The PhylProM database. Extending the use of phylogenetic profiles and their applications for membrane proteins.

Anna Thorén

October 24, 2000

Contents

1 Introduction

In the emerging era of functional genomics, a major goal is to attach functions to as many proteins as possible. Until recently, the only non-experimental method to achieve functional knowledge about a protein was by inference from the known function of a homologous protein. However, in a pioneering study Eisenberg and coworkers [Pellegrini *et al.*, 1999] introduced the method of phylogenetic profiles as a non-homology based method. This was the first of a number of methods that managed to go beyond the traditional sequence matching technique, providing functional relationships information for non-homologous proteins [Pellegrini *et al.*, 1999, Marcotte *et al.*, 1999a, Marcotte *et al.*, 1999b, Enright *et al.*, 1999].

While the concept of phylogenetic profiles is easy to grasp, the construction of comprehensive profiles for a large set of genomes is a computationally demanding task. In order to facilitate the access to phylogenetic profiles, an easily search- and browse-able database has therefore been created (http://www.sbc.-su.se/PhylProM/). The database includes phylogenetic profiles based on 24 prokaryotic and 2 eukaryotic genomes. PhylProM is particularly useful for the study of integral membrane proteins, but all encoded proteins are included.

Although the method of phylogenetic profiles predicts protein functions on a non-homology basis, it still requires a first step where a query protein finds some homolog. Depending on the parameters used in this initial protein sequence comparison, some organisms may be erroneously excluded from a given phylogenetic profile, while others may be erroneously included. This finding along with the fact that it is possible that a specific biochemical function, perhaps due to lateral transfer, has been replaced by another protein in an organism, suggests that the grouping of proteins according to their phylogenetic profiles may benefit from considering not only identical but also "closely related" profiles. Two different methods to define these closely related phylogenetic profiles have therefore been utilized. One is simply based on the number of organisms that differ between profiles, while the other is based on their differential parsimony. The PhylProM database has been set up to allow searches using "degenerate" phylogenetic profiles.

All phylogenetic profiles might not contain the same amount of functional information. For instance, proteins that only exist in one organism or proteins that exist in all organisms are not very likely to be functionally related. On the other hand, if proteins are found in only a few distantly related organisms with similar ecological and metabolic constraints, it is more likely that they are functionally linked. To explore this, all profiles have been sorted into groups depending on their number of organisms and on their total parsimony. In general, profiles that contain an intermediate number of organisms or that are of intermediate to high total parsimony are found to be the best guide to functional linkages. Higher total parsimony trees are indicative of either multiple lateral transfer events or selective gene loss, both of which are dictated by divergent selective pressures.

The method of phylogenetic profiles might be of different use for different types of functional similarities. Functions that are present in almost all organisms or in only a single organism are most likely not detectable, while functions present in a subset of organisms should be easier to detect. These functions involve specific biochemical pathways only necessary for some organisms.

2 Background and theory

2.1 Proteins and protein functions

In all living organisms genes are the units of genetic information. Genes encode proteins that play crucial roles in virtually all biological processes. All proteins are constructed from the same set of 20 amino acids, but each protein has its own unique combination. The amino acid building blocks are joined together by peptide bonds into chains of amino acids, each chain with a free NH_2 group at one end and a free COOH group at the other. The ends are referred to as the N-terminal and the C-terminal, and the amino acids of each protein are conventionally written in the N- to C-terminal direction.

Proteins that are bound to membrane surfaces or that span through membranes are called membrane proteins. Membranes are organized and sheetlike assemblies that consist mainly of proteins and lipids. They build up highly permeable barriers around cells in all types of organism and around organelles within cells in eukaryotic organisms. In order to appropriately respond to changes in the external environment, like for instance changes in nutrient and hormone levels, specific signals are transmitted across the membranes. These transmissions events would not be possible without the help of membrane proteins.

Membrane proteins can be classified as either peripheral or integral, depending on whether they are bound to the membrane surface or span through the inner section of the membrane. A common feature for integral membrane proteins is the presence of at least one transmembrane region. This consists of an alpha helical stretch of hydrophobic amino acids, and can therefore be seen as a region of low compositional complexity. The hydrophobicity is imposed by the constraint that residues buried in the membrane need to perform hydrophobic interactions with the membrane lipids [Sonnhammer *et al.*, 1998b]. An integral membrane protein with domains exposed on both sides of the membrane is called a transmembrane protein. The length of the tail, either N- or C-terminal, that protrudes from the membrane surface and the number of regions spanning the membrane vary from one transmembrane protein to another.

Since each amino acid can be either acidic or basic, hydrophobic or hydrophilic, or have a straight, branched or aromatic chain, the alphabet of amino acids in all different types of proteins allows for great diversity in protein function. The term function can be described in different ways [Bork *et al.*, 1998]. For instance, each protein can be associated with a molecular function. This somewhat elementary function is connected to the presence of specific binding sites for substrates, nucleic acids or other proteins, and includes functions like for instance the catalyse of metabolic reactions and the transmission of signals to other proteins. Furthermore, since proteins rarely work alone, sets of cooperating proteins can be seen as responsible for physiological functions or cellular functions. These include for instance metabolic pathways and signal transduction cascades, many of which are crippled by the loss of any of its individual components [Marcotte, 2000]. Finally, the physiological subsystems and their interplay with various environment stimuli, can be seen to determine the phenotypic function, the morphology and physiology of the organism and its behavior.

2.2 Functional predictions

A major goal in the emerging era of functional genomics is to find functions for as many proteins as possible. Until recently, the only non-experimental way to achieve functional knowledge about a newly sequenced protein was to search for some homolog that had already been characterized, and assign its function to the uncharacterized protein. However, in many cases functional similarity could not be reflected in sequence similarity [Yona *et al.*, 1999]. This was true for cases in which, for instance, non-homologous proteins cooperated in the same function within an organism or proteins in different organisms performed the same biological functions but lacked a common evolutionary origin. The lack of sequence similarity resulted in a failure for the traditional method, and hence begged for complementary methods for functional prediction not based on homology.

In a pioneering study Eisenberg and coworkers introduced the method of phylogenetic profiles for functional predictions [Pellegrini *et al.*, 1999]. This was the first of several following methods that managed to go beyond the traditional sequence matching techniques, providing functional relationships information for non-homologous proteins [Pellegrini *et al.*, 1999, Marcotte *et al.*, 1999a, Marcotte *et al.*, 1999b, Enright *et al.*, 1999].

The method of phylogenetic profiles focuses on the correlation of protein inheritance across different species. The idea is that functionally linked proteins evolve in a correlated fashion, and hence have homologs in the same set of organisms. Since it is most likely that an organism that requires a specific function carries most or all of the genes for its protein components, while any organism that lacks the same function also lacks all of the component genes, the idea of the method seems promising.

An example of another strategy for detecting protein functions on a non-homology basis is given by the Rosetta Stone method, which focuses on the correlation of protein domains across different species. It infers protein interaction from genome sequences, based on the finding that some pairs of interacting proteins have homologs that are fused into a single protein in another organism. The single protein is referred to as a Rosetta Stone sequence, and it is believed to imply functional linkage between corresponding not fused proteins in other organism.

Since it is not obvious that all proteins that interact in one organisms have proteins that are fused in another organism, the method of phylogenetic profiles is the most powerful of the two new methods [Huynen *et al.*, 2000]. However, both still have the power to group functionally related proteins, even when the involved proteins have never been characterized. Furthermore, since they make their functional analysis on a non-homology basis, they have the ability to not only find the elementary functions for individual proteins, but, also the higher level functions for entire complexes and pathways [Bork & Eisenberg, 2000].

2.3 Sequence similarity search

Although direct sequence similarity is not the basis for the analysis in the non-homologous functional prediction methods, many of them, including the method of phylogenetic profiles, still require the use of an efficient search method for a query protein to find some sequence homolog.

Today there are several different similarity search tools to choose between, with the NCBI BLAST 2.0 version as one example [Altschul *et al.*, 1997]. The algorithms used in the BLAST 2.0 version all seek for local as opposed to global alignment, which means that proteins are compared by small sequence fragments, instead of by total sequences. The power of using local alignments is that it makes it possible to find relationships among sequences that share only isolated regions of similarity.

Two different algorithms offered by the NCBI BLAST 2.0 version are blastall and blastpgp. A search using either of these begins by looking for segment pairs of equal length within the compared protein sequences. All found alignments that have a score value above a certain threshold score T, and whose scores cannot be improved by extension or trimming are called 'high scoring segments pairs' or HSPs.

To assess whether a given HSP constitute evidence for homology, blastall and blastpgp gives a measure of how strong the alignment can be expected from chance alone. The measure is called the "expect value", or E-value, and is given by:

$$E = Kmne^{-\lambda S} \tag{1}$$

where m and n are the lenghts of the compared protein sequences and K and λ can be thought of simply as natural scales for the search space size and the scoring system respectively. The necessary for a sufficient similarity threshold is different for each task [Yona *et al.*, 1999], and the E-value can therefore be set by the user ahead of each search.

