/DEAD PFAM domain,

a HELICc SMART/Helicase_C PFAM domain (Letunic, Doerks and Bork, 2009; Finn et al., 2010), and a K-rich region. All three RNA helicases consist of the already described DEXDc SMART domain (Letunic, Doerks and Bork, 2009). However, in the case of *E. cuniculi* S2 and *P. blakesleeanus* S1 this is the only shared domain. In contrast, the *P. blakesleeanus* S2 proteins consists of all domains that are present in the *E. cuniculi* RNA helicase. Namely, they share the DEXDc SMART/DEAD PFAM domain, the HELICc SMART/Helicase_C PFAM, and the K-rich stretch. Moreover, the features are in the same order. Thus, the RNA helicase encoded in the microsporidian gene cluster shares a very similar feature architecture with the *P. blakesleeanus* RNA helicase S2.

The evolutionary history of the genes in the zygomycete sex related region (gene cluster S1) and its shared syntenic counterpart in the microsporidia (gene cluster S2) has been investigated before (Lee et al., 2010a). However, the authors did not decisively conclude whether the corresponding genes in the two clusters are paralogs or extremely diverged orthologs. We followed a tripartite approach to solve this issue. Orthology predictions using both InParanoid (Berglund et al., 2008) and a less stringent reciprocal best Blast hit search failed to recognize the S1 RNA helicase of *P. blakesleeanus* and the S2 RNA helicase of *E. cuniculi* as orthologs, and the same applies to the TPTs. A phylogenetic tree reconstruction placed S1 and S2 sequences in distinct clades where each clade contained sequences from fungi, animals, and plants. This already suggests an early separation of the S1 and S2 genes that predates the split of microsporidia and fungi. However, the validity of conclusions drawn from both orthology assignment and phylogenetic tree reconstruction can be compromised by the high evolutionary rate particularly of microsporidian proteins (e.g. Brinkmann et al. 2005). Hence, we added the analysis of evolutionarily conserved characteristic sites as a third line of evidence. We found that the microsporidian sequences share substantially more characteristic sites with the S2 sequences then with the S1 sequences. This finding seamlessly integrates with the results from the ortholog search and the tree reconstruction. Thus, all evidences point towards a common ancestry of the microsporidian genes and the respective other S2 genes to the exclusion of the S1 genes. On the contrary, they are not compatible with the hypothesis that the microsporidian genes are extremely diverged orthologs of the genes in the zygomycete sex related cluster, as it has been suggested before (Lee et al., 2010a).

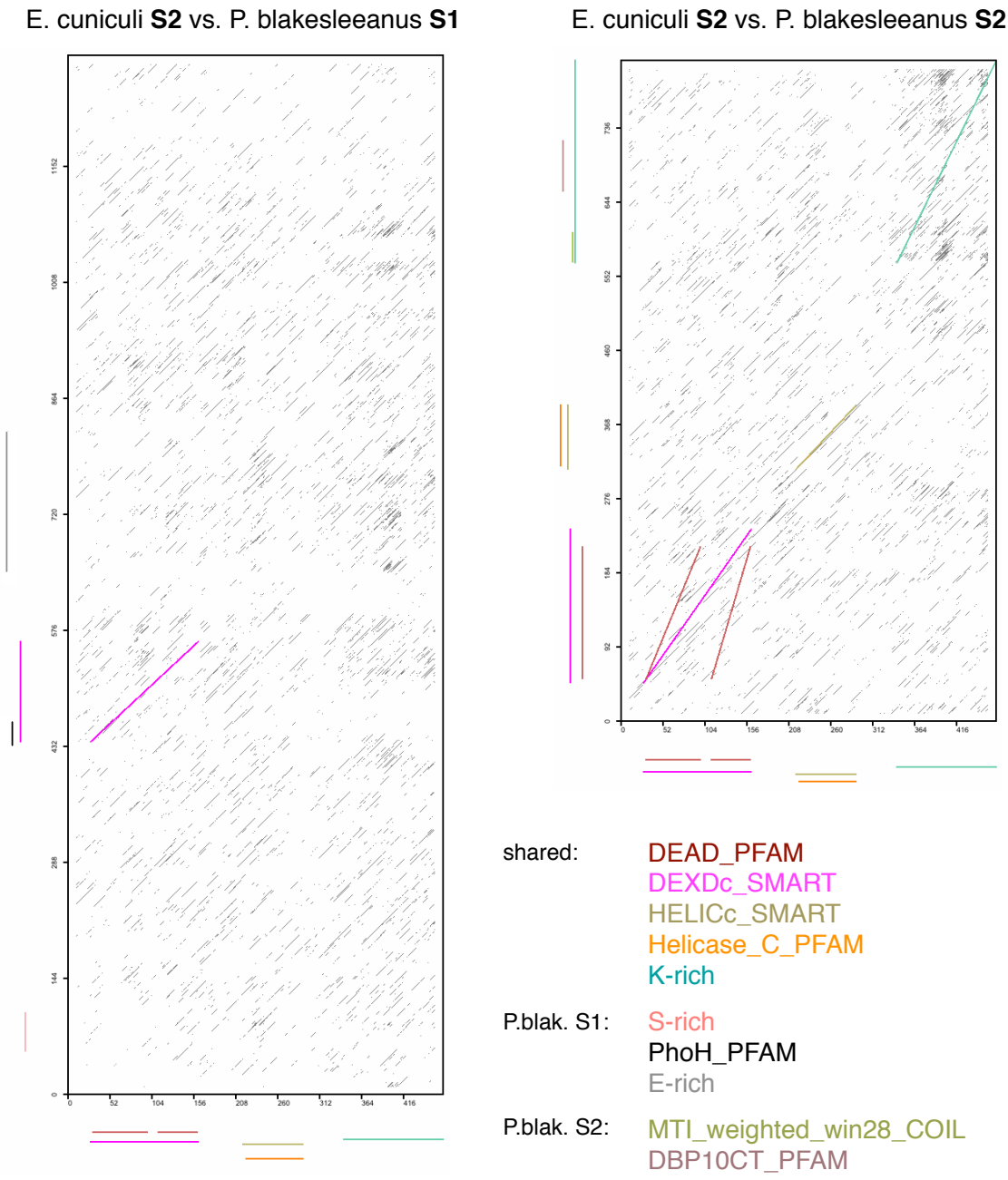


Figure 5.4: Feature dotplot contrasting the feature architecture of the *E. cuniculi* RNA helicase and the *P. blakesleeanus* RNA helicase S1 (left hand side) and S2 (right hand side), respectively. The feature architecture of the *E. cuniculi* protein is shown on the both x-axes and the *P. blakesleeanus* proteins are shown on the y-axes.

5.3.2 The Implications of Shared Synteny

Our analyses have revealed that both gene pairs, RNA helicase-S1/S2 and TPT-S1/S2, separated early in eukaryotic evolution and long before the split of zygomycetes and microsporidia. How can this result be reconciled with the proposed shared synteny of the genomic regions the genes reside in? To do so, we need to assume that an ancestral TPT-HMG-RNA helicase gene cluster existed already in the common ancestor of plants, animals and fungi. This gene cluster gave rise to two independently evolving copies in this primordial species. One copy each was retained with conserved gene order in microsporidia and zygomycetes and the second copy was reciprocally lost in the two lineages. In all other taxa analyzed so far both copies of the ancestral gene cluster were lost. Thus, conditioned on the shared synteny assumption we provide now for the first time evidence that the evolutionary history of the zygomycete MAT locus can be traced back to the early days of eukaryote evolution. It would be therefore old enough to represent the archetype of fungal and animal sex determination.

However, a considerable number of evolutionary events need to be assumed to uphold the initial assumption of shared synteny. It is, therefore, worthwhile to consider an alternative hypothesis. The two gene clusters in zygomycetes and microsporidia may have been assembled twice and independently during evolution and are not shared syntenic. In this case the presence of a microsporidian HMG type transcription factor flanked by a TPT and an RNA helicase allows no conclusions about the evolutionary history of HMG driven sex determination. Although convergent evolution appears on the first sight unlikely we will now show that it is not implausible. *P. blakesleeanus* has 132 different proteins with a DEXDc smart domain (Letunic, Doerks and Bork, 2009), the characteristic feature of the RNA helicase-S1. Further 30 proteins contain a HMG box, and 17 TPTs exist. The numbers for *Batrachochytrium dendrobatidis* (chytridiomycetes) and *Saccharomyces cerevisiae* (ascomycetes) are similar (RNA helicases: 115/123; HMG: 10/9; TPT: 8/11). This indicates that these genes were as abundant in the last common ancestor of all fungi. Microsporidia evolved from the ancestor shared with the fungi by undergoing a massive genome compaction and an associated loss of genes (Katinka et al., 2001). Still *E. cuniculi* has retained 48 helicases, 3 TPTs and 2 HMG type proteins. It can be easily imagined that the re-organization of the microsporidian genome during its evolution has just by chance placed any of the genes encoding RNA helicases, TPTs, and HMG type transcription factors next to each other. By that a gene cluster emerged resembling that of the zygomycete sex

related region.

At the moment it is impossible to decide which of the two scenarios, ancient relationships or convergent evolution, applies to the two gene clusters. Presumably only an in-depth functional analysis of the individual proteins in microsporidia, with a focus on the HMG type transcription factor will help to shed further light on this matter.

5.3.3 Are Microsporidia and Zygomycetes Monophyletic?

How do our findings relate to the debate about the phylogenetic position of microsporidia? The conservation of gene order, and in particular the presence of the microsporidian gene cluster resembling the zygomycete sex related locus, has served as argument to place the microsporidia next to the zygomycetes in the fungal tree of life (Lee et al., 2008, 2010b). However, we have shown that the suggested shared synteny traces the two gene clusters back to an ancient gene cluster in the common ancestor of plants, animals and fungi. In cladistic terms they represent a shared ancestral character, or a plesiomorphy. Plesiomorphies are phylogenetically not informative (Hennig, Davis and Zangerl, 1966). Hence, they cannot serve as supporting evidence for the proposed monophyly of microsporidia and zygomycetes (but see Lee et al. 2008, 2010a). This emphasizes that phylogenetic inferences based on gene order conservation are problematic when the exact evolutionary relationships of the genes remain uninvestigated. Unfortunately, the only quantitative analysis of gene order conservation to determine the phylogenetic position of microsporidia used a unidirectional BlastP search (E-value cutoff 10E-5) for homology inference (Lee et al., 2008). A comparison between several orthology prediction methods has shown that orthology assignments based on unidirectional BlastP searches are wrong in 50% of the cases (Chen et al., 2007). This bears the risk that a considerable fraction of the identified zygomycete-microsporidia gene pairs comprise paralogs. To assess whether this has any consequences for the conclusions of this study we re-investigated the extent of gene order conservation between microsporidia and zygomycetes. In brief, we used InParanoid (Remm, Storm and Sonnhammer, 2001) for orthology prediction. In contrast to the unidirectional BlastP search, InParanoid has a reported false positive rate of only 7% (Chen et al., 2007). We established the evolutionary relationships between the genes of *Encephalitozoon cuniculi* and two zygomycete species, *Phycomyces blakesleeanus* and *Rhizopus oryzae*. Four further fungal species (*Batrachochytrium dendrobatidis*, *Sporobolomyces roseus*, *Laccaria bicolor*, *Aspergillus niger*) and human

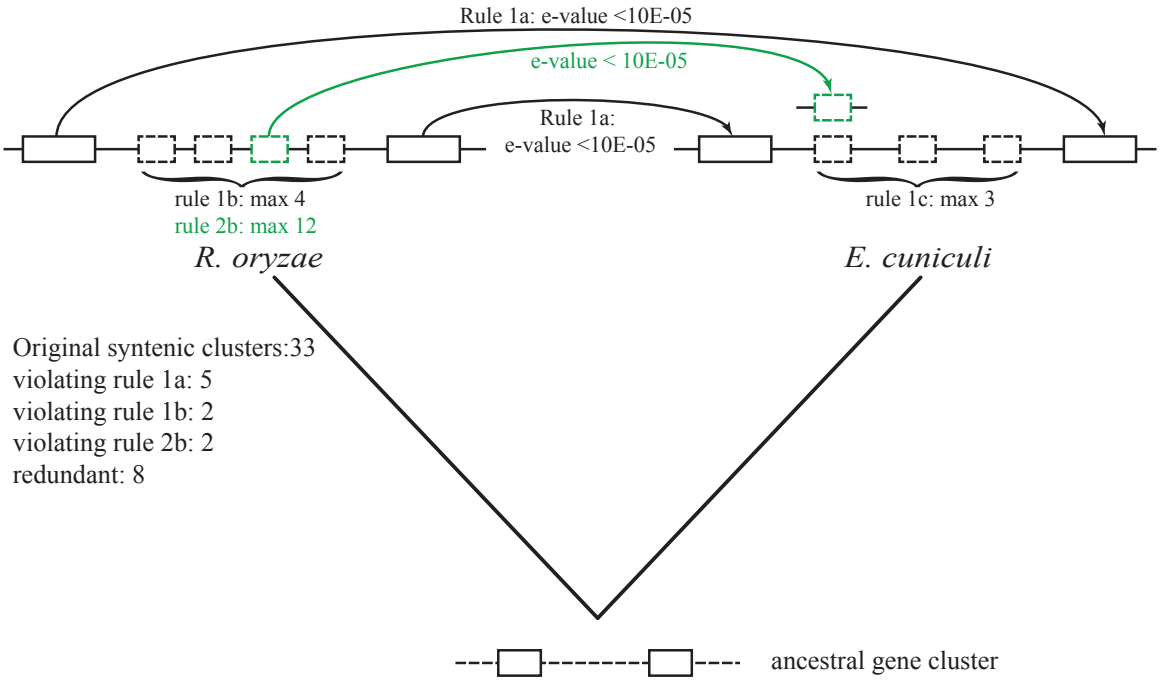


Figure 5.5: Re-analysis of the proposed shared syntenic gene clusters between *R. oryzae* and *E. cuniculi*. Two decision rules were used to assign shared synteny (c.f., figure S5 in Lee et al. 2008). Rule 1 requires that (1a) two *R. oryzae* genes have a Blast hit with an E-value < 10E-05 in *E. cuniculi*, (1b) the two *R. oryzae* genes must be separated by no more than 4 genes, and (1c) the corresponding *E. cuniculi* genes must be separated by no more than 3 genes. Rule 2 extends rule 1 if one of the intervening *R. oryzae* genes also has a Blast hit (E-value < 10E-05) in *E. cuniculi*. In this case, the *R. oryzae* genes must be separated by no more than 12 genes (2b). Please note that in the original publication the decision rules where described for a search in the opposite direction, i.e., with the *E. cuniculi* proteins as query. However, the data presented in table S1 of Lee et al. (2008) (cf. also Appendix table C.2) are not compatible with this direction of the search. Hence, we adjusted the decision rules to fit the data.