Since blastall and blastpgp treat gaps, like deletions and insertions, in different ways, the alignments returned from each of the algorithms often differ. The blastall algorithm does not allow gaps to be introduced into the alignments. Similar sequences are therefore often broken up into several segments. The blastpgp algorithm, however, does allow the introduction of deletions and insertions. Since these events have been common in evolutionary history, the scoring returned from a gapped search using blastpgp tends to reflect biological relationships more closely.

Significant alignments returned from a blastpgp search can be further used in an iterative PSI-BLAST search. In this search the information in the alignments are used to construct a position specific score matrix. In essence, this means that every amino acid is assigned a score at each position in the derived matrix, that gives information on how well conserved the amino acid is at the particular position. The matrix is then used to replace the query sequence in a next round of database search. The power of the method is further enhanced through iteration of the search procedure, which may be done until no new significant alignments are found. The reason for using position-specific score matrices is that it has been found that these can help to detect distant similarities that otherwise can only be found through the comparison of three-dimensional structures.

As many as one fourth of all residues in proteins occur within regions of highly biased amino acid composition. Since the unusual composition of these low-complexity regions may give rise to artifactual hits due to segment composition instead of residue order, they may create problems in a sequence similarity search. Because of this, filters that remove the biased regions ahead of each search, can be used in PSI-BLAST. The filtering can be turned off, however, which is useful when query sequences contain large regions of low compositional complexity, as in the case with membrane proteins.

The current version of PSI-BLAST has one limitation, in that it requires the finding of some similar sequence in the initial search to work. The use of a strict E-value threshold may therefore result in that some protein homologs are missed.

2.4 Evolution

Several of the non-homology based method, including the method of phylogenetic profiles, take advantage of the genetic variation among different organisms to achieve functional knowledge about proteins. All these variations are due to different types of evolution, for instance traditional evolution and convergent evolution. In traditional means of evolution, functionally interacting non-homologous proteins go through a common ancestral pathway and cooperate in the same function in different species. In convergent evolution, however, proteins that perform redundant functions in different organisms lack any evolutionary origin.

In evolutionary history genes are affected by a number of evolutionary events. These are for instance deletions, insertions and duplications, which arise from the removal, the addition and the multiplication of gene fragments respectively. Since all genomes evolved from a common ancestral genome at the root of the evolutionary tree of life, the different evolutionary events that proteins within each genome have experienced can be analyzed using a phylogenetic tree. The root of such a tree divides into three major branches representing the primary groups; eubacteria, archaeabacteria and eukaryotes. Each major branch is further divided into a branching pattern based on the phylogenetic relationship between all genomes within that group, resulting in a non-overlapping hierarchical arrangement. The branches most far apart from the root of the tree, terminate into nodes representing the different genomes.

When examining the different evolutionary events that a protein has experienced, the first thing to do is to find the most parsimonious tree, i.e. the tree that requires the smallest number of substitutions along all its branches. The reason for this is that this tree is expected to have a high degree of correspondence to true phylogeny.

3 Methods

3.1 Sequence file construction

The method of phylogenetic profiles only works on completely sequenced genomes. The genomes included in this study were therefore 18 fully sequenced eubacterial genomes: Chlamydia pneumoniae (n), Chlamydia trachomatis (i), Synechocystis PCC6803 (c), Bacillus subtilis (b), Mycoplasma genitalium (g), Mycoplasma pneumoniae (p), Mycobacterium tuberculosis (r), Aquifex aeolicus (q), Thermotoga maritima (v), Campylobacter jejuni (w), Escherichia coli (e), Haemophilus influenzae (h), Helicobacter pylori J99 (j), Helicobacter pylori 26695 (u), Rickettsia prowazekii (x), Borrelia burgdorferi (o), Treponema pallidum (l) and Deinococcus radiodurans (f), six fully sequenced archaebacterial genomes: Aeropyrum pernix (d), Archaeoglobus fulgidus (a), Methanobacterium thermoautotrophicum (t), Methanococcus jannaschii (m), Pyrococcus abyssi (s) and Pyrococcus horikoshii (k), and two fully sequenced eukarytotic genomes: Saccharomyces cerevisiae (y) and Caenorhabditis elegans (C).

All protein sequences were downloaded from the public database of Entrez at http://www.ncbi.nlm.nih.gov/Entrez. The database is provided by the National Center for Biotechnology Information (NCBI) and contains protein sequences from the Swiss-Prot, PITR, PRF, PDB and translated protein sequences from the DNA sequence databases GenBank, EMBL and DDBJ. The downloaded sequences were placed in a large file of a total number of 70435 proteins. Based on knowledge that the all-against-all similarity search to follow was to be a very time- and memory consuming process, all protein sequences were also subdivided into smaller sequence files, one per genome, which allowed them to be run simultaneously against the larger file and save a lot of time and memory.

3.2 Membrane protein predictions

The method of phylogenetic profiles can be optimized for the study of different types of proteins. In this project the phylogenetic profiles were to be particularly useful for the study of integral membrane proteins. In order to predict the presence or absence of membrane spanning alpha helices, TMHMM was run on each protein. The server is available at http://www.cbs.dtu.dk/-services/TMHMM, and the method behind the server is carefully described in [Sonnhammer *et al.*, 1998b]. Although the resulting phylogenetic profiles were to be particularly useful in integral membrane protein studies, all encoded proteins were included in the project.

3.3 Pairwise comparisons

A basic step in the method of phylogenetic profiles is to search for all proteins that are related. In this project this search was done by running an all-against-all search on all protein sequences using PSI-BLAST [Altschul *et al.*, 1997]. In this was allowed for a maximum number of 10 iterations. To better detect membrane proteins no filters for low compositional complexity regions were used, but instead a more stringent E-value for inclusion in each iteration round as well as a more stringent E-value for reporting matches, i.e. homologs. Suitable E-values were found by running PSI-BLAST, with different parameter settings, for E.coli proteins involved in the well studied membrane embedded Sec and Tat protein translocation systems [Sargent *et al.*, 1998]. As found in a previous membrane protein study [Jones *et al.*, 1994], an E-value around 1e-15 turned out to be a good choise. The 32 best matches for SecY in E.coli in a PSI-BLAST search with an E-value of 1e-15 for inclusion in each iteration round, an E-value of 10 for reporting matches and a maximum number of 10 iterations, is shown in Table 1.

Protein	Annotation	E-value
gi|1789696	putative ATPase subunit of translocase	e-200
gi|1573808	preprotein translocase SecY subunit (secY)	e-178
gi|2982812	preprotein translocase SecY	e-165
gi|3861179	preprotein translocase SecY	e-162
gi|4155830	preprotein translocase SecY	e-160
gi|2314464	preprotein translocase subunit (secY)	e-160
Cj1688c	secY	e-158
gi|3328947	Translocase	e-157
gi|4376927	Translocase	e-156
gi|4982044	preprotein translocase SecY subunit	e-155
gi|2632403	preprotein translocase subunit	e-153
gi|6459912	preprotein translocase, SecY subunit	e-153
gi|3322478	preprotein translocase subunit (secY)	e-152
gi|2911006	secY	e-152
gi|1652409	preprotein translocase SecY subunit	e-149
gi|2688393	preprotein translocase subunit (secY)	e-138
gi|3844764	preprotein translocase (secY)	e-130
gi|1674352	SecY	e-130
gi|5457750	protein translocase subunit (secY)	e-109
gi|3258185	hypothetical preprotein translocase secY subunit	e-109
gi|2621063	preprotein translocase SecY	e-107
gi|2826286	protein translocase, subunit SEC61 alpha (secY)	e-102
gi|5104652	hypothetical preprotein translocate secY subunit	e-102
gi|2648637	protein translocase, subunit SEC61 alpha (secY)	e-101
gi|3881191	similar to protein transport protein SEC61	2e-95
gi|6323411	Sec61p	3e-87
gi|6319760	Probable SEC61 protein homolog; Ssh1p	2e-70
gi|5458506	sodium-dependent phosphate transporter	3e-05
gi|2634591	similar to hypothetical proteins	0.002
gi|3861164	unknown	0.003
gi|6320075	vacuolar H+/Ca2+ exchanger; Vcx1p	0.003
gi|2688416	conserved hypothetical integral membrane protein	0.010

Table 1: The 32 best matches for SecY in E.coli in a PSI-BLAST search with an E-value of 1e-15 for inclusion in each iteration, an E-value of 10 for reporting matches and a maximum number of 10 iterations. No filters for low compositional complexity were used.

In this it can be seen that true homologs, i.e. other SecY proteins, separated from other proteins in the particular PSI-BLAST search. Since a similar result was seen for the other Sec and Tat proteins, the E-value for inclusion in each iteration round was set to 1e-15. True homologs were in all results, generally, found to have lower E-values than 1e-15. The threshold for reporting matches were, therefore, also set to 1e-15.