were analyzed to investigate whether the extent of gene order conservation to *E. cuniculi* varies between species. For 461 *E. cuniculi* genes an ortholog was present in all seven taxa. In *E. cuniculi* these 461 genes are arranged in 674 gene pairs with no more than 3 genes in-between. We then considered a microsporidian gene pair as conserved if its orthologs in the non-microsporidian species are separated by no more then 9 intervening genes. The results are summarized in table 5.3. From the 674 *E. cuniculi* gene pairs only 4 were recovered in *R. oryzae*, 5 were present in *B. dendrobatidis*, a chytridiomycete, and 3 in humans. In essence, no marked conservation of gene order between *E. cuniculi* and zygomycetes is seen. However, again in the light that orthology prediction for the fast evolving microsporidia is hard, our approach bears the risk of being overly stringent. We thus may lack the sensitivity for a meaningful analysis of gene order conservation. To address this point, we re-analyzed the existing data that were obtained with the unidirectional Blast searches (table S1 in Lee et al. 2008). The results are summarized in figure 5.5 and in Appendix table C.2. In their relaxed stringency analysis Lee et al. (2008) found 33 clusters with conserved gene order in *E. cuniculi* and *R. oryzae*. Of these clusters 5 do not fulfill the E-value cutoff of 10E-5 in the original data. Further 4 clusters exceed the maximally allowed number of intervening genes in *R. oryzae*. Of the remaining 24 clusters, 13 *R. oryzae* clusters point to only five clusters in *E. cuniculi*. The corresponding cluster must have duplicated on the *R. oryzae* lineage after the split from the microsporidia. Hence, they can be counted only once each. This reduces the number of independent shared syntenic regions between the two species to 16 what would be expected by chance (Lee et al., 2008). Thus, a proper analysis of Lee et al.'s (2008) data provides no evidence that the gene order is more conserved between microsporidia and zygomycetes than between microsporidia and any other fungal taxon or even humans. As a consequence, the proposed placement of microsporidia as a sister to the zygomycetes receives no support by the data. The question remains therefore open where to confidently place this enigmatic taxon in the fungal tree of life.

In summary, our study has revealed what can and what cannot be inferred from the observation that microsporidia harbor a gene cluster closely resembling the sex related region of zygomycetes. If we take shared synteny for granted, our results trace the zygomycete sex related region back to the early days of eukaryote evolution. It may therefore indeed comprise the archetype of animal and fungal sex determination. However, the evolutionary scenario that has to be invoked is complex. Thus, sacrificing the shared synteny assumption may lead to a more parsimonious hypothesis, i.e., that

		Number of intervening genes**									
		0	1	2	3	4	5	6	7	8	9
E. cuniculi: 674*	*Batrachochytrium dendrobatidis*	3		1	1						
	Aspergillus niger	1			1		1				
	Laccaria bicolor	1						1			
	Sporobolomyces roseus	1			1			1			
	Rhizopus oryzae	2	1	1							
	Phycomyces blakesleeanus						1				
	Homo sapiens	1				1					1

Table 5.3: Number of *E. cuniculi* gene pairings recovered in six fungi and humans. *Gene pairs separated by no more than three genes with orthologs in 6 fungal taxa and H. sapiens. **Number of intervening genes between two orthologs to an *E. cuniculi* gene pair.

the two gene clusters arose independently through convergent evolution. Independent of the true evolutionary relationships of the two gene clusters, however, one observation stands out. Their presence in zygomycetes and microsporidia represent, at best, a plesiomorphy and provide no information about the phylogenetic relationships of zygomycetes and microsporidia. As there is no further evidence for a significant conservation of gene order between the two taxa, the proposed alliance of microsporidia and zygomycetes remains speculative.

Chapter 6

Summary and Outlook

The general purpose of this thesis was to investigate and develop methods to identify functionally equivalent proteins. We, therefore, set of by introducing strategies for ortholog detection for a given query protein (chapter 1.1). The resulting presence-absence pattern of orthologs across the analyzed species constitutes the phylogenetic profile of this protein (Pellegrini et al., 1999). The functional similarity between these orthologs can then be assessed with a dotplot that contrasts the feature architecture of two proteins (presented in chapter 2). We then described how searches for proteins of similar feature architecture with FACT (chapter 2) complement phylogenetic profiles in regions of the phylogeny where orthologs were not found. Still, ortholog searches are the prevalent method to address questions about the evolutionary history of individual proteins for instance when during evolution they presumably emerged or when they may have been secondarily lost. However, we have to cope with one essential detail: Not detecting a protein is not synonymous to the true absence of a protein. A central question in this thesis was, therefore, how to interpret the apparent absence of homologs to a query protein in a given species. In particular, we were interested in the question over what evolutionary distances can orthologs be found at all with the method at hand. We started our approach to address this problem by developing a software, REvolver, that facilitates the simulation of protein evolution considering domain constraints. We investigated cases where a protein domain was not detected due to a lack of sensitivity in the domain search (chapter 4). This analysis revealed domains that are evolutionarily stable and those that lose their domain characteristics already after a few mutations. The latter ones might cause problems when they have to be identified over large evolutionary distances.

The questions 'Over what evolutionary distances can orthologs be found?' is, how-

ever, not yet answered. This problem may be solved by extending the approach to evaluate the evolutionary stability of individual protein domains (chapter 4) to the entire protein. To investigate whether we expect to identify an existing ortholog p_B in species B to a particular protein p_A in species A, we have to consider three additional issues. First, the definition of the term half-life, introduced in chapter 4, needs to be adjusted to the detection of orthologs via sequence similarity. Second, the evolutionary distance between two species A and B needs to be assessed. Since fast evolving proteins aggregate many times more mutations than slowly evolving proteins over the same timespan, we, third, need to assess the specific rate of evolution k_{p_A} for p_A. The half-life can then be scaled according to the protein's specific evolutionary rate. The resulting evolutionary traceability of p_A represents an estimate whether we expect to find p_B in B or whether the protein is assumed to have accumulated too many mutation to be identified via sequence similarity based methods. As a result, the phylogenetic profile of any protein p can be interpreted in a more meaningful way. In the following we outline a possible approach to estimate the evolutionary traceability of a particular protein.

First, we calculate the pairwise distances $d(A, S)$ between species A and any species $S \in \mathcal{S}$, where $B \in \mathcal{S}$. A method to obtain the evolutionary distance between a pair of species is exemplified in Appendix D where we first search for orthologs, then calculate the evolutionary distances between the orthologs, and finally build the average. The average evolutionary distance between the orthologs is taken as $d(A, S)$. We then calculate the evolutionary traceability of p_A and evaluate whether the probability to find p_B is higher than the probability to miss it as follows (cf. figure 6.1):

1. Determine $\mathcal{P}$, the set of detectable orthologs to p_A in $\mathcal{S}$. $\mathcal{C} \subset \mathcal{S}$ consists of those species for which an orthologs was found. If $B \in \mathcal{C}$ we are done.
2. Estimate protein specific parameters based on $\mathcal{P} \cup \{p_A\}$:
   a) Compute $d(p_A, p)$ for all $p \in \mathcal{P}$.
   b) Calculate the scaling factor

$$k_{p_A} = \frac{\sum_{S \in \mathcal{C}} d(A, S)}{\sum_{p \in \mathcal{P}} d(p_A, p)}. \tag{6.1}$$

If $k_{p_A} < 1$, p_A is faster evolving that the average protein and if $k_{p_A} > 1$, p_A

is more slowly evolving than the average protein.

c) Compute a multiple sequence alignment for $\mathcal{P} \cup \{p_A\}$ and estimate the protein specific insertion rate λ_I and deletion rate λ_D using the parsimony approach described in chapter 4.

d) Annotate protein p_A with domains (e.g. Pfam; Finn et al. 2010).

3. For 100 times:

a) Use REvolver to simulate the evolution of protein p_A under domain constraints (chapter 3).

b) Perform a Blast search using the simulated sequence as query against the proteome of species A and check whether p_A is the best Blast hit.

4. Calculate the half-life for p_A (HL_{p_A}) as the number of substitutions where less than 50% of the simulated sequences identify p_A as the best Blast hit.

5. The traceability for protein p_A is finally

$$\tau_{p_A} = HL_{p_A} * k_{p_A}. \tag{6.2}$$

Thus, τ_{p_A} is now scaled according to the distances between the species in $\mathcal{C}$.

Assuming that p_B is present in species B, we conclude as follows: If $d(A, B) \leq \tau_{p_A}$ the probability to find p_B is ≥ 0.5. Otherwise the probability to find p_B is < 0.5 and we, thus, have a high chance to miss p_B in an ortholog search.

To illustrate the procedure, we calculated the traceability of the *S. cerevisiae* SUS1 protein. SUS1 is involved in the re-location of activated genes to the nuclear pore complex (Rodriguez-Navarro et al., 2004). Our approach obtained the following results: $HL_{\text{SUS1}} = 4.00$; $k_{\text{SUS1}} = 0.41$; $\tau_{\text{SUS1}} = 1.65$ (see Appendix D). Thus, after 4 substitutions per site, less than 50% of the simulated SUS1 sequences obtained the yeast SUS1 as best Blast hit. The scaling factor of 0.41 shows that SUS1 is faster evolving than the average protein. Finally, the calculated τ_{SUS1} tells us that we expect to find an ortholog to SUS1 in any other species only, if its evolutionary distance to yeast is ≤ 1.65 substitutions per site. To demonstrate the results also graphically, we mapped the traceability of the yeast SUS1 onto a phylogenetic tree of 244 eukaryotes (figure 6.2). Among these, the *E. bieneusi* (microsporidia) proteins have with 2.37 substitutions per site the longest average evolutionary distance to its yeast orthologs. *E. bieneusi* is, therefore, the most diverged species in this phylogeny. Thus, we do

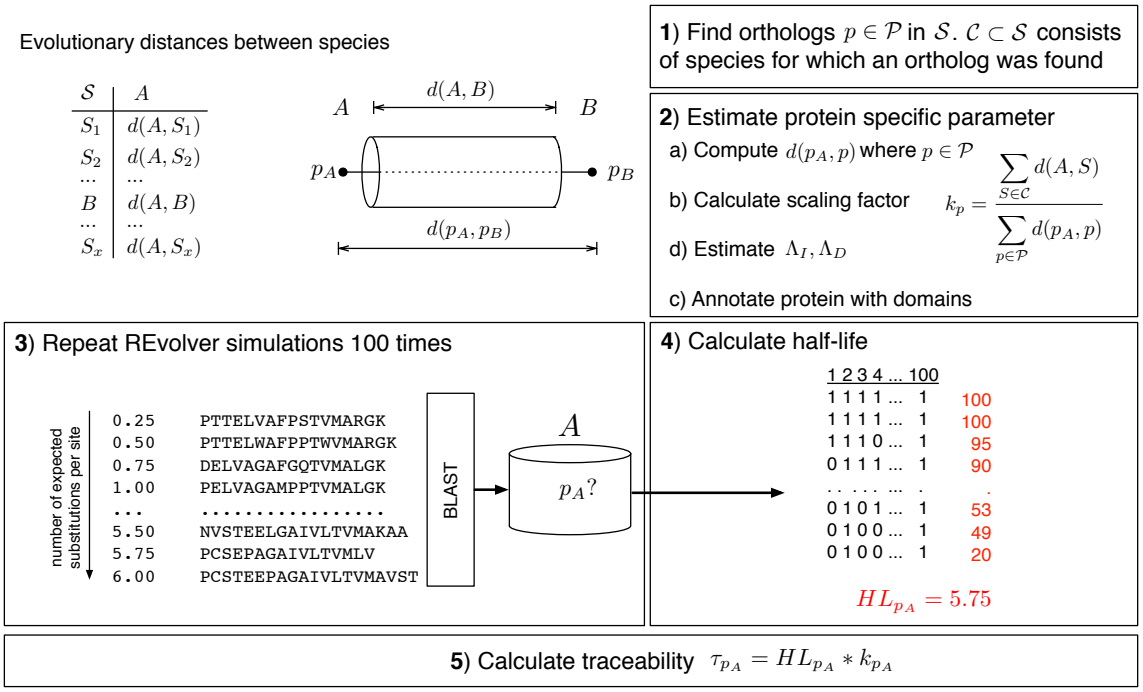


Figure 6.1: Workflow to estimate the evolutionary traceability for p_A. Given the evolutionary distances between species A and every species in $\mathcal{S}$, we determine whether or not we expect to find an ortholog p_B in species B to protein p_A. The individual steps are numbered consecutively from 1 to 5 and described in detail in the main text. The table in step 4 shows whether the best Blast hit using the simulated sequence as query was p_A (1) or not (0). The row sums are given in red. After 5.75 substitutions per site only 49 out the 100 simulated sequences identified p_A as best Blast hit. The half-life of p_A in this example is therefore 5.75. In step 5, the traceability is calculated and if $d(A, B) \leq \tau_{p_A}$, p_B is expected to be found.

not expect to find orthologs to the yeast protein in distantly related species like microsporidia, cryptophyta, or other protists (black labels in figure 6.2). SUS1 orthologs were mainly identified in several fungi and animals (red labels in figure 6.2). However, no orthologs where, for instance, found in the Caenorhabditis clade or in the pezizomycotina. However, the reason for not finding orthologs may be different. For the pezizomycotina, we expect to find SUS1 orthologs, since the average protein distance between any species of this systematic group and yeast is below the traceability of SUS1 (green labels). Thus, it is likely that pezizomycotina lost SUS1. In contrast, in the Caenorhabditis clade we do not expect to find SUS1 orthologs (black labels) since too many substitutions have accumulated. Consequently, we still do not know whether the protein is truly absent in Caenorhabditis species. More sensitive searches for SUS1 proteins (e.g. via FACT; see chapter 2) together with confirmative experiments are necessary to get a final answer.