Homologs found in a PSI-BLAST search can either be orthologs or paralogs [Tatusov *et al.*, 1997]. Orthologs are proteins in different species that have evolved from a common ancestral protein, whereas paralogs correspond to proteins that are related by one or several duplication events within an organism. While orthologs are found to normally retain the same function in the course of evolution, paralogs are believed to evolve new functions. Paralogs, therefore, constitute a potential problem in the search for homologs with identical functions. However, in this project this problem was not considered as necessarely that important. This was based on the fact that paralogous proteins still perform related functions that can be thought of as the same [Benner *et al.*, 2000] to the resolution of protein functions used in this project.

A PSI-BLAST search result is not necessarily symmetric. Two proteins might be identified as homologs when the search is performed starting with one protein as query, but not with the other protein as a query. Furthermore, even if two proteins are reported as homologs in both search directions their E-values usually differ. In order to make the results from the all-against-all search symmetric each case where the matches were not symmetric, were replaced with the most significant E-value for that case.

3.4 Phylogenetic profile construction

The method of phylogenetic profiles presents the occurance of related proteins in different genomes by constructing a phylogenetic profile for each protein. In this project the profiles were constructed based on the symmetric all-against-all PSI-BLAST search result.

A phylogenetic profile is basicly a string with p positions. Each position corresponds to a specific genome, and the number of positions is therefore identical to the number of genomes used. For a specific protein the presence of a homolog in the pth genome was, in this project, indicated with a single letter of the alphabet, assigned to represent the genome, in the pth position. If no homolog was found the position was marked with a dash.

According to the idea of the method of phylogenetic profiles functionally linked proteins have homologs in the same set of genomes. In order to find these proteins, all proteins with identical phylogenetic profiles were therefore grouped. Within the different profile groups, several of the proteins did not only share profile similarity, but similarity in sequence as well. Since the method of phylogenetic profiles aims to detect protein function on a non-homology basis, all these homologous proteins had to be further sorted into families of sequence neighbors. This was done by using an algorithm for the selection of sets of dissimilar proteins [Hobohm *et al.*, 1992].

The selection aimed to find a set of non-homologous proteins, that could represent all the different families of sequence neighbors within each profile group. From a total list of protein sequence pair relations, collected from the symmetric all-against-all PSI-BLAST search result, a successive thinning out of clusters of similar proteins was done. For a given profile group, this thinning out process started by removing the protein with the largest number of sequence neighbors. This protein was then no longer counted as a neighbor to any of the remaining proteins, and so the number of neighbors in the new situation had to be reassessed. Again the protein with the largest number of neighbors was removed. The process was repeated until the proteins left over had no more sequence neighbors. These proteins were then chosen to represent all the different families of sequence neighbors. In case more than one protein shared the largest number of neighbors in the thinning out process, proteins were removed in the alphabetical order of their organisms.

3.5 Phylogenetic profile reconstruction

Homologous proteins have in some cases evolved too far to be recognized by sequence comparison [Tauer & Benner, 1997], especially when strict similarity search parameters are used. This fact along with the possibility of protein replacement events in specific biochemical functions, suggests that the grouping of proteins according to their phylogenetic profiles may benefit from considering not only identical but also "closely related" profiles. In this project the closely related profiles were detected using two different approaches. In the first approach the number of organisms that differed between profiles were simply counted. Phylogenetic profiles that differed in less than three positions were considered as closely related.

In the second approach the finding of related profiles was instead based on the differential parsimony between profiles. The idea behind this approach was that it might be more accurate to detect closely related profiles using a model in which all the evolutionary events that had resulted in the appearances of the different phylogenetic profiles was taken into account. Phylogenetic profiles with a parsimony difference less than three were here defined as closely related.

Differential parsimony values were obtained by comparing the most parsimonious tree for different profiles. Each tree was found by using the Fitch algorithm [Fitch, 1971]. To begin with, a basic evolutionary tree was constructed from the phylogenetic relationships between the 26 complete genomes. This information was downloaded from the NCBI at http://www.ncbi.nlm.nih.gov/. Using this tree for each phylogenetic profile, the 26 nodal sets corresponding to the different genomes were set to 1 in case the genome was present in the profile, and to 0 otherwise.

The reconstruction of preliminary ancestral nodes then proceeded from the descendent nodal sets by formulating an ancestral nodal set for an immediate ancestor based on the intersection of its immediate descendent nodal sets. In case the intersection did not exist, the nodal sets were instead formed by a union of the immediate descendent nodal sets. The reconstruction worked successively backwards, one node at a time, until the preliminary ancestral set for the most distant ancestor, i.e. the common ancestor of all species at the root of the tree of life, had been formulated.

The preliminary ancestral nodal sets were reconstructed to possible ancestral sets, working in the reverse direction. First the preliminary set for the ultimate node was made the final set for that node. Then the construction proceeded to the penultimate node, finding the final set for this node by applying some specific rules. These rules said that if the nodal set contained all the sets in the final nodal set of its immediate ancestor, all sets not present in the final nodal set was to be removed and what was left was to be made the final set for the penultimate node. The reconstruction of possible ancestral sets could then continue to the following preliminary nodal set. Otherwise, i.e. if the preliminary nodal set did not contain all the sets in the final set of its immediate ancestor, the way the preliminary phase had been formulated decided what was to be done. If it had been formulated by a union of its immediate descendents, any state present in the final nodal set of its immediate ancestor, but missing in the preliminary set, was to be added to the preliminary nodal set. If it had not been formulated by a union, any state not already present, provided that they were present in the final set of its immediate ancestor as well as in at least one of the immediately descendent preliminary nodal sets, was to be added to the preliminary nodal set. In either cases the nodal sets were, after the addition, to be considered as final nodal sets, and the reconstruction could continue to the following preliminary nodal set. In order to distinguish the states added in the final phase from those already present in the preliminary phase, all states added in the final phase were included in a parenthesis.

When the reconstruction of all nodal sets had been done, the next step was to find all permitted links, or transitions, between the states of the different nodal sets, and exclude all that were not permitted. The applied rule was that a state not included in a parenthesis, was to be the only valid terminal point of a link which originated from the same state in its immediate ancestor, meaning that going from 1 to 1, from 0 to 0, from (1) to 1 or from (0) to 0 was to be obligatory as long as the descendent characters 1 or 0 existed. However, in case they did not exist all possible transitions were permitted apart from the ones from 1 to (0), 0 to (1), (1) to (0) or (0) to (1).

Each permitted link was associated with a certain probability, or node value. This probability was defined based on two assumptions; all permitted states present in the ultimate nodal set are equiprobable, meaning that if a state is permitted to be either of character 1 or 0, the probabilities $P({}^{5}0)$ and $P({}^{5}1)$, where ${}^{k}N$ denoted the state in the kth node, are both of value 0.5, and every permitted link from a given nodal state, $L({}^{k}N)$, is equiprobable. From these two suppositions the node values were defined as:

$$P[L({}^{k}N)] = P({}^{k}N)/n \tag{2}$$

where n corresponded to the number of descending links from ${}^{k}N$. From this a value for each branch was found by:

$$bv = t - b \tag{3}$$

where bv was the branch value, obtained by subtracting the node value at the bottom of the branch (b) from the node value at the top of the branch (t). From this, two trees could be compared for similarity as:

$$dp = \sum |bv(A) - bv(B)| \tag{4}$$

to measure the differential parsimony between phylogenetic trees A and B.

In addition to using the constructed phylogenetic trees for finding closely related phylogenetic profiles, they were also made use of in the search for the phylogenetic profiles that contained the largest amounts of functional information.

In this all profiles were sorted into groups according to their total parsimony, defined for a single profile as:

$$tp = \sum |bv(A)| \tag{5}$$

An example of a phylogenetic profile with two closely related profiles is given in Figure 1. In this a sample phylogenetic profile is given (kstmadhejuwnibrcfvq), with a total parsimony of 14.5. Two additional phylogenetic profiles are given to illustrate the difference between the methods for finding closely related profiles. Since both profiles have a differential parsimony of 1, they are considered as closely related using the differential parsimony approach. However, one of the profiles has a bitwise difference of 6 and is therefore not considered as closely related using the organism difference approach.

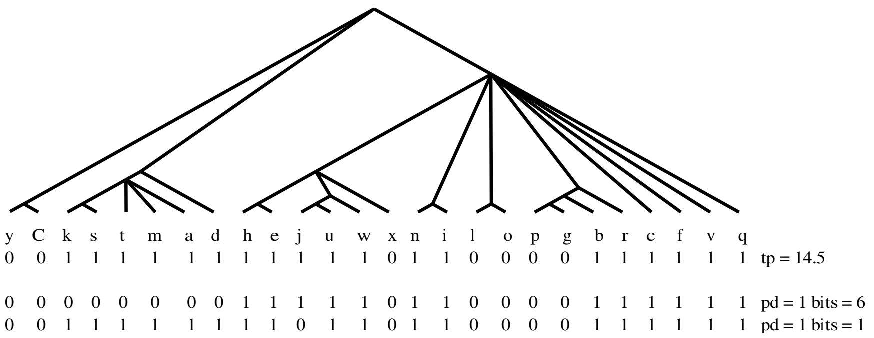


Figure 1: Descriptive figure of bitwise difference and differential parsimony.

In the search for profiles with valuable functional information, phylogenetic profiles were sorted depending on their number of organisms as well.

3.6 Keyword analysis

Functional knowledge about proteins can only be found if the used method is able to provide functional information. In order to examine the performance of the method of phylogenetic profiles three different tests were carried out in this project. In the performance tests it was necessary to have some knowledge about the function of each protein. This knowledge was found in the Swissprot [Bairoch & Apweiler, 1996] keyword lines, after excluding the keywords hypothetical, membrane protein, transmembrane, inner membrane, outer membrane and 3D-structure, since they were not believed to provide any significant functional information. Altogether this resulted in 291 different keywords. Using these, functionally similar proteins were defined as proteins that shared at least one keyword.

All three performance tests were carried out on the E.coli genome. The reason for this was that E.coli is one of the genomes with the largest fraction of proteins classified in the Swissprot database. Prior to any of the tests, proteins found only within the E.coli genome itself were excluded, based on the assumption that a large fraction of proteins found only within one single organism is not very likely to be functionally related. The remaining E.coli proteins were divided into groups of linked proteins, i.e. proteins that had either been reported as homologs in the all-against-all PSI-BLAST search or that shared identical or "closely related" phylogenetic profiles.

Previous studies using the method of phylogenetic profiles had pointed out that the functional predictive power of the method was dependent on the function that was examined [Pellegrini *et al.*, 1999]. The first test in this project was therefore done in order to examine if the method could be used to predict protein functions, and if so, what types of functions the method could detect. In this test the "keyword recovery rate", i.e. the number of linked proteins that shared a specific keyword divided by the number of randomly chosen protein that shared this keyword, was measured for the 291 keywords.

As mentioned, all phylogenetic profiles were not assumed to contain the same amount of functional information. In order to find functionally significant phylogenetic profiles, profile groups with identical total parsimony values as well as identical number of organisms were in the second test compared with the keyword recovery rates.

Previous studies using phylogenetic profiles had focused on the study of non-homologous functionally cooperating proteins [Pellegrini *et al.*, 1999]. However, in addition to proteins like this, there are proteins that also lack a common evolutionary origin and also perform the same functions but in different organisms. In case these proteins are involved in a function present in all organisms, they can possibly be detected using inverse phylogenetic profiles. In the third test the performance of inverse phylogenetic profiles were therefore examined, using modified keyword recovery rates. The modified keyword recovery rates were found by dividing the number of proteins with inverse profiles that shared a particular keyword, with the number of randomly chosen proteins that shared this keyword.

Although the PhylProM database had been optimized for membrane proteins, the performance of it was examined mainly using non-membrane proteins. The reason for this was that the use of stringent E-values in the similarity search was not believed to significantly change the search results for globular proteins. Furthermore, since many more non-membrane proteins existed, the comparisons would be more reliable using these proteins as well.

4 Results and discussion

4.1 The PhylProM database

Based on an all-against-all PSI-BLAST search of all proteins encoded in 24 prokaryotic and 2 eukaryotic genomes and topology prediction of the potential membrane proteins in the collection (see Methods), the PhylProM database at http://www.sbc.su.se/PhylProM/ has been developed. The database is the first publicly available easily accessible database for phylogenetic profiles. PhylProM is primary intended for the study of integral membrane proteins (which have not been properly treated in previous phylogenetic profile analyses), and the PSI-BLAST search parameters have also chosen with this in mind. However, the database can be used to study all encoded proteins, although the strict parameters used mean that some weak but real sequence similarities between globular proteins may have been missed.

Altogether the database includes data for 70003 proteins. This number is somewhat lower than the number of sequences downloaded at the opening (see Methods). The lower number is due to short operon leader sequences, that do not even report themselves as homologs in a PSI-BLAST similarity search with parameters chosen as in this project. Since these proteins give rise to phylogenetic profiles that only contain dashes, they are not included in the PhylProM database.

Running TMHMM on all proteins, result in a fraction of 25% of predicted integral membrane proteins. This total fraction lies within the range of a 20 to 30% integral membrane proteins suggested in a previous membrane protein study [Wallin & von Heijne, 1998]. In the same study, a fairly good correlation between the fraction of integral membrane proteins and the genome size was found. In Figure 2 the fraction of TMHMM predicted integral membrane proteins for each genome is plotted versus the size of the genome. Since TMHMM predicted transmembrane regions in the N-terminal region sometimes turn out to be signal peptides, two different fractions are given for each genome; the fraction of all predicted integral membrane proteins (circles) and the fraction of predicted integral membrane proteins with more than one transmembrane region (stars). True integral membrane protein content probably lies somewhere in between.

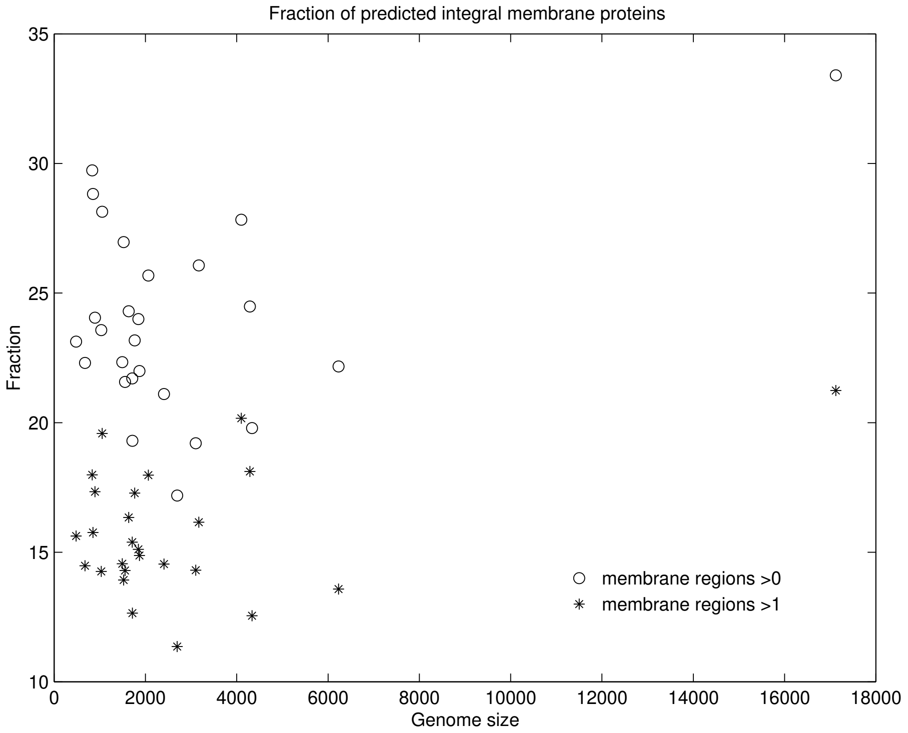


Figure 2: Plot of fraction (in %) of all predicted integral membrane proteins (circles) and of predicted integral membrane proteins with more than one membrane region (stars) versus genome size.

In Figure 2 it can be seen that although the C.elegans genome, with the largest number of encoded proteins, also has the largest fraction of predicted integral membrane proteins, the correlation between genome size and integral membrane protein content is not that clear. This might be due to the fact that all genomes except C.elegans are uni-cellular organisms, and that these organisms might not show the same correlation as multi-cellular organisms.

All proteins included in PhylProM are distributed over 4497 different phylogenetic profiles, of which 1541 contain predicted integral membrane proteins. The number of protein families within each profile group greatly varies. Phylogenetic profile groups with more than 100 protein families are shown in Table 2. In this it can be seen that the largest number of proteins families, 6353, is found within the C.elegans profile. This might not be that surprising, since the C.elegans genome contains approximately three times as many proteins as

Phylogenetic profile	Protein families
`-------------------------C`	6353
`----y---------------------`	2828
`---------------------d----`	1660
`---------b----------------`	1472
`----------r---------------`	1373
`----------------------f---`	1225
`--------e-----------------`	1171
`-------c------------------`	1154
`----y--------------------C`	784
`a-------------------------`	733
`--t-----------------------`	533
`---k----------------------`	483
`-m------------------------`	459
`------v-------------------`	427
`------------------------w-`	364
`-----------------l--------`	346
`------------uj------------`	329
`-----q--------------------`	321
`-----------h--------------`	289
`---k-------------------s--`	274
`----------------o---------`	243
`------------------in------`	232
`--------------------x-----`	223
`--------e--h--------------`	201
`amtkyqvcebrhujgpolinxdfswC`	194
`-------------------n------`	177
`--------------gp----------`	143
`-----------------------s--`	116
`------------uj----------w-`	112

Table 2: Number of protein families within phylogenetic profile groups with more than 100 families.

the second largest genome; S.cerevisiae, and consequently even more times as many as the rest of the included genomes. In addition to the C.elegans profile, several other single organism phylogenetic profiles are found to contain a large number of protein families. The number of proteins unique to one organism might indicate the wide variety and adaptability of the bacterial world, and the ability to metabolize a variety of chemical compounds.