In summary, the protein traceability is a first approach to predict over what evolutionary distances orthologous proteins can be identified via sequence similarity. The protein traceability facilitates an important step towards a more reliable interpretation of phylogenetic profiles. Moreover, it also highlights the limits of sequence similarity based approaches and depicts those cases where more sensitive methods to identify proteins should be applied. To conclude, the presented approaches and ideas contribute to our understanding in phylogenetic profiles, protein evolution, and organismal and functional evolution.

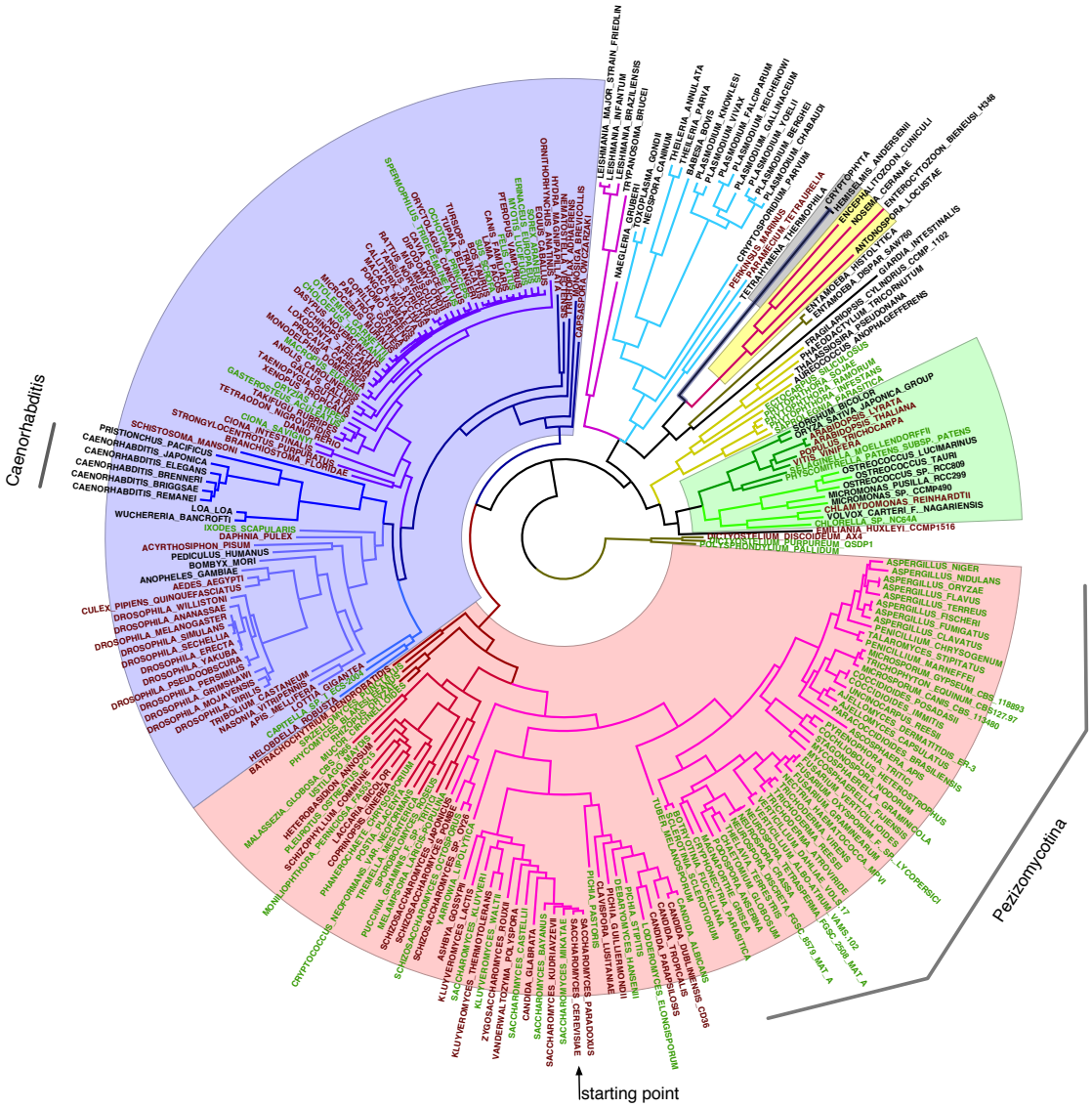


Figure 6.2: Unrooted phylogenetic tree of 244 eukaryotes. Metazoan and their two closest relatives *C. owczarzaki* and *M. brevicollis* are highlighted in blue, fungi in red, plants in green, microsporidia in yellow, and cryptophyta are highlighted in grey. Red labels indicate species where an ortholog to the SUS1 protein in *S. cerevisiae* (arrow) was found. Green labels indicate species where a SUS1 ortholog was not found, however the probability to find it, if it is present, is ≥ 0.5. In the remaining species (black labels) neither an ortholog was found nor it is expected to be found.

Acknowledgments

I take this opportunity to thank those who made this thesis possible. Special thanks go to both of my supervisors Ingo and Arndt for their invaluable guidance over the last years. Thank you Ingo, for constantly showing interest in my work and for help, advice, motivating words at any time and not least for the nice chats over a cup of coffee. Among many other things, you taught me to give apparently negative results a chance to turn into something really interesting.

I want to thank Arndt for his continuous support and for giving me enough space to develop my own ideas and follow my interests. Thank you for your advice and for giving me the opportunity to attend several conferences where I also got to meet many researchers from a variety of different areas.

I thank all CIBIV members for a nice working environment and a lot of fun at the one or other happy hour. Anne, Minh Anh, and Wolfgang have been the best office-mates I could think of. I acknowledge the technical support from Wolfgang, Minh for adapting his code for my purposes that saved me a lot of time, and Ines for proofreading parts of this thesis.

I want to thank my family and friends who supported me in any way. Not least, I want to thank Daniel for simply being there and I deeply appreciate the time outside work with you.

Curriculum Vitae

Tina Köstler

Center for Integrative Bioinformatics Vienna (CIBIV)
Max F. Perutz Laboratories
Dr. Bohr Gasse 9
1030 Vienna, Austria
Phone: +43 1 / 4277 24027
Email: tina.koestler@univie.ac.at, tina.koestler@gmail.com
Homepage: www.cibiv.at/∼tina

Date of birth November 14, 1984

Place of birth Braunau am Inn, Austria

Nationality Austria

Education

2008 Diploma in Bioinformatics

2004-2008 Student at the University of Applied Sciences, Hagenberg, Austria

2004 Matura focused on Mechatronics

1999-2004 Student at the secondary school (HTL Braunau), Austria

Research Experience

November 2008 - present PhD Student at the CIBIV

Supervisors:

Prof. Arndt von Haeseler

Dr. Ingo Ebersberger

January 2008 - July 2008 Diploma student at the CIBIV

Supervisors:

Dr. Ingo Ebersberger

Mag. Gerald Lirk

August 2007 - December 2007 Internship at the Stockholm Bioinformatics Center

Supervisor:

Dr. Erik Sonnhammer

Publications

Journal Articles

(i) T. Koestler, A. von Haeseler, and I. Ebersberger (2012) Modeling sequence evolution under domain constraints. *Molecular Biology and Evolution*, in press.

(ii) T. Koestler and I. Ebersberger (2011) Zygomycetes, microsporidia, and the evolutionary ancestry of sex determination. *Genome Biology and Evolution*, (3):186-194.

(iii) T. Koestler, A. von Haeseler, and I. Ebersberger (2010) FACT: Functional annotation transfer between proteins with similar feature architecture. *BMC Bioinformatics*, **11**(1):417.

(iv) G. Ostlund, T. Schmitt, K. Forslund, T. Köstler, D.N. Messina, S. Roopra, O. Frings, E.L. Sonnhammer (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, 38(Database issue):D196-203

(v) C. Dessimoz, T. Gabaldon, D. S. Roos, E. Sonnhammer, J. Herrero, and the Quest for Orthologs Consortium* (2012) Toward Community Standards in the Quest for Orthologs *Bioinformatics*, in press.

*Members of the Quest for Orthologs Consortium: A. Altenhoff, R. Apweiler, J. Blake, B. Boeckmann, A. Bridge, E. Bruford, M. Cherry, M. Conte, D. Dannie, R. Datta, C. Dessimoz, J.-B. Domelevo Entfellner, I. Ebersberger, T. Gabaldón, M. Galperin, J. Herrero, J. Joseph, T. Koestler, E. Kriventseva, O. Lecompte, J. Leunissen, S. Lewis, B. Linard, M. S. Livstone, H.-C. Lu, M. Martin, R. Mazumder, V. Miele, M. Muffato, G. Perriére, M. Punta, D. Roos, M. Rouard, T. Schmitt, F. Schreiber, A. Silva, K. Sjlander, N. Skunca, E. Sonnhammer, E. Stanley, R. Szklarczyk, P. Thomas, I. Uchiyama, M. Van Bel, K. Vandepoele, A. J. Vilella, A. Yates, and E. Zdobnov.

Bibliography

Abascal F, Posada D, Zardoya R. 2007. MtArt: a new model of amino acid replacement for arthropoda. Molecular Biology and Evolution. 24:1–5.

Adachi J, Hasegawa M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. Journal of Molecular Evolution. 42:459–468.

Adachi J, Waddell PJ, Martin W, Hasegawa M. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. Journal of Molecular Evolution. 50:348–358.

Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. 2010. OMA 2011: orthology inference among 1000 complete genomes. Nucleic Acids Research. 39:D289–D294.

Altschul SF, Madden TL, Schäffer 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 Research. 25:3389–3402.

Ashburner M, Ball CA, Blake JA, et al. (20 co-authors). 2000. Gene ontology: tool for the unification of biology. Nature Genetics. 25:25–29.

Bartlett GJ, Borkakoti N, Thornton JM. 2003. Catalysing new reactions during evolution: economy of residues and mechanism. Journal of Molecular Biology. 331:829–860.

Basu MK, Carmel L, Rogozin IB, Koonin EV. 2008. Evolution of protein domain promiscuity in eukaryotes. Genome Research. 18:449–461.

Baurain D, Brinkmann H, Petersen J, Rodríguez-Ezpeleta N, Stechmann A, Demoulin V, Roger AJ, Burger G, Lang BF, Philippe H. 2010. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. Molecular Biology and Evolution. 27:1698–1709.

Benner SA, Cohen MA, Gonnet GH. 1993. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. Journal of Molecular Biology. 229:1065–1082.

Berglund A, Sjölund E, Ostlund G, Sonnhammer ELL. 2008. InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Research. 36:D263–266.

Berriman M, Ghedin E, Hertz-Fowler C, et al. (102 co-authors). 2005. The genome of the african trypanosome trypanosoma brucei. Science. 309:416–422.

Boulesteix A. 2010. Over-optimism in bioinformatics research. Bioinformatics (Oxford, England). 26:437–439.

Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Systematic Biology. 54:743–757.

Buljan M, Bateman A. 2009. The evolution of protein domain families. Biochemical Society Transactions. 37:751–755.

Butler G, Kenny C, Fagan A, Kurischko C, Gaillardin C, Wolfe KH. 2004. Evolution of the MAT locus and its ho endonuclease in yeast species. Proceedings of the National Academy of Sciences of the United States of America. 101:1632–1637.

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. 2009. AmiGO: online access to ontology and annotation data. Bioinformatics (Oxford, England). 25:288–289.

Cavalier-Smith T. 1986. The kingdoms of organisms. Nature. 324:416–417.

Chang MSS, Benner SA. 2004. Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. Journal of Molecular Biology. 341:617–631.

Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PloS One. 2:e383.

Choi JH, Lou W, Vancura A. 1998. A novel membrane-bound glutathione s-transferase functions in the stationary phase of the yeast saccharomyces cerevisiae. The Journal of Biological Chemistry. 273:29915–29922.

Corradi N, Keeling PJ. 2009. Microsporidia: a journey through radical taxonomical revisions. Fungal Biology Reviews. 23:1–8.

Darwin C. 1959. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London.

Dayhoff M, Schwartz R, Orcutt B. 1978. A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure. 5:345–352.

Dessimoz C, Gabaldón T, Roos DS, Sonnhammer E, Herrero J, the Quest for Orthologs Consortium. 2012. Toward community standards in the quest for orthologs. Bioinformatics. in press.

Diao A, Rahman D, Pappin DJC, Lucocq J, Lowe M. 2003. The coiled-coil membrane protein golgin-84 is a novel rab effector required for golgi ribbon formation. The Journal of Cell Biology. 160:201–212.

Dietrich FS, Voegeli S, Brachat S, et al. (14 co-authors). 2004. The ashbya gossypii genome as a tool for mapping the ancient saccharomyces cerevisiae genome. Science (New York, N.Y.). 304:304–307.

Dyer P. 2008a. Evolutionary biology: Microsporidia sex — a missing link to fungi. Current Biology. 18:R1012–R1014.

Dyer PS. 2008b. Evolutionary biology: Genomic clues to original sex in fungi. Current Biology. 18:R207–R209.

Ebersberger I, Strauss S, von Haeseler A. 2009. HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evolutionary Biology. 9:157.

Eddy SR. 1998. Profile hidden markov models. Bioinformatics (Oxford, England). 14:755–763.

Embley TM, Hirt RP. 1998. Early branching eukaryotes? Current Opinion in Genetics & Development. 8:624–629.

Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen M, Pieper U, Sali A. 2006. Comparative protein structure modeling using modeller. Current Protocols in Bioinformatics. Chapter 5.