Another finding in Table 2 is that a number of profiles with species that are close evolutionary related in theory contain large numbers of protein families. The result reveals similarities between the involved organisms, and address the existing theories of the evolutionary relationship between the involved organisms.

The phylogenetic profile group with protein families present in all 26 organisms has also a large number of families. This finding indicate that many functional pathways are conserved through evolution.

As many as 3641 of the total number of 4497 different phylogenetic profiles in PhylProM contain only a single protein family. Unfortunately, this makes the non-homology based finding of functionally related proteins impossible within these profile groups. However, if the functional analyze is expanded to include closely related phylogenetic profiles as well (See Methode), the single protein family profiles may be used in the possible detection of links between non-homologous proteins.

4.2 Identical phylogenetic profiles

In previous studies using the phylogenetic profile method it has been pointed out that the information in phylogenetic profiles depends on the function that is examined [Pellegrini *et al.*, 1999]. To identify what type of functions the method of phylogenetic profiles is able to detect, the keyword recovery rates for 291 Swissprot keywords were analyzed for the E.coli genome.

In Figure 3 the number of keywords with a particular keyword recovery rate is plotted against the recovery rate. For a few very specific functions a majority of the non-homologous proteins found within a phylogenetic profile group share keywords in Swissprot, which results in a high keyword recovery rate. This is the case for the keyword coupled to the existence of flagella or not. Flagella is the propeller by which most mobile bacteria move. As many as 63% of all non-homologous proteins within the phylogenetic profile (qvebujolw) share a keyword coupled to this function. For a larger group of functional keywords the phylogenetic profile groups increase the recovery rate of functionally related proteins, but still only a small fraction of all proteins that have the same profile actually share the same function. This is the case for keywords like for instance Ribosome, Transcription and tRNA synthetase/ligase. For an even larger group, with Iron as an example, no significant increase can be seen in the rate.

The empty bars in Figure 3 corresponds to the keyword recovery rates for the traditional homology based method. Using this method a similar recovery rate distribution trend as above can be seen. However, in this case a significant larger proportion of the keywords have high keyword recovery rates. Standard homology based methods hence seem to be better than phylogenetic profiles. This might not be that surprising. Computer analysis of complete prokaryotic genomes have, for instance, shown that microbial proteins are in general highly conserved, with about 70% of them containing ancient conserved re-

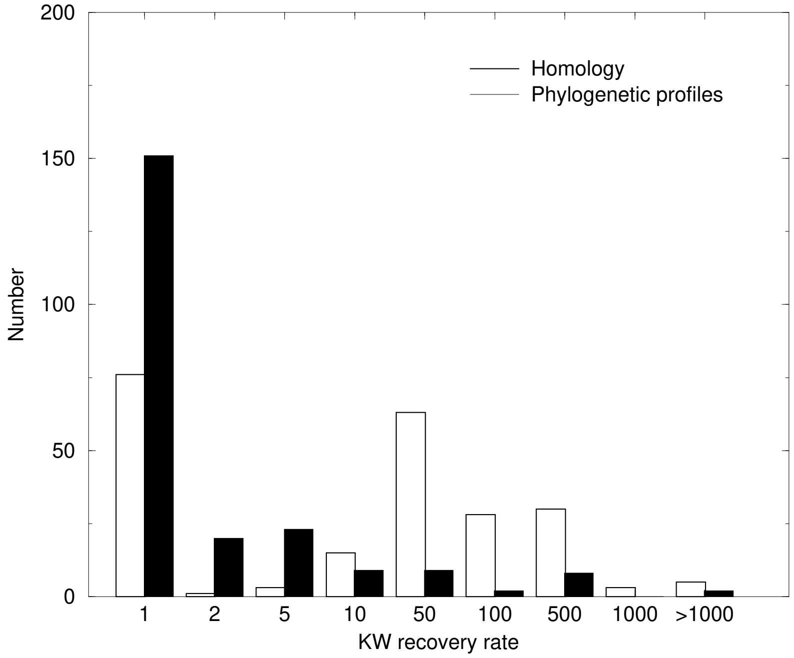


Figure 3: The number of keywords plotted against the keyword recovery rate for all keywords in E.Coli. The empty bars are for standard homology searches and the filled bars for exact phylogenetic profile matches.

gions [Koonin *et al.*, 1998]. However, while sequence homology reveals proteins of identical or equivalent functions, the method of phylogenetic profile aims to identify interacting proteins, proteins with related functions or proteins operating in the same cellular context [Marcotte, 2000]. The methods can, therefore, be seen to complement each other in functional prediction processes.

The keywords with a keyword recovery rate higher than five for the exact profiles are shown in Table 3. Interestingly, it can be seen in this that many of the functions for which exact phylogenetic profiles (Exact in Table 3) are most successful are involved in biosynthesis of small molecules. Several other are involved in specific functions unique for some organisms, such as chemotaxis, cell wall or flagella. The most successful example from Table 3 is that of Organic Radical. This refers to enzymes that enable bacteria to metabolize

Keyword	Family	Exact	Profiles	Parsimony
Organic Radical	1030	1441	3	131
Selenocysteine	0	1439	239	103
Lysine biosynthesis	0	481	8	13
Pyridoxine biosynthesis	0	361	11	12
Selenium	0	360	40	30
Topoisomerase	360	192	5	11
Leucine biosynthesis	481	175	30	27
DNA integration	0	155	4	4
Riboflavin biosynthesis	201	144	80	57
Tryptophan biosynthesis	0	144	7	19
Sensorytransduction	95	96	49	77
Flagella	22	84	50	57
Lipid biosynthesis	65	37	15	23
Histidine biosynthesis	50	31	22	14
Ubiquinone biosynthesis	107	30	5	9
Peptidoglycan synthesis	84	21	3	3
Phosphotransferase system	77	18	4	3
Chemotaxis	32	16	9	10
Cell wall	53	15	4	3
Multifunctional enzyme	11	11	2	2
Aminoacyl-tRNA synthetase	130	11	6	3
Phosphorylation	38	10	6	4
Folate biosynthesis	0	9	4	3
Protein biosynthesis	46	8	4	4
Ubiquinone	18	7	8	7
Excision nuclease	0	7	3	2
ATP-binding	15	7	4	2
Fattyacid biosynthesis	22	6	3	1
Purine biosynthesis	31	6	4	1
SOS response	7	6	3	2

Table 3: Keyword recovery rate for the keywords with most information from exact phylogenetic profiles.

hydrocarbons, including alkylbenzenes, alkanes, and alkenes, using a radical intermediate [Heider *et al.*, 1999]. Organisms that share this keyword are all capable of metabolizing hydrocarbons anaerobically, a trait that likely arose convergently under selective pressure [Castresana & Moreira, 1999].

4.3 Closely related profiles

The use of strict similarity search parameters as well as the possibility of protein replacement events in specific biochemical functions, suggests that functional prediction might benefit from considering closely related profiles. The similarity between two profiles can be measured either by counting the number of organisms that are present in one profile but not in the other, or by counting their differential parsimony.

The results from the analyze using closely related profiles are presented in Table 3. In this it can be seen that the functional prediction performance for the differential parsimony approach (Parsimony in Table 3) as well as for the organism difference approach (Profiles Table 3) is worse than for the exact phylogenetic profiles. Hence, considering closely related profiles does not improve the functional prediction power for the method of phylogenetic profiles. However, in many cases the differential parsimony values and the simple comparisons of non-exact profiles still show large keyword recovery rates. In these cases they might provide significant functional information.

It is hard to decide from the results in Table 3 which of the two approaches that are the best one to use when searching for related profiles. The differential parsimony values are higher in some cases, while the simple comparisons of non-exact profiles are better in other. This might indicate that the most important factor for non-identical phylogenetic profiles are simple search misses. However, the differential parsimony approach has a potential problem in that of lateral gene transfer, i.e. the passing of genetic material not from a parent to daughter cell but across species lines. Lateral gene transfer is common in eubacterial and archaebacterial species, especially among housekeeping genes [Doolittle, 1999, Jain *et al.*, 1999], and make individual trees less parsimonious. In case interacting proteins are transferred together this should not effect the differential parsimony values. However, in the future the development of branch length weighted parsimony approaches based upon the network of life [Strimmer & Moulton, 2000] may become a more ideal model for finding the closely related profiles.

4.4 Informative phylogenetic profiles

Phylogenetic profiles might not contain the same amount of functional information. In order to detect the most informative profiles, the keyword recovery rates for profiles sorted into groups depending on their total number of organisms as well as on their total parsimony values were analyzed for the E.coli genome. In Figure 4 each group is plotted against the average keyword recovery rates for all proteins within the group. As predicted, profiles with a low total parsimony or few organisms have a lower recovery rate. The highest recovery rate, when considering total parsimony values, is not obtained for the extreme total parsimony rates, but for rates about 10. However, this is probably largely due to a few profiles. When considering number of organisms, phylogenetic profiles

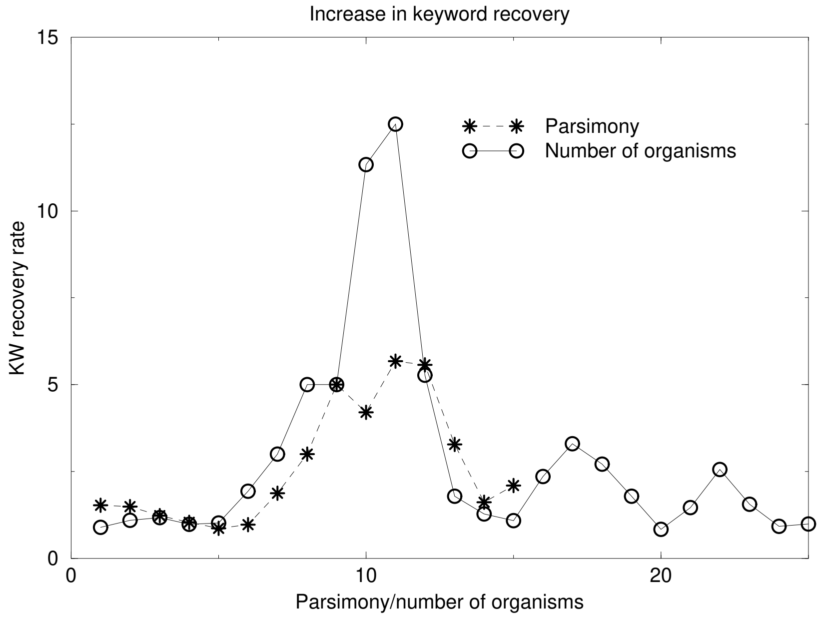


Figure 4: Plot of average increase in keyword recovery versus average total parsimony (stars) and number of organisms in the profile (circles).

that contain approximately half the organisms have a higher keyword recovery rate than other profiles. Although the large height of the peak may be due to a few profiles, it is clear that profiles with an intermediate number of organisms contain most information.

4.5 Inverse phylogenetic profiles

In previous comparisons of the archaeal, eukaryotic, and bacterial genomes, a specific type of non-homologous proteins have been found [Galperin *et al.*, 1998]. Although the proteins have been shown to lack a common evolutionary origin, they still perform the same biological function in different species. The most likely mechanism for the convergent evolution of these unrelated proteins, appears to be recruitment of existing proteins that take over new functions by virtue of changed substrate specificity or a modified mechanism. If the biological function is present in all organisms, the involved proteins might be detected by examining inverse phylogenetic profiles.

When this investigation is done, three keywords; DNA-directed DNA polymerase (recovery rate=49), DNA repair (recovery rate=9), Isomerase (recovery rate=8), are found to have a recovery rate higher than two. DNA replication systems have been shown to present prominent examples of convergent evolution. Comparison of the amino acid sequences of DNA polymerases has demonstrated a great deal of evolutionary conservation between a variety of prokaryotes, eukaryotes and viruses, forming the basis for grouping them into four divergent families with differences in functional activity [Ito & Braithwaite, 1991, Ito & Braithwaite, 1993, Kodra, 1998]. It has been suggested [Leipe *et al.*, 1999] that DNA replication has evolved twice independently and that the last common ancestor of bacteria and eukarya/archeas had a genetic system that contained RNA and DNA, with the latter produced by reverse transcription only. DNA repair and Isomerase have also been mentioned as functions that might have evolved through convergent evolution [Leipe *et al.*, 1999, Galperin *et al.*, 1998]. These findings demonstrate that examining inverse phylogenetic profiles can be useful for detecting functionally complementary proteins that have evolved through one of several mechanisms of convergent evolution. In general the existence of an inverse profile should imply either functional similarity, as detected here, or alternatively mutually exclusive functionalities.

4.6 PhylProM user interface

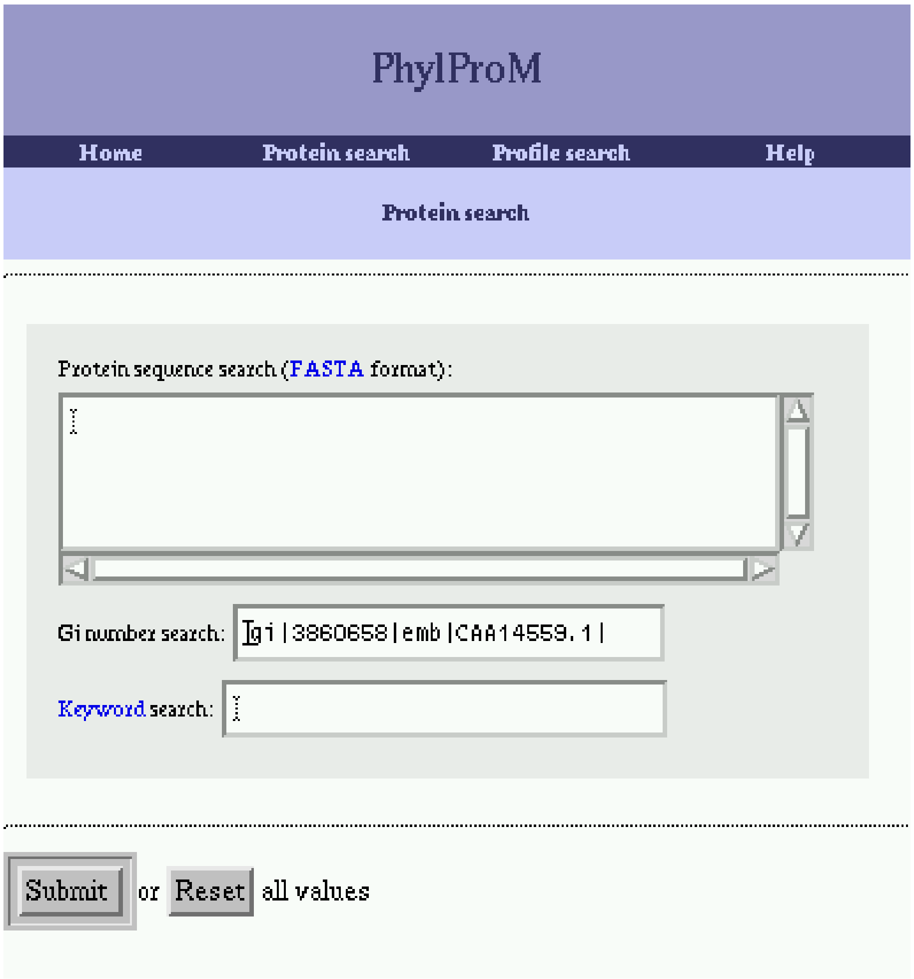
The PhylProM database has been designed to be easy to use. Browsing begins either by searching for a protein or a phylogenetic profile using the "Protein search" of "Profile search" respectively.

The protein search routine is shown in Figure 5. In this is can be seen that it is possible to search for a protein by sequence, name or keyword. A protein is searched for by sequence in the "Protein sequence search" area. Protein sequences can be presented in several different sequence file formats. One of the most common, and one that is also required in this search is the FASTA format. A protein sequence in FASTA format begins with a single line description of the protein, followed by lines of sequence data. The single line description always begins with a greater-than (">") symbol. In case a search is specified by sequence, PhylProM runs a BLAST search against all included proteins with the given protein as query. From this, all matches with an E-value below 1e-5 are reported. Generally, the lowest E-value match corresponds to the specified protein.

In the "Gi number search" area a protein can be searched for by name, i.e. gi number, corresponding to the molecular biology database identifier. In case a search is specified by name, PhylProM tells whether the protein is included in the database or not.