Felsenstein J. 2004. Inferring phylogenies. Sinauer Associates.

Finn RD, Mistry J, Schuster-Böckler B, et al. (13 co-authors). 2006. Pfam: clans, web tools and services. Nucleic Acids Research. 34:D247–251.

Finn RD, Mistry J, Tate J, et al. (14 co-authors). 2010. The pfam protein families database. Nucleic Acids Research. 38:D211–222.

Finn RD, Tate J, Mistry J, et al. (11 co-authors). 2008. The pfam protein families database. Nucleic Acids Research. 36:D281–288.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. Systematic Zoology. 19:99–113.

Fitch WM. 1971. Toward defining the course of evolution: Minimum change for a specific tree topology. Systematic Zoology. 20:406–416.

Fletcher W, Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. Molecular Biology and Evolution. 26:1879 –1888.

Forslund K, Henricson A, Hollich V, Sonnhammer ELL. 2008. Domain tree-based analysis of protein architecture evolution. Molecular Biology and Evolution. 25:254–264.

Forslund K, Pekkari I, Sonnhammer EL. 2011. Domain architecture conservation in orthologs. BMC Bioinformatics. 12:326.

Forslund K, Sonnhammer ELL. 2008. Predicting protein function from domain content. Bioinformatics. 24:1681–1687.

Galperin MY, Walker DR, Koonin EV. 1998. Analogous enzymes: independent inventions in enzyme evolution. Genome Research. 8:779–790.

Gilks WR, Audit B, de Angelis D, Tsoka S, Ouzounis CA. 2005. Percolation of annotation errors through hierarchically structured protein sequence databases. Mathematical Biosciences. 193:223–234.

Gill EE, Fast NM. 2006. Assessing the microsporidia-fungi relationship: Combined phylogenetic analysis of eight genes. Gene. 375:103–109.

Gillespie DT. 1977. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry. 81:2340–2361.

Gough J. 2005. Convergent evolution of domain architectures (is rare). Bioinformatics (Oxford, England). 21:1464–1471.

Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. Journal of Molecular Biology. 313:903–919.

Grahnen JA, Kubelka J, Liberles DA. 2011. Fast side chain replacement in proteins using a coarse-grained approach for evaluating the effects of mutation during evolution. Journal of Molecular Evolution. 73:23–33.

Grahnen JA, Nandakumar P, Kubelka J, Liberles DA. 2011. Biophysical and structural considerations for protein sequence evolution. BMC Evolutionary Biology. 11:361.

Haber JE. 1992. Mating-type gene switching in saccharomyces cerevisiae. Trends in Genetics: TIG. 8:446–452.

Haimel M, Pröll K, Rebhan M. 2009. ProteinArchitect: protein evolution above the sequence level. PloS One. 4:e6176.

Hammesfahr B, Odronitz F, Hellkamp M, Kollmar M. 2011. diArk 2.0 provides detailed analyses of the ever increasing eukaryotic genome sequencing data. BMC Research Notes. 4:338.

Haqq CM, King CY, Donahoe PK, Weiss MA. 1993. SRY recognizes conserved DNA sites in sex-specific promoters. Proceedings of the National Academy of Sciences of the United States of America. 90:1097–1101.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America. 89:10915–10919.

Hennig W, Davis D, Zangerl R. 1966. Phylogenetic systematics. University of Illinois Press.

Henricson A, Forslund K, Sonnhammer ELL. 2010. Orthology confers intron position conservation. BMC Genomics. 11:412.

Hollich V, Sonnhammer ELL. 2007. PfamAlyzer: domain-centric homology search. Bioinformatics (Oxford, England). 23:3382–3383.

Hulo N. 2006. The PROSITE database. Nucleic Acids Research. 34:D227–D230.

Idnurm A, Walton FJ, Floyd A, Heitman J. 2008. Identification of the sex genes in an early diverged fungus. Nature. 451:193–196.

James TY, Kauff F, Schoch CL, et al. (70 co-authors). 2006. Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature. 443:818–822.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS. 8:275–282.

Kanehisa M, Araki M, Goto S, et al. (11 co-authors). 2008. KEGG for linking genomes to life and the environment. Nucleic Acids Research. 36:D480–484.

Karlin S, Taylor H. 1975. A first course in stochastic processes. In: A First Course in Stochastic Processes, 2nd ed. New York: Academic Press.

Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA. 2009. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. Genome Research. 19:1404–1418.

Katinka MD, Duprat S, Cornillot E, et al. (17 co-authors). 2001. Genome sequence and gene compaction of the eukaryote parasite encephalitozoon cuniculi. Nature. 414:450–453.

Katoh K, Kuma Ki, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research. 33:511–518.

Keeling P. 2009. Five questions about microsporidia. PLoS Pathog. 5:e1000489.

Keeling PJ. 2003. Congruent evidence from alpha-tubulin and beta-tubulin gene phylogenies for a zygomycete origin of microsporidia. Fungal Genetics and Biology: FG & B. 38:298–309.

Keeling PJ, Luker MA, Palmer JD. 2000. Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. Molecular Biology and Evolution. 17:23–31.

Kent WJ. 2002. BLAT–the BLAST-like alignment tool. Genome Research. 12:656–664.

Kim J, Sinha S. 2010. Towards realistic benchmarks for multiple alignments of non-coding sequences. BMC Bioinformatics. 11:54.

Kimura M. 1979. The neutral theory of molecular evolution. Scientific American. 241:98–100, 102, 108 passim.

Koestler T, von Haeseler A, Ebersberger I. 2010. FACT: functional annotation transfer between proteins with similar feature architectures. BMC Bioinformatics. 11:417.

Kosiol C, Goldman N. 2005. Different versions of the dayhoff rate matrix. Molecular Biology and Evolution. 22:193–199.

Kotz S, Johnson NL, Balakrishnan N. 2000. Continuous Multivariate Distributions: Models and applications. John Wiley and Sons.

Kumar S. 2005. Molecular clocks: four decades of evolution. Nature Reviews Genetics. 6:654–662.

Laity JH, Lee BM, Wright PE. 2001. Zinc finger proteins: new insights into structural and functional diversity. Current Opinion in Structural Biology. 11:39–46.

Lakner C, Holder MT, Goldman N, Naylor GJP. 2011. What's in a likelihood? simple models of protein evolution and the contribution of structurally viable reconstructions to the likelihood. Systematic Biology. 60:161–174.

Lartillot N, Philippe H. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology and Evolution. 21:1095–1109.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Molecular Biology and Evolution. 25:1307 –1320.

Lee B, Lee D. 2009. Protein comparison at the domain architecture level. BMC Bioinformatics. 10:S5.

Lee SC, Corradi N, Doan S, Dietrich FS, Keeling PJ, Heitman J. 2010a. Evolution of the sex-Related locus and genomic features shared in microsporidia and fungi. PLoS ONE. 5:e10539.

Lee SC, Corradi N, III EJB, Torres-Martinez S, Dietrich FS, Keeling PJ, Heitman J. 2008. Microsporidia evolved from ancestral sexual fungi. Current Biology. 18:1675–1679.

Lee SC, Ni M, Li W, Shertz C, Heitman J. 2010b. The evolution of sex: a perspective from the fungal kingdom. Microbiology and Molecular Biology Reviews: MMBR. 74:298–340.

Lengeler KB, Fox DS, Fraser JA, Allen A, Forrester K, Dietrich FS, Heitman J. 2002. Mating-Type locus of cryptococcus neoformans: a step in the evolution of sex chromosomes. Eukaryotic Cell. 1:704–718.

Leslie CS, Eskin E, Cohen A, Weston J, Noble WS. 2004. Mismatch string kernels for discriminative protein classification. Bioinformatics (Oxford, England). 20:467–476.

Letunic I, Doerks T, Bork P. 2009. SMART 6: recent updates and new developments. Nucleic Acids Research. 37:D229–232.

Lin K, Zhu L, Zhang D. 2006. An initial strategy for comparing proteins at the domain architecture level. Bioinformatics (Oxford, England). 22:2081–2086.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science (New York, N.Y.). 285:751–753.

Margoliash E. 1963. PRIMARY STRUCTURE AND EVOLUTION OF CYTOCHROME c. Proceedings of the National Academy of Sciences of the United States of America. 50:672–679.

Müller T, Vingron M. 2000. Modeling amino acid replacement. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology. 7:761–776.

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

Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, Pond SLK. 2007. HIV-specific probabilistic models of protein evolution. PloS One. 2:e503.

Notredame C. 2007. Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol. 3:e123.

Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, Frings O, Sonnhammer ELL. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Research. 38:D196–203.

Palczewski K, Kumasaka T, Hori T, et al. (12 co-authors). 2000. Crystal structure of rhodopsin: A g Protein-Coupled receptor. Science. 289:739–745.

Pang A, Smith AD, Nuin PAS, Tillier ERM. 2005. SIMPROT: using an empirically determined indel distribution in simulations of protein evolution. BMC Bioinformatics. 6:236.

Parisi G, Echave J. 2001. Structural constraints and emergence of sequence patterns in protein evolution. Molecular Biology and Evolution. 18:750–756.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proceedings of the National Academy of Sciences of the United States of America. 96:4285–4288.

Philippe H. 2000. Early–branching or fast–evolving eukaryotes? an answer based on slowly evolving positions. Proceedings of the Royal Society of London. Series B: Biological Sciences. 267:1213–1221.

Poy F, Yaffe MB, Sayos J, Saxena K, Morra M, Sumegi J, Cantley LC, Terhorst C, Eck MJ. 1999. Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition. Molecular Cell. 4:555–561.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS One. 5:e9490.

Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, Hamodrakas S, Sander C, Ouzounis CA. 2000. CAST: an iterative algorithm for the complexity analysis of sequence tracts. complexity analysis of sequence tracts. Bioinformatics (Oxford, England). 16:915–922.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research. 35:D61–65.

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J. 2001. The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Research. 29:159–164.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. CABIOS. 13:235–238.

Ramirez IB, de Graffenried CL, Ebersberger I, Yelinek J, He CY, Price A, Warren G. 2008. TbG63, a golgin involved in golgi architecture in trypanosoma brucei. Journal of Cell Science. 121:1538–1546.

Rastogi S, Reuter N, Liberles DA. 2006. Evaluation of models for the evolution of protein sequences and functions under structural constraint. Biophysical Chemistry. 124:134–144.

Remm M, Storm CE, Sonnhammer EL. 2001. Automatic clustering of orthologs and inparalogs from pairwise species comparisons. Journal of Molecular Biology. 314:1041–1052.

Rodriguez-Navarro S, Fischer T, Luo M, Antúnez O, Brettschneider S, Lechner J, Pérez-Ortn JE, Reed R, Hurt E. 2004. Sus1, a functional component of the SAGA histone acetylase complex and the nuclear Pore-Associated mRNA export machinery. Cell. 116:75–86.

Rost B. 2002. Enzyme function less conserved than anticipated. Journal of Molecular Biology. 318:595–608.

Satoh A, Wang Y, Malsam J, Beard MB, Warren G. 2003. Golgin-84 is a rab1 binding partner involved in golgi structure. Traffic (Copenhagen, Denmark). 4:153–161.

Schmidt HA, von Haeseler A. 2007. Maximum-likelihood analysis using TREE-PUZZLE. Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]. Chapter 6:Unit 6.6.

Schuster-Bockler B, Schultz J, Rahmann S. 2004. HMM logos for visualization of protein families. BMC Bioinformatics. 5:7.

Schwartz JH, Maresca B. 2006. Do molecular clocks run at all? a critique of molecular systematics. Biological Theory. 1:357–371.

Shah AR, Oehmen CS, Webb-Robertson B. 2008. SVM-HUSTLE–an iterative semi-supervised machine learning approach for pairwise protein remote homology detection. Bioinformatics (Oxford, England). 24:783–790.

Smith T, Waterman M. 1981. Identification of common molecular subsequences. Journal of Molecular Biology. 147:195–197.

Söding J. 2005. Protein homology detection by HMM–HMM comparison. Bioinformatics (Oxford, England). 21:951 –960.

Song N, Sedgewick RD, Durand D. 2007. Domain architecture comparison for multidomain homology identification. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology. 14:496–516.

Sonnhammer EL, Wootton JC. 2001. Integrated graphical analysis of protein sequence features predicted from sequence composition. Proteins. 45:262–273.

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

Stoye J, Evers D, Meyer F. 1998. Rose: generating sequence families. Bioinformatics. 14:157–163.

Strope CL, Abel K, Scott SD, Moriyama EN. 2009. Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen version 2.0. Molecular Biology and Evolution. 26:2581 –2593.

Strope CL, Scott SD, Moriyama EN. 2007. indel-Seq-Gen: a new protein family simulator incorporating domains, motifs, and indels. Molecular Biology and Evolution. 24:640 –649.

Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. Nature Reviews. Genetics. 3:137–144.

The UniProt Consortium. 2010. Ongoing and future developments at the universal protein resource. Nucleic Acids Research. 39:D214–D219.

Thomarat F, Vivarès CP, Gouy M. 2004. Phylogenetic analysis of the complete genome sequence of encephalitozoon cuniculi supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes. Journal of Molecular Evolution. 59:780–791.

Thomas JO, Travers AA. 2001. HMG1 and 2, and related 'architectural' DNA-binding proteins. Trends in Biochemical Sciences. 26:167–174.

Thorne JL, Kishino H, Felsenstein J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. Journal of Molecular Evolution. 33:114–124.

Tian W, Skolnick J. 2003. How well is enzyme function conserved as a function of pairwise sequence identity? Journal of Molecular Biology. 333:863–882.

Tusnády GE, Simon I. 2001. The HMMTOP transmembrane topology prediction server. Bioinformatics. 17:849–850.

Velankar S, Alhroub Y, Alili A, et al. (33 co-authors). 2011. PDBe: protein data bank in europe. Nucleic Acids Research. 39:D402–410.

Vinayagam A, König R, Moormann J, Schubert F, Eils R, Glatting K, Suhai S. 2004. Applying support vector machines for gene ontology based gene function prediction. BMC Bioinformatics. 5:116.

Viterbi A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. Information Theory, IEEE Transactions on. 13:260–269.

von Haeseler A, Schöniger M. 1998. Evolution of DNA or amino acid sequences with dependent sites. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology. 5:149–163.

Vossbrinck CR, Maddox JV, Friedman S, Debrunner-Vossbrinck BA, Woese CR. 1987. Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. Nature. 326:411–414.

Weiner r January, Beaussart F, Bornberg-Bauer E. 2006. Domain deletions and substitutions in the modular protein evolution. The FEBS Journal. 273:2037–2047.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution. 18:691–699.

Wilgenbusch JC, Swofford D. 2003. Inferring evolutionary trees with PAUP*. Current Protocols in Bioinformatics. Chapter 6:Unit 6.4.

Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Molecular Biology and Evolution. 15:1600–1611.

Zuckerkandl E, Pauling L, Kasha M, Pullman B. 1962. Molecular disease, evolution, and genetic heterogeneity. In: Horizons in Biochemistry, Academic Press, pp. 189–225.

Appendix A

Supplementary Tables and Figures to Chapter 2

p-value	max	rank 1	unique rank 1
$< 10^{-0}$	9570	8014	7091
$< 10^{-1}$	9536	8007	7084
$< 10^{-2}$	8981	7740	6849
$< 10^{-3}$	7961	6979	6185
$< 10^{-4}$	5703	5154	4574
$< 10^{-5}$	4018	3712	3290
$< 10^{-6}$	3045	2864	2513
$< 10^{-7}$	2387	2278	1992
$< 10^{-8}$	1934	1859	1620
$< 10^{-9}$	1558	1508	1317
$< 10^{-10}$	1325	1286	1114
$< 10^{-11}$	1073	1047	892
$< 10^{-12}$	909	891	760
$< 10^{-13}$	758	743	635
$< 10^{-14}$	650	636	541
$< 10^{-15}$	558	548	465

Table A.1: Impact of p-value thresholds on the coverage of FACT (*FACT* score). 'max' denotes the number of searches resulting in a highest scoring protein with a p-value below the given threshold ('p-value'). 'rank 1' denotes the number of searches were the highest scoring protein has the same EC annotation as the query. 'unique rank 1' denotes the number of searches were the highest scoring protein has the same EC annotation as the query and is uniquely highest scoring.

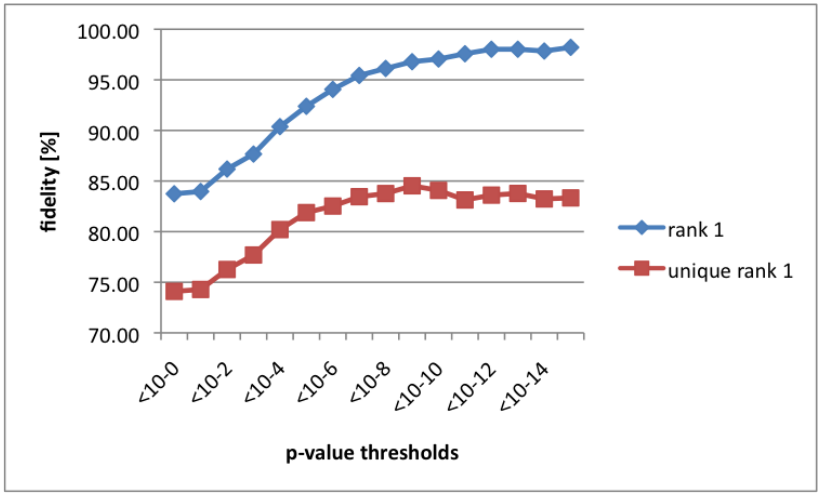


Figure A.1: Impact of p-value thresholds on the fidelity of FACT (*FACT* score). 'fidelity' denotes the fraction of FACT searches where a protein with the same EC number as the query is top scoring (blue graph). The red line represents the percentage of correctly and uniquely top ranked proteins with FACT (unique rank 1). The coverage of FACT for the p-value thresholds are given in table A.1.

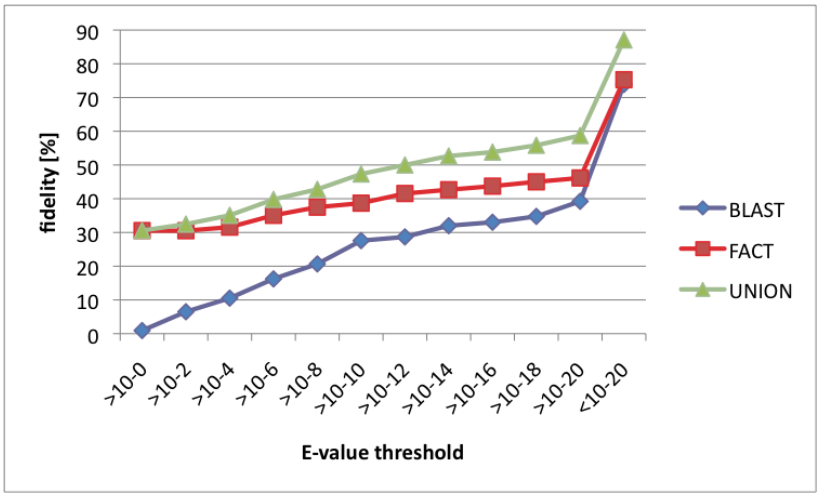


Figure A.2: Cumulative fidelity along E-value thresholds for FACT (*FACT* score), Blast, and the union of FACT and Blast.

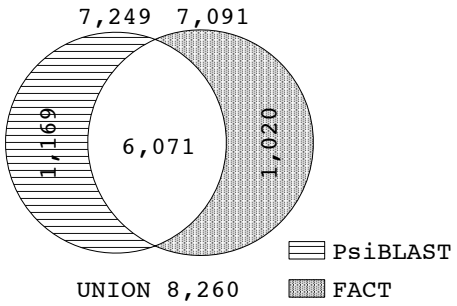


Figure A.3: Venn diagram contrasting the performance of FACT (*FACT* score) and PsiBlast. Given are the the numbers of uniquely top ranking proteins having the same EC number as the query.

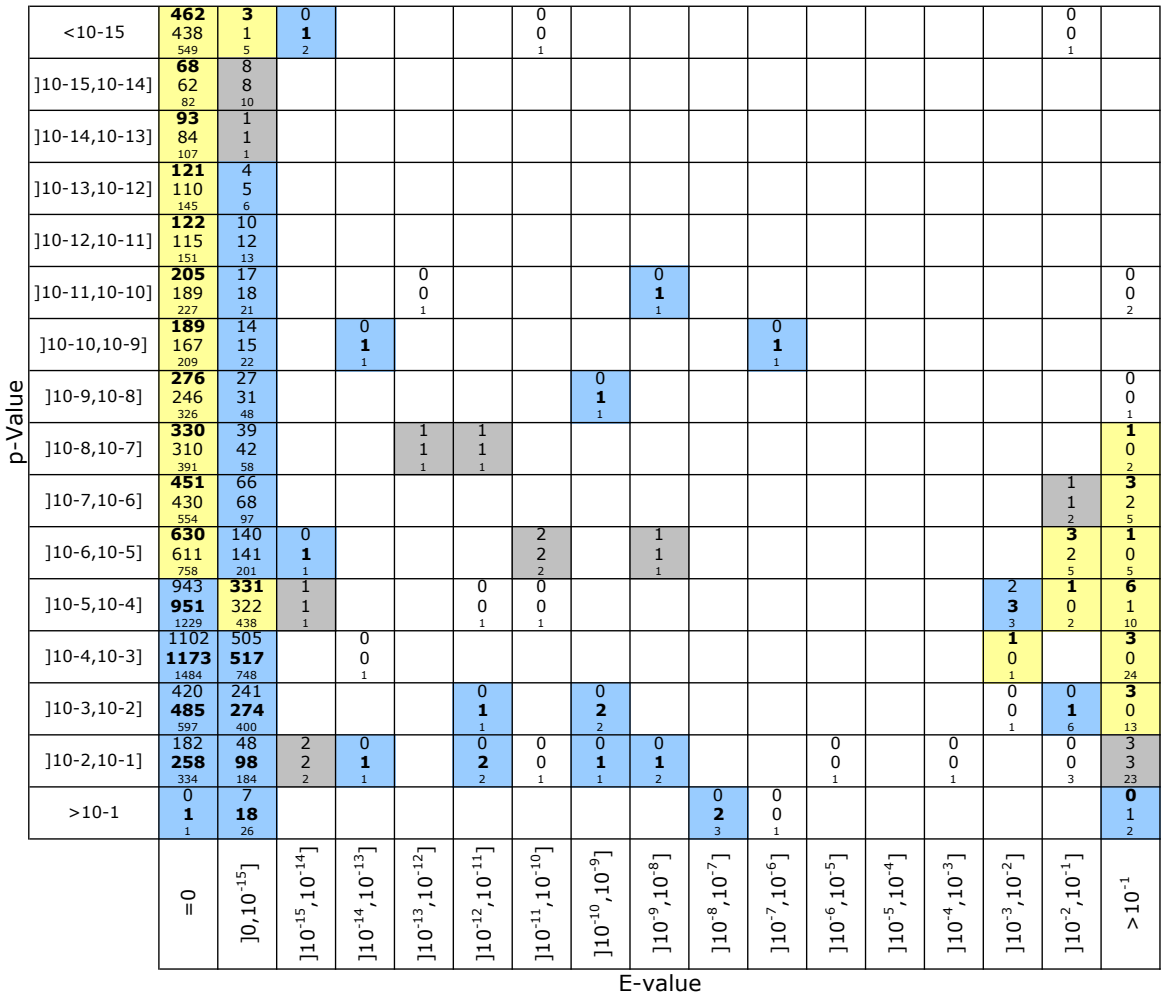


Figure A.4: Contrast of PsiBlast and FACT (*FACT* score) for different E-value/p-value combinations. The matrix bins the 9,570 proteins according to the E-value and the p-value of the best hit when used as query for PsiBlast and FACT, respectively. The total number of proteins for a E-value/p-value combination is given by the bottom number in the corresponding cell. The two further numbers in a cell give the number of searches FACT (top) and PsiBlast (middle) had a functional equivalent as top scoring protein. The number for the better performing tool is given in bold face. Yellow cells show E-value/p-value combinations where FACT identified more functional equivalents than PsiBlast, whereas the blue cells indicate a higher fidelity of PsiBlast. Grey cells mark ties.

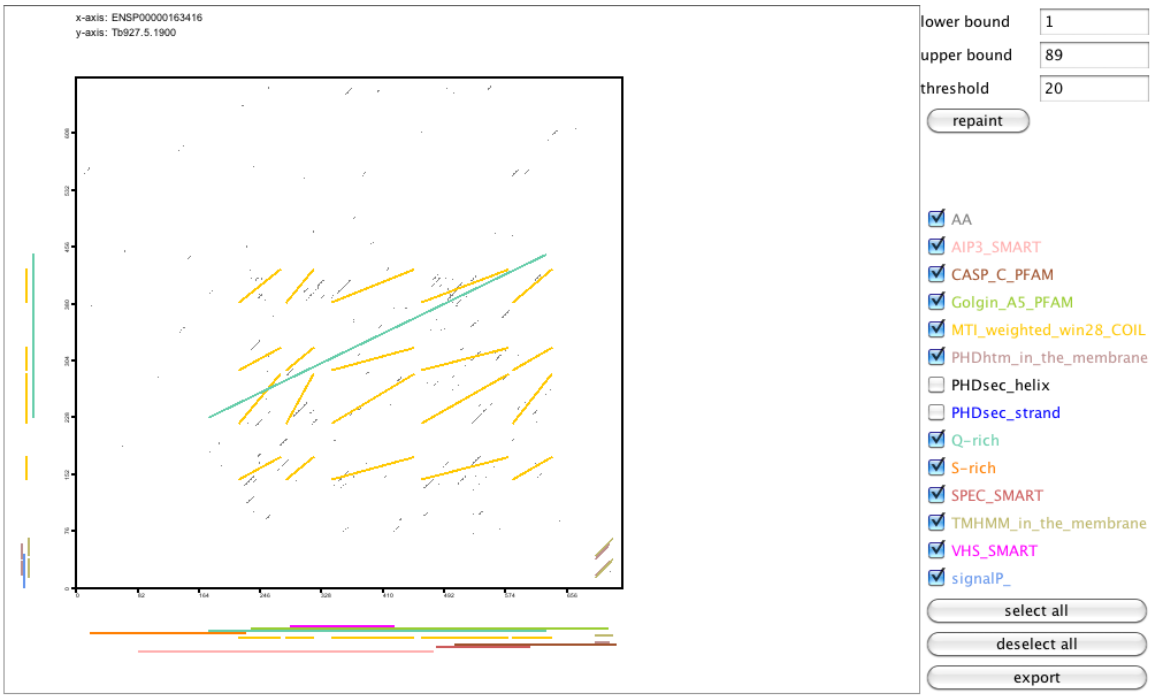


Figure A.5: FDP of the human GolgA5 and the highest scoring hit (MLS) in *T. brucei*: Tb927.5.1900