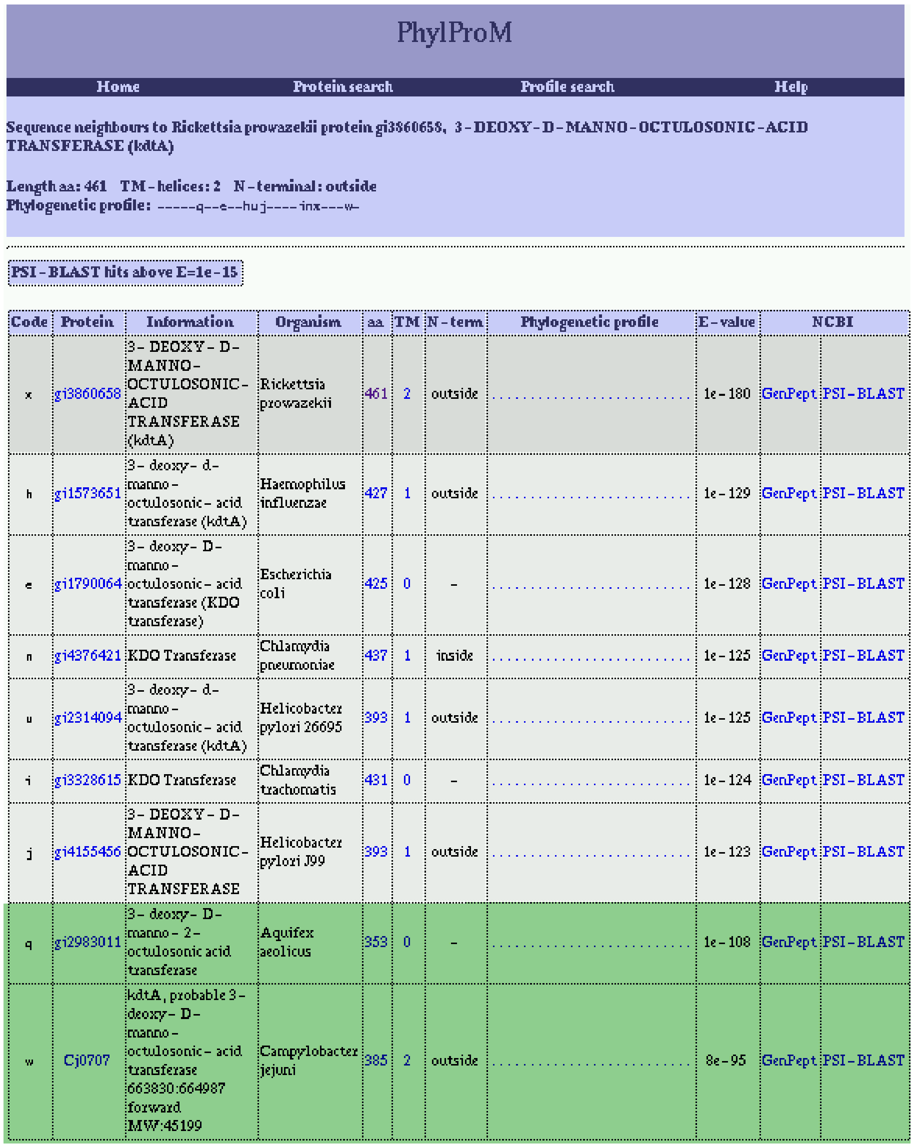
A protein is searched for by one or several keywords in the "Keyword search" area. In case a search is specified like this, PhylProM returns all proteins with matching words in their annotations. Whether the search is specified by amino acid sequence, gi number of keywords, all matches are reported as links to corresponding protein pages.

In Figure 5 the search is specified by name, using a gi number corresponding to a protein found in the PhylProM database. This particular search, hence, result in a match. The protein page for this match is shown in Figure 6.



Gi number search: gi|3860658|emb|CAA14559.1|

Figure 5: The Protein search routine. In this specific case the search is specified by gi number.



PhylProM

Sequence neighbours to Rickettsia prowazekii protein gi3860658, 3-DEOXY-D-MANNO-OCTULOSONIC-ACID TRANSFERASE (kdtA)

Length aa: 461 TM-helices: 2 N-terminal: outside
Phylogenetic profile: -----q--e--huj----inx---w-

PSI-BLAST hits above E=1e-15

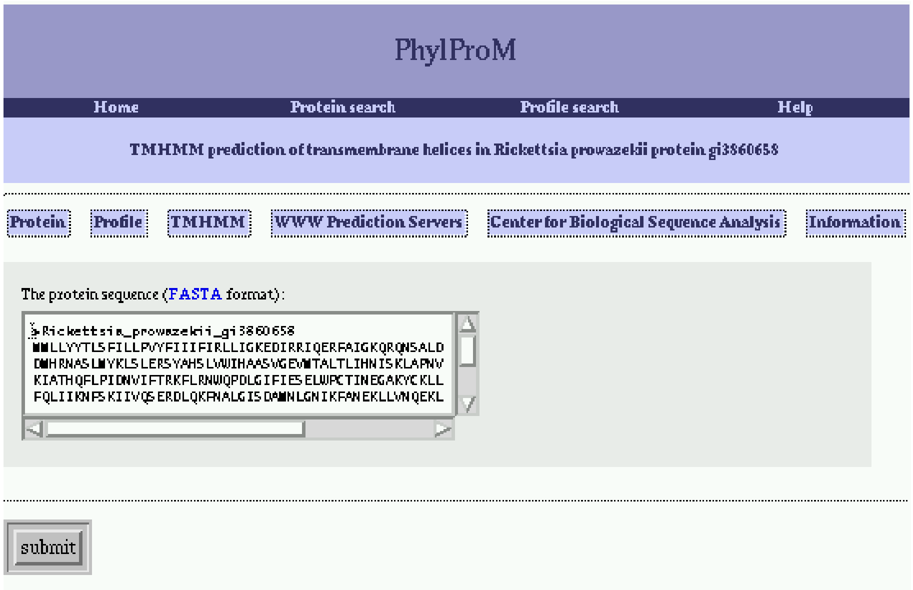
Code	Protein	Information	Organism	aa	TM	N-term	Phylogenetic profile	E-value	NCBI	
x	gi3860658	3-DEOXY-D-MANNO-OCTULOSONIC-ACID TRANSFERASE (kdtA)	Rickettsia prowazekii	461	2	outside	....	1e-180	GenPept	PSI-BLAST
h	gi1573651	3-deoxy-d-manno-octulosonic-acid transferase (kdtA)	Haemophilus influenzae	427	1	outside	....	1e-129	GenPept	PSI-BLAST
e	gi1790064	3-deoxy-D-manno-octulosonic-acid transferase (KDO transferase)	Escherichia coli	425	0	-	....	1e-128	GenPept	PSI-BLAST
n	gi4376421	KDO Transferase	Chlamydia pneumoniae	437	1	inside	....	1e-125	GenPept	PSI-BLAST
u	gi2314094	3-deoxy-d-manno-octulosonic-acid transferase (kdtA)	Helicobacter pylori 26695	393	1	outside	....	1e-125	GenPept	PSI-BLAST
i	gi3328615	KDO Transferase	Chlamydia trachomatis	431	0	-	....	1e-124	GenPept	PSI-BLAST
j	gi4155456	3-DEOXY-D-MANNO-OCTULOSONIC-ACID TRANSFERASE	Helicobacter pylori J99	393	1	outside	....	1e-123	GenPept	PSI-BLAST
q	gi2983011	3-deoxy-D-manno-2-octulosonic acid transferase	Aquifex aeolicus	353	0	-	....	1e-108	GenPept	PSI-BLAST
w	Cj0707	kdtA, probable 3-deoxy-D-manno-octulosonic-acid transferase 663830:664987 forward MW:45199	Campylobacter jejuni	385	2	outside	....	8e-95	GenPept	PSI-BLAST

Figure 6: The protein page for gi3860658 in Rickettsia prowazekii.

All included proteins in PhylProM have protein pages like this, that list all homologs found within the database. Since the the lists are limited to matches in PSI-BLAST with an E-value below 1e-15 (see Methode), it is possible to get lists of hits with higher E-values through the link to PSI-BLAST hits above 1e-15.

Various information are presented for the listed proteins. First of all, a single letter of the alphabet, assigned to represent each of the 26 genomes, indicate in which genome each protein exists. These letters are the same that are used in the phylogenetic profiles. Each protein has furthermore a specific gi number that is linked to its corresponding protein page, and a line of text that describe the protein some. The line of text came along with the protein sequences when they were downloaded from http://www.ncbi.nlm.nih.gov/Entrez. In addition to the code of the genome, the full name is also presented. For each protein is also given its certain number of amino acids, linked to corresponding protein pages in http://www.ncbi.nlm.nih.gov/Entrez, where the unique combination of the building blocks can be viewed.

Since PhylProM is particularly useful for integral membrane proteins studies, it presents the predicted number of membrane regions in all protein. The number of regions for each protein is linked to a page like the one shown in Figure 7. In this it can be seen that the sequence of the protein, corresponding to the selected number, is already pasted in the search area. TMHMM can, therefore, be run directly and make it possible to view the positions of predicted transmembrane regions, as well as whether the N-terminal is on the outside of the membrane or on the inside.