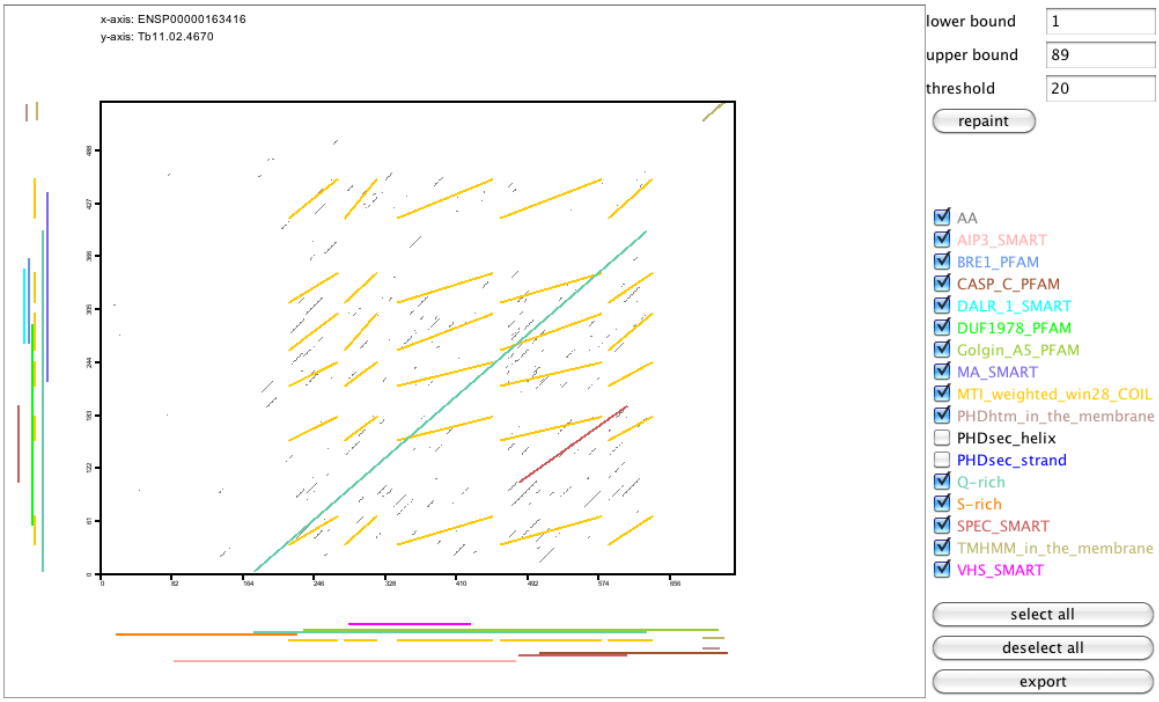


Figure A.6: FDP of the human GolgA5 and the highest scoring hit (MS_{uni}) in *T. brucei*: Tb11.02.4670

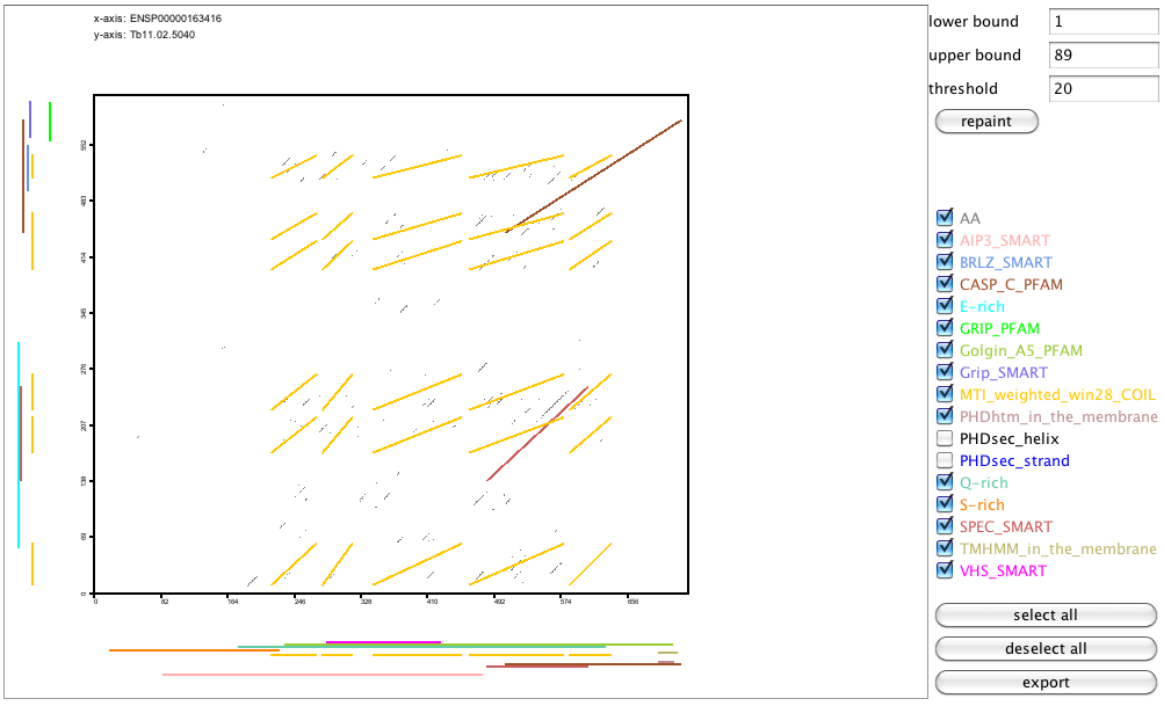


Figure A.7: FDP of the human GolgA5 and the highest scoring hit ($MS_{st}/FACT$ score) in *T. brucei*: Tb11.02.5040

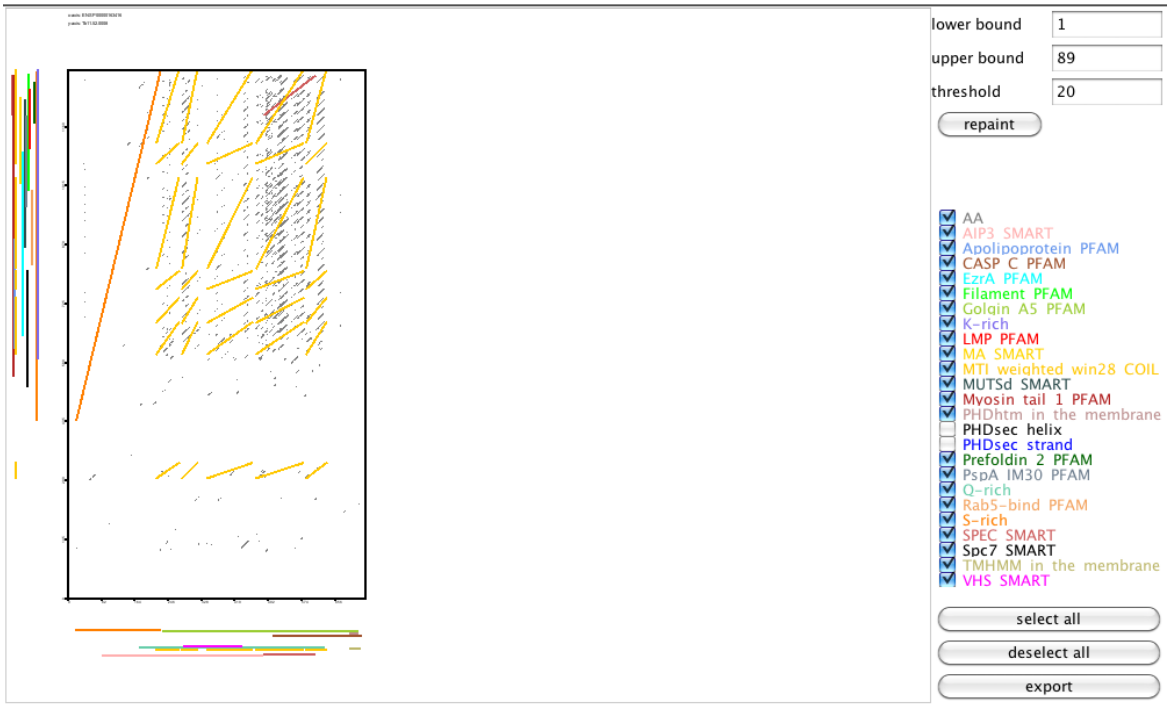


Figure A.8: FDP of the human GolgA5 and the best Blast hit in *T. brucei*: Tb11.52.0008

Appendix B

Supplementary Tables and Figures to Chapter 3

$\Lambda_I = \Lambda_D = 0.1$				
T	unconstrained	$\alpha = 1$	$\alpha = 0.5$	constrained
0	45738 (1.000)	45738 (1.000)	45738 (1.000)	45738 (1.000)
0.1	38273 (0.837)	38542 (0.843)	38600 (0.844)	42267 (0.924)
0.5	16223 (0.355)	18702 (0.409)	20361 (0.445)	39149 (0.856)
1	1670 (0.037)	3728 (0.082)	5496 (0.120)	34960 (0.764)
1.5	31 (0.001)	378 (0.008)	1034 (0.023)	30837 (0.674)
$\Lambda_I = \Lambda_D = 0.05$				
T	unconstrained	$\alpha = 1$	$\alpha = 0.5$	constrained
0	45738 (1.000)	45738 (1.000)	45738 (1.000)	45738 (1.000)
0.1	39574 (0.865)	39565 (0.865)	39727 (0.869)	42763 (0.935)
0.5	22457 (0.491)	24665 (0.539)	26462 (0.579)	40817 (0.892)
1	5700 (0.125)	10151 (0.222)	13185 (0.288)	38504 (0.842)
1.5	484 (0.011)	2942 (0.064)	5572 (0.122)	36230 (0.792)
$\Lambda_I = \Lambda_D = 0$				
T	unconstrained	$\alpha = 1$	$\alpha = 0.5$	constrained
0	45738 (1.000)	45738 (1.000)	45738 (1.000)	45738 (1.000)
0.1	40772 (0.891)	40811 (0.892)	41061 (0.898)	43288 (0.946)
0.5	29602 (0.647)	31834 (0.696)	33326 (0.729)	42646 (0.932)
1	15742 (0.344)	22017 (0.481)	26064 (0.570)	42150 (0.922)
1.5	6727 (0.147)	15134 (0.331)	20510 (0.448)	41943 (0.917)

Table B.1: Number (fraction) of preserved Pfam domains in the course of simulated evolution with and without imposing domain constraints. All human proteins were taken as root sequences and evolved up to 0.1, 0.5, 1.0 and 1.5 substitutions per site. In the unconstrained case we additionally modeled substitution rate heterogeneity using two different α-values for the gamma distribution. The simulations were performed with three different insertion and deletion rates (0.1; 0.05; 0).

$\Lambda_I = \Lambda_D = 0.1$				
T	unconstrained	$\alpha = 1$	$\alpha = 0.5$	constrained
0	32289 (1.000)	32289 (1.000)	32289 (1.000)	32289 (1.000)
0.1	25721 (0.797)	25836 (0.800)	25883 (0.802)	29536 (0.915)
0.5	8695 (0.269)	10274 (0.318)	11279 (0.349)	26095 (0.808)
1	746 (0.023)	1755 (0.054)	2694 (0.083)	22547 (0.698)
1.5	33 (0.001)	170 (0.005)	482 (0.014)	19484 (0.603)
$\Lambda_I = \Lambda_D = 0.05$				
T	unconstrained	$\alpha = 1$	$\alpha = 0.5$	constrained
0	32289 (1.000)	32289 (1.000)	32289 (1.000)	32289 (1.000)
0.1	26726 (0.828)	26874 (0.832)	26949 (0.835)	29965 (0.928)
0.5	12734 (0.394)	14491 (0.449)	15603 (0.483)	28019 (0.868)
1	2566 (0.079)	4999 (0.155)	6610 (0.205)	25767 (0.798)
1.5	259 (0.008)	1307 (0.040)	2567 (0.080)	23846 (0.739)
$\Lambda_I = \Lambda_D = 0$				
T	unconstrained	$\alpha = 1$	$\alpha = 0.5$	constrained
0	32289 (1.000)	32289 (1.000)	32289 (1.000)	32289 (1.000)
0.1	27826 (0.862)	28036 (0.868)	28201 (0.873)	30349 (0.940)
0.5	17818 (0.552)	19619 (0.608)	20917 (0.648)	29905 (0.926)
1	7730 (0.239)	11995 (0.371)	15086 (0.467)	29572 (0.916)
1.5	2754 (0.085)	7265 (0.225)	10739 (0.333)	29467 (0.913)

Table B.2: Number (fraction) of preserved SMART domains in the course of simulated evolution with and without imposing domain constraints. All human proteins were taken as root sequences and evolved up to 0.1, 0.5, 1.0 and 1.5 substitutions per site. In the unconstrained case we additionally modeled substitution rate heterogeneity using two different α-values for the gamma distribution. The simulations were performed with three different insertion and deletion rates (0.1; 0.05; 0).

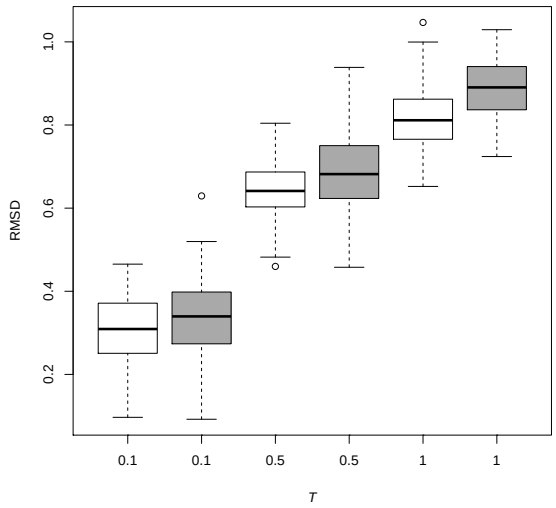
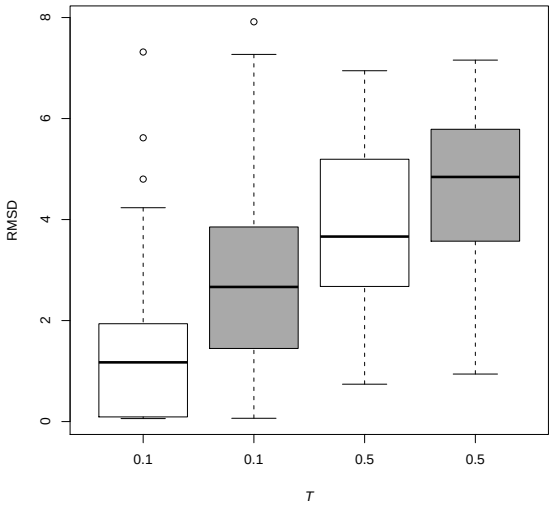
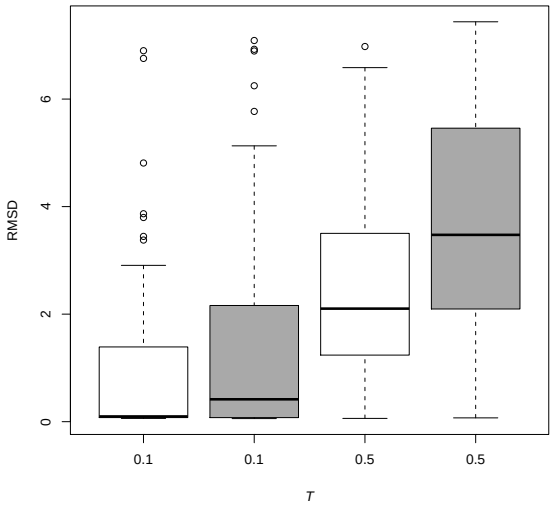


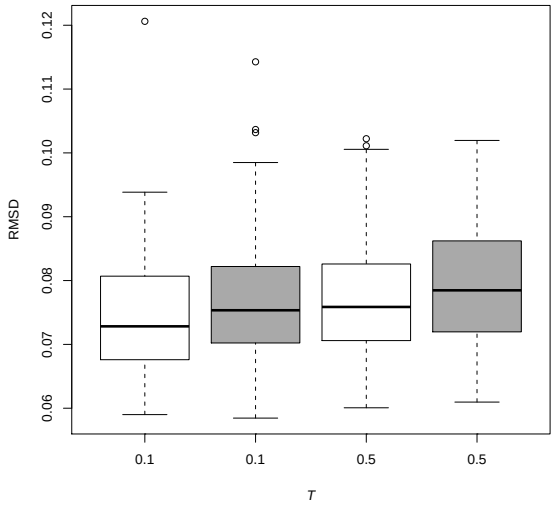
Figure B.1: RMSD of the side chains between sequences simulated with domain constraints and the native structure (white boxes), and between sequences simulated without domain constraints and the native structure (gray boxes). The mean RMSD for all sequences simulated under domain constraints are significantly smaller than for sequences simulated without domain constraints (t-test). The p-values are as follows. $T = 0.1 : p = 0.022$; $T = 0.5 : p = 2e^{-5}$; $T = 1 : p = 3e^{-12}$.



(a) $\Lambda_I = \Lambda_D = 0.1$; $p_{0.1} = 5e^{-9}$; $p_{0.5} = 9e^{-5}$



(b) $\Lambda_I = \Lambda_D = 0.05$; $p_{0.1} = 0.029$; $p_{0.5} = 6e^{-6}$



(c) $\Lambda_I = \Lambda_D = 0$; $p_{0.1} = 0.024$; $p_{0.5} = 0.058$

Figure B.2: RMSD of the backbone between sequences simulated with domain constraints and the native structure (white boxes), and between sequences simulated without domain constraints and the native structure (gray boxes). $p_{0.1}$ and $p_{0.5}$ below each figure show the p-values for the t-test between the resulting mean RMSD from simulations with domain constraints and simulations without domain constraints.

	REvolver
tm regions	6.13 ± 1.27
Bit score (7tm_1)	20.71
Bit score (custom pHMM)	170.30
Top n BlastP hits	
25	117.4
100	109.5
250	101.8

Table B.3: Results for the simulated evolution of a consensus GPCR sequence. The consensus sequence was annotated with a custom pHMM built from the alignment of 29 GPCR proteins. 1,000 independent simulations were performed. 'tm regions' denotes the average number of transmembrane regions in the simulated sequences. 75% of the simulated sequences show a significant similarity to the Pfam domain 7tm_1 and all simulated sequences are significantly similar to the custom pHMM. For 99% of the simulated sequences all 250 BlastP hits each were members of the GPCR family. Only in 302 cases (1%) a non-GPCR protein was among the top 250 BlastP hits. The mean bit scores of the first *25*, *100*, and *250* BlastP hits are shown below 'Top n BlastP hits'.

Appendix C

Supplementary Figures and Tables to Chapter 5

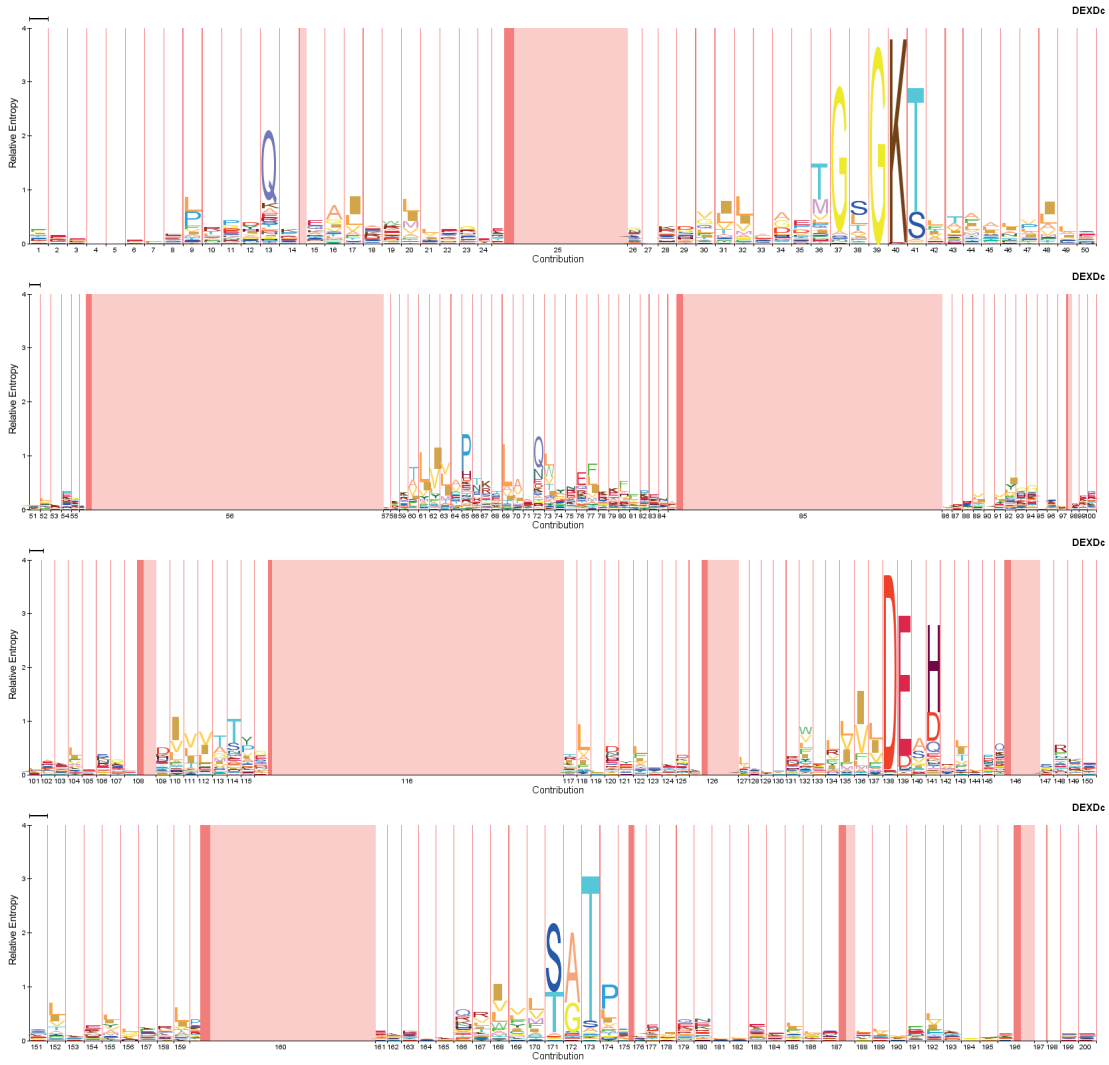


Figure C.1: Logo of the DEXDc pHMM.

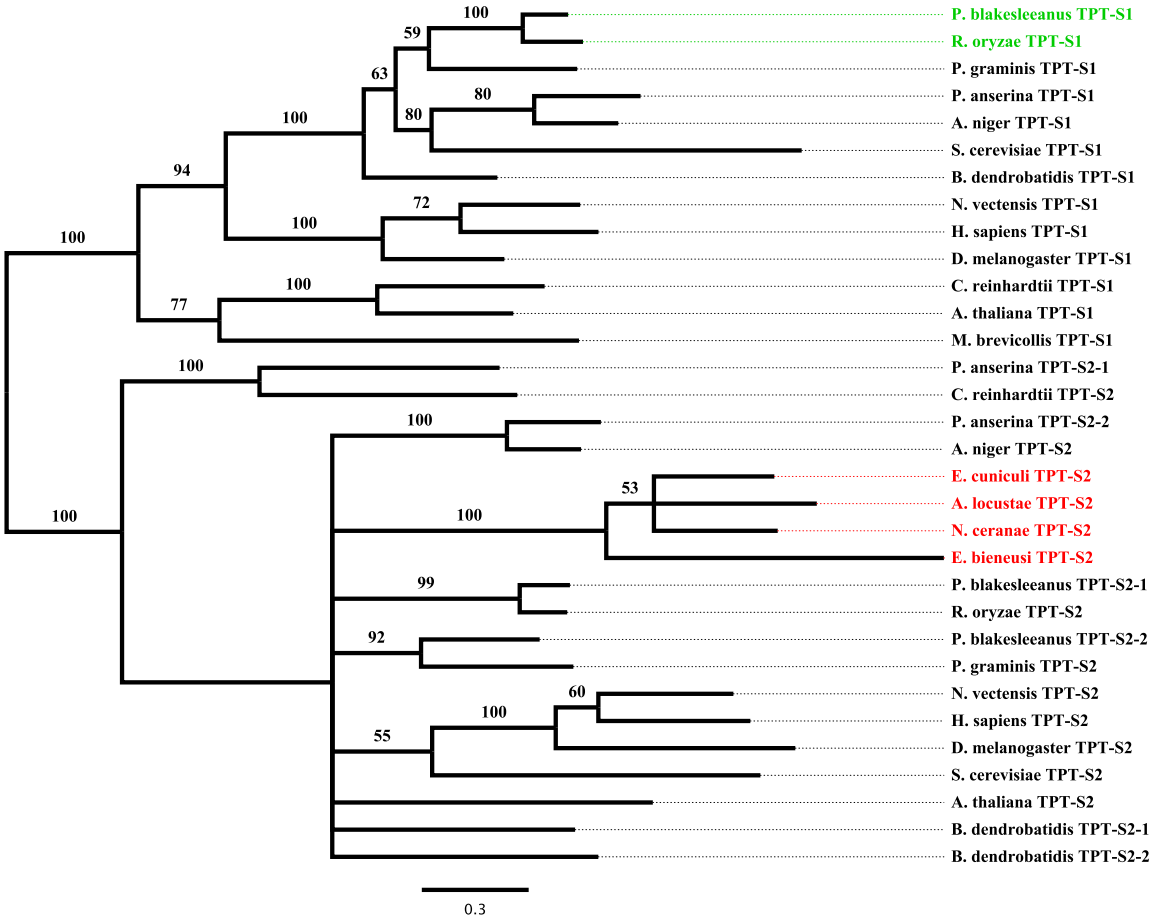


Figure C.2: Maximum likelihood tree of the S1 and S2 TPTs. Sequences in the zygomycete sex related region are labeled in green, sequences in the corresponding region of the microsporidia are labeled in red. Branch labels denote bootstrap support values.

	TPT-S1	TPT-S2	RNA helicase-S1	RNA helicase-S2
P. blakesleeanus	gw1.41.68.1	gw1.4.230.1 e_gw1.8.104.1	estExt_fgeneshPB_pg.C_410079	e_gw1.1.921.1
R. oryzae	predicted*	RO3G_03053	RO3G_01291	RO3G_01291
A. niger	estExt_GeneWisePlus.C_11009	estExt_GeneWisePlus.C_4037	e_gw1.5.51.1	fgenesh1_pg.C_scaffold_3000464
P. anserina	Pa_6_5520	Pa_3_8480 Pa_1_11780	Pa_1_16040	Pa_6_3090
P. graminis	PGTG_02172	PGTG_17729		PGTG_11063
S. cerevisiae	YOR307C	YML038C		YDL031W
B. dendrobatidis	e_gw1.1.368.1	fgenesh1_pg.C_scaffold_1001144 estExt_Genewise1.C_70123		
E. cuniculi		Q8SV84		Q8SRN8
A. locustae		1098		
E. bieneusi		XP_002651458.1		XP_002650711.1
N. ceranae		EEQ82475.1		EEQ81838.1
H. sapiens	ENSP00000248069	ENSP00000361301	ENSP00000360817	ENSP00000304072
D. melanogaster	FBpp0077036	FBpp0073002	FBpp0083940	FBpp0072475
N. vectensis	estExt_gwp.C_320251	estExt_gwp.C_840152	e_gw.488.2.1	e_gw.29.18.1
M. brevicollis	fgenesh2_pg.scaffold_16000087			estExt_Genewise1.C_40354
A. thaliana	Q9ZSR7	Q8H184		O49289
C. reinhardtii	205633	192150	127992	123188

Table C.1: Gene IDs of the RNA helicases and the TPTs. Data sources: Joint Genome Institute (`http://genome.jgi-psf.org`): *B. dendrobatidis*, *P. blakesleeanus*, *A. niger*, *N. vectensis*, *C. reinhardtii*, *M. brevicollis*; Broad Institute (`http://www.broad.mit.edu/`): *R. oryzae*, *P. graminis*; EBI (`http://www.ebi.ac.uk/integr8`): *E. cuniculi*, *A. thaliana*; ENSEMBL (`http://www.ensembl.org`): *H. sapiens*, *D. melanogaster*, *S. cerevisiae*; Podospora anserina Genome Project (`http://podospora.igmors.u-psud.fr/`): *P. anserina*; Josephine Bay Paul Center (`http://gmod.mbl.edu`): *A. locustae*.
*this sequence has been inferred directly from the *R. oryzae* genome sequence since the corresponding gene is not present in the annotated gene set of this species (see also Idnurm et al. 2008).