TMHMM prediction of transmembrane helices in Rickettsia prowazekii protein gi3860658

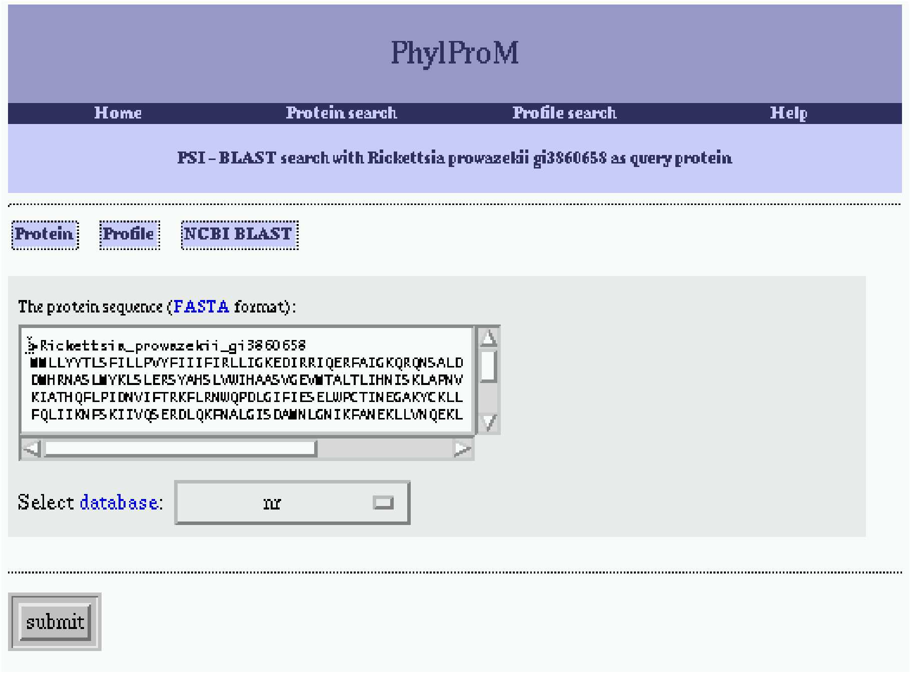
The protein sequence (FASTA format):

```
>Rickettsia_prowazekii_gi3860658
MMLLYYTLSFILLPVYFIIIFIRLLIGKEDIRRIQERFAIGKQRQNSALD
DMHRNASLMYKLSLERSYAHSLVWIHAASVGEVMTALTLIHNISKLAPNV
KIATHQFLPIDNVIFTRKFLRNWQPDLGIFIESELWPCTINEGAKYCKLL
FQLIIKNFSKIIVQSERDLQKFNALGISDAMNLGNIKFANEKLLVNQEKL
```

Figure 7: The TMHMM page for gi3860658 in Rickettsia prowazekii.

As mentioned (in Background), E-values indicate how strong each match is in a PSI-BLAST search. For each listed homolog the value that was reported when the search was done with the protein that owns the page as query, is presented.

In addition to amino acid sequences, http://www.ncbi.nlm.nih.gov/Entrez provides more information about all proteins, and "GenPept" is a link to this. NCBI:s database also provides different computational tools, with the PSI-BLAST similarity search algorithm as one example. Each listed homolog is linked to this algorithm, through a page like the one shown in Figure 8. In this it can be seen that the amino acid sequence for the selected protein is already pasted in the search area, and PSI-BLAST can be run directly with the protein as query. In this it is possible to choose which sequence database to search in, but the parameters are already set to an E-value of 1e-15 for inclusion in the first iteration and an E-value of 10 for reporting matches. The filter for low compositional complexity is turned off.



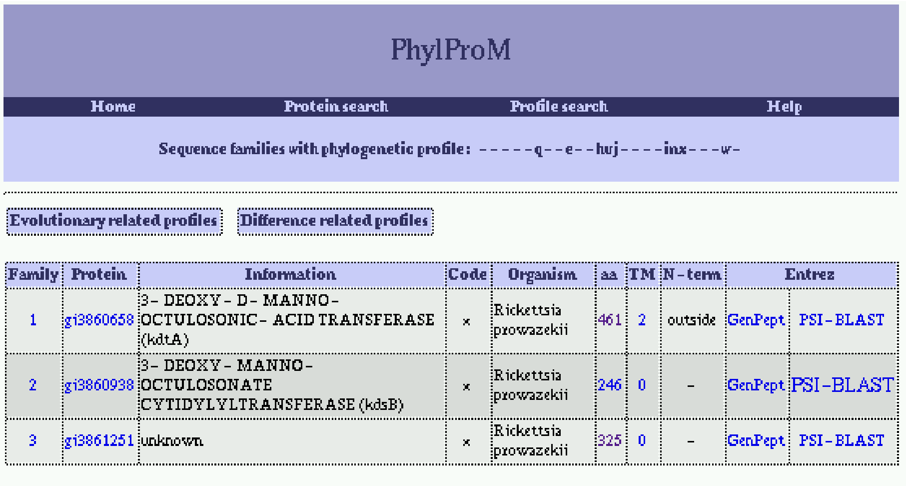
The protein sequence (FASTA format):

```
>Rickettsia_prowazekii_gi3860658
MMLLYYTLSFILLPVYFIIIFIRLLIGKEDIRRIQERFAIGKQRQNSALD
DMHRNASLMYKLSLERSYAHSLVWIHAASVGEVMTALTLIHNISKLAPNV
KIATHQFLPIDNVIFTRKFLRNWQPDLGIFIESELWPCTINEGAKYCKLL
FQLIIKNFSKIIVQSERDLQKFNALGISDAMNLGNIKFANEKLLVNQEKL
```

Select database: nr

Figure 8: The PSI-BLAST page for gi3860658 in Rickettsia prowazekii.

The phylogenetic profile for each listed protein is compared to the phylogenetic profile for the protein that owns the page. In the result of this comparison, all positions in the profile are either given a dot, a letter or a dash. Dots correspond to positions that look identical in the two profiles that are being compared. Letters indicate the presence of homologs in genomes in which the owner of the page has no homologs. Dashes indicate the absence of homologs in genomes in which the owner of the page has homologs. All phylogenetic profiles are linked to their corresponding profile pages, and for the protein gi3860658 this page is shown in Figure 9. Each phylogenetic profile has a page like this in PhylProM, that lists all protein families within the profile group and provide links to "closely related" profiles. The listed proteins in Figure 9 serve to



Sequence families with phylogenetic profile: -----q--e--huj----inx---w-

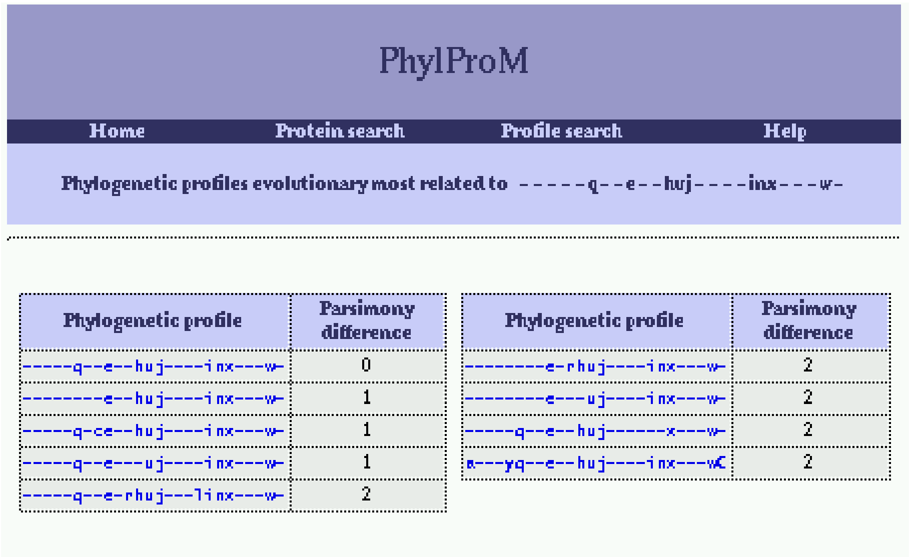
Family	Protein	Information	Code	Organism	aa	TM	N-term	Entrez	
1	gi3860658	3-DEOXY-D-MANNO-OCTULOSONIC-ACID TRANSFERASE (kdtA)	×	Rickettsia prowazekii	461	2	outside	GenPept	PSI-BLAST
2	gi3860938	3-DEOXY-MANNO-OCTULOSONATE CYTIDYLYLTRANSFERASE (kdsB)	×	Rickettsia prowazekii	246	0	-	GenPept	PSI-BLAST
3	gi3861251	unknown	×	Rickettsia prowazekii	325	0	-	GenPept	PSI-BLAST

Figure 9: Protein families within the profile group (qehujinxw).

represent each of the different families within the profile group, and are further linked to lists presenting all the other family members.

It is possible to reach two different lists of related profiles through the links “Evolutionary related profiles” and “Difference related profiles”. Evolutionary related profiles are profiles that have a parsimony difference less than three. The profiles related in this way to (qehujinxw) are shown in Figure 10. Difference related profiles are profiles that differ in less than three positions. In Figure 11 the profiles that fulfil this criteria when compared to (qehujinxw