Cluster-Id[$]	*R. oryzae* protein	*E. cuniculi* Blast hit[△%]	Blast e-value	*E. cuniculi* ortholog[△]	Chromosome
1[&]	RO3G_00003	19068537	7.0E-30	-	I
	RO3G_00004	19069234	7.0E-05	-	VII
	RO3G_00005			-	
	RO3G_00006	19068534	3.0E-09	-	I
2*	RO3G_00009	19170836	2.0E-28	19170836	V
	RO3G_00010	19170838	5.0E-09	19170838	V
3	RO3G_01355	19171027	2.0E-23	19171027	VIII
	RO3G_01356	19171031	1.0E-06	-	VIII
	RO3G_01357			-	
	RO3G_01358			-	
	RO3G_01359			-	
	RO3G_01360	19171024	1.0E-67	-	VIII
4	RO3G_02423	19168763	1.0E-17	19168763	III
	RO3G_02424			-	
	RO3G_02425	19069007	3.0E-10	-	VI
	RO3G_02426			-	
	RO3G_02427			-	
	RO3G_02428	19168761	6.0E-12	-	III
5	RO3G_03308	19068592	4.0E-44	19068592	II
	RO3G_03309	19068589	9.0E-09	-	II
6*[&]	RO3G_03530	19068611	2.0E-63	19068611	II
	RO3G_03531			-	
	RO3G_03532	19068606	3.0E-16	19068606	II
	RO3G_03533			-	
	RO3G_03534	19170972	3.0E-15	19170972	VIII
	RO3G_03535	19068608	1.0E-10	19068608	II
7	RO3G_04457	19168720	9.0E-29	19168720	III
	RO3G_04458			-	
	RO3G_04459	19168723	5.0E-24	-	III
8[&]	RO3G_04463	19068607	6.0E-13	19068607	II
	RO3G_04464			-	
	RO3G_04465			-	
	RO3G_04466			-	

	RO3G_04467			-	
	RO3G_04468			-	
	RO3G_04469			-	
	RO3G_04470	19171321	5.0E-13	19171321	IX
	RO3G_04471			-	
	RO3G_04472	19068605	2.0E-15	-	II
1[&]	RO3G_05485	19068537	1.0E-32	-	I
	RO3G_05486	19171356	1.0E-17	19171356	IX
	RO3G_05487	19068534	1.0E-84	19068534	I
1	RO3G_05592	19068534	2.0E-08	-	I
	RO3G_05593	19068537	9.0E-24	-	I
9	RO3G_05835	19069147	5.0E-20	-	VII
	RO3G_05836	19069144	9.0E-83	19069144	VII
○	RO3G_05865	19168645	9.0E-04	-	III
	RO3G_05866	-		19168649	III
	RO3G_05867			-	
	RO3G_05868	19168651	8.0E-30	-	III
#	RO3G_06214	19168723	1.0E-49	-	III
	RO3G_06215			-	
	RO3G_06216			-	
	RO3G_06217			-	
	RO3G_06218			-	
	RO3G_06219			-	
	RO3G_06220			-	
	RO3G_06221	19168720	4.0E-30	19168720	III
#[&]	RO3G_07828	19068701	1.0E-13	-	II
	RO3G_07829			-	
	RO3G_07830			-	
	RO3G_07831			-	
	RO3G_07832			-	
	RO3G_07833			-	
	RO3G_07834			-	
	RO3G_07835			-	
	RO3G_07836			-	
	RO3G_07837			-	

	RO3G_07838			-	
	RO3G_07839	19069299	2.0E-09	-	X
	RO3G_07840			-	
	RO3G_07841			-	
	RO3G_07842	19068700	2.0E-06	-	II
2*	RO3G_08158	19170838	6.0E-26	19170838	V
	RO3G_08159	19170836	6.0E-29	19170836	V
2*	RO3G_08282	19170836	2.0E-28	19170836	V
	RO3G_08283	19170838	6.0E-26	19170838	V
7	RO3G_08889	19168720	3.0E-11	19168720	III
	RO3G_08890			-	
	RO3G_08891			-	
	RO3G_08892	19168723	6.0E-22	-	III
o[&]	RO3G_09031	19069361	6.0E-08	-	X
	RO3G_09032			-	
	RO3G_09033			-	
	RO3G_09034	19068933	2.0E-11	-	VI
	RO3G_09035	19069362	2.0E-04	-	X
10	RO3G_10613	19170970	4.0E-48	-	VIII
	RO3G_10614			-	
	RO3G_10615			-	
	RO3G_10616	19170971	9.0E-10	-	VIII
11*	RO3G_10956	19069314	1.0E-113	19069314	X
	RO3G_10957	19069313	7.0E-07	19069313	X
o[&]	RO3G_11255	19168644	9.0E-15	-	III
	RO3G_11256			-	
	RO3G_11257	19068709	7.0E-08	-	II
	RO3G_11258			-	
	RO3G_11259			-	
	RO3G_11260			-	
	RO3G_11261			-	
	RO3G_11262			-	
	RO3G_11263			-	
	RO3G_11264			-	
	RO3G_11265			-	

	RO3G_11266			-	
	RO3G_11267			-	
	RO3G_11268	19168645	2.0E-04	-	III
#[&]	RO3G_12238	19069100	3.0E-24	-	VII
	RO3G_12239			-	
	RO3G_12240			-	
	RO3G_12241			-	
	RO3G_12242			-	
	RO3G_12243			-	
	RO3G_12244			-	
	RO3G_12245			-	
	RO3G_12246			-	
	RO3G_12247			-	
	RO3G_12248	19068936	3.0E-18	-	VI
	RO3G_12249			-	
	RO3G_12250			-	
	RO3G_12251			-	
	RO3G_12252	19069102	2.0E-12	-	VII
12[&]	RO3G_12682	19168744	4.0E-36	19168744	III
	RO3G_12683			-	
	RO3G_12684	19069355	1.0E-178	19069355	X
	RO3G_12685			-	
	RO3G_12686			-	
	RO3G_12687			-	
	RO3G_12688			-	
	RO3G_12689	19168742	1.0E-13	-	III
13[&]	RO3G_12888	19069579	1.0E-119	19168679	III
	RO3G_12889				
	RO3G_12890				
	RO3G_12891	19069662	5.0E-23	19069662	XI
	RO3G_12892				
	RO3G_12893				
	RO3G_12894				
	RO3G_12895	19069580	2.0E-10		III
6*	RO3G_13173	19068611	1.0E-60	19068611	II

	RO3G_13174	19068608	1.0E-10	19068608	II
14[&]	RO3G_13322	19069361	7.0E-06	-	X
	RO3G_13323	19170962	4.0E-33	-	VIII
	RO3G_13324			-	
	RO3G_13325			-	
	RO3G_13326	19069362	3.0E-09	-	X
○[&]	RO3G_14144	19068528	3.0E-94	-	I
	RO3G_14145			-	
	RO3G_14146			-	
	RO3G_14147	19170857	2.0E-38	19170857	V
	RO3G_14148			-	
	RO3G_14149			-	
	RO3G_14150			-	
15	RO3G_14309	19069097	2.0E-08	-	VII
	RO3G_14310	19069100	2.0E-25	-	VII
○	RO3G_14965	19068691	5.0E-04	-	II
	RO3G_14966			-	
	RO3G_14967	19068690	1.0E-15	19068690	II
2*	RO3G_15055	19170836	2.0E-28	19170836	V
	RO3G_15056	19170838	6.0E-26	19170838	V
16	RO3G_15651	19069192	7.0E-47	19069192	VII
	RO3G_15652	19069191	7.0E-10	-	VII
16	RO3G_15944	19069192	6.0E-46	19069192	VII
	RO3G_15945	19069191	3.0E-09	-	VII
#	RO3G_16080	19168720	3.0E-30	19168720	III
	RO3G_16081			-	
	RO3G_16082			-	
	RO3G_16083			-	
	RO3G_16084			-	
	RO3G_16085			-	
	RO3G_16086			-	
	RO3G_16087			-	
	RO3G_16088	19168723	3.0E-40	-	III

Table C.2: Re-analysis of the proposed 33 syntenic cluster between *Rhizopus oryzae* and *Encephalitozoon cuniculi*.
$ Conserved gene cluster between *R. oryzae* and *E. cuniculi*. Cluster involving the same *E. cuniculi* genes are counted only once.
$^{\triangle}$ NCBI gi ID of the *E. cuniculi* proteins.
$^{\%}$ Blast hit reported by Lee et al. (2008).
* Syntenic cluster involving orthologous genes.
$^{\&}$ *E. cuniculi* genes assigned to a *R. oryzae* gene cluster but reside on different chromosomes are marked in yellow. These clusters indicate that the data is not compatible with the decision rules described in figure S5 of Lee et al. (2008). There, two gene cluster are defined to be syntenic if two genes in *E. cuniculi* that have at most three intervening genes have homologs in *R. oryzae* that are separated by at most 4 genes. If one of the intervening genes in *E. cuniculi* also has a homolog in *R. oryzae*, then the *R. oryzae* genes must be in a window of less than 15 gene (Figure 4, rule 2).The yellow marked *E. cuniculi* genes are not intervening and can only be found when the search was performed with the *R. oryzae* genes.
Number of intervening genes in *R. oryzae* exceeded.
$\circ$ E-value limit of 10.0E-05 exceeded

Appendix D

Supplementary Text to Chapter 6

Tree reconstruction For the tree reconstruction, we took 598 OMA (Altenhoff et al., 2010) groups of orthologs for which at least one species from each of four systematic categories is represented. The four systematic categories are given in the following where the numbers of ortholog groups in which a species is represented is given in parenthesis: Group1 (Metazoa): *Lottia gigantea* (463), *Daphnia pulex* (448), *Homo sapiens* (504), *Caenorhabditis elegans* (429), *Nematostella vectensis* (438); Group2 (Fungi): *Cryptococcus neoformans* (403), *Yarrowia lipolytica* (462); Group3 (Viridiplantae): *Arabidopsis thaliana* (466), *Oryza sativa* (427), *Ostreococcus lucimarinus* (389); Group4: *Leishmania major* (306), *Cryptosporidium parvum* (227), *Plasmodium falciparum* (200), *Trypanosoma brucei* (227), *Dictyostelium discoideum* (481), *Giardia lamblia* (136). The 598 core-ortholog groups were used to perform a subsequent broader ortholog search with HaMStR (Ebersberger, Strauss and von Haeseler, 2009) in 235 proteomes. The strict version of the HaMStR search was chosen, which uses all species from the core-ortholog group as reference species. 113 ortholog groups were retained for which an ortholog was found in at least 80% of the proteomes. These groups were complemented with orthologs from 9 further species whose annotated genome became available in the course of the analysis and were then used for the tree reconstruction. First, the sequences were aligned with MAFFT –linsi (Katoh et al., 2005). The alignments were then concatenated and columns with more than 20% gaps or an X in the sequences were removed. The final alignment comprised 244 species and 34,540 amino acids. We used raxmlHPC v. 7.2.2 and the PROTGAMMAILGF model of sequence evolution (Stamatakis, 2006) for tree reconstruction, performing 100 bootstrap replicates. The bootstrap consensus tree was computed with tree puzzle (Schmidt and von Haeseler, 2007).

Traceability calculation of SUS1 We followed the procedure described in chapter 6 to estimate the evolutionary traceability for the yeast SUS1 protein. To this end, we took the patristic distance (sum of branches connecting two taxa) from the reconstructed phylogeny (see previous paragraph) as average pairwise distances between proteins from yeast and from each of the other 243 species under study. We aligned the OMA (Altenhoff et al., 2010) group of orthologs including the yeast SUS1 protein with MAFFT –linsi (Katoh et al., 2005), reconstructed a tree with FastTree (Price, Dehal and Arkin, 2010), and again took the patristic distances as pairwise distances between the yeast SUS1 and its orthologs. We annotated the yeast SUS1 with Pfam domains (Finn et al., 2010) using hmmscan from the HMMER3 software package (`http://hmmer.janelia.org/`) and simulated its evolution with REvolver (chapter 3). Insertion and deletion rates where estimated according to the parsimony approach described in chapter 4. Blast (Altschul et al., 1997) using the simulated sequences as queries was used to search for the protein with the highest score in yeast.

Phylogenetic profiling The search for orthologs to the yeast SUS1 protein was performed with HaMStR (Ebersberger, Strauss and von Haeseler, 2009). The core-ortholog set was constructed in an iterative way as follows: First, the initial core-ortholog set consisted only of the yeast protein. A pHMM was constructed based on the single sequence (hmmbuild; `http://hmmer.janelia.org/`) and then used to search for the most similar protein in any of the 243 species under study. To this end, all significant hmmsearch hits (default settings) served as query for a Blast search (Altschul et al., 1997) against yeast. If an ortholog candidate found the protein from the core-ortholog set as best Blast hit, it was added to the list of candidate core-ortholog proteins. A Smith-Waterman alignment (Smith and Waterman, 1981) between each protein from the list of candidate core-orthologs and the yeast protein from the core-ortholog set was computed. The protein with the highest similarity to the yeast protein was added to the set. The new core-ortholog set was aligned with MAFFT –linsi (Katoh et al., 2005) and a new pHMM was constructed. In the next iteration, all hmmsearch hits were used as query for Blast searches against each species that was already present in the core-ortholog set. Those proteins that found the respective protein from the core-ortholog set as best Blast hit, were added to the list of candidate core-orthologs. Of these, only the one with the highest average similarity to the proteins in the core-ortholog set was added to the set. This procedure of identifying and adding new proteins to the core-ortholog set was repeated till the

core-ortholog set consisted of 5 sequences plus the initial yeast protein (SUS1). The final pHMM was constructed based on the set of 6 core-orthologs. With this pHMM we performed a final HaMStR search in all species. Again, the strict version of HaMStR was used where all 6 species from the core-ortholog set are reference species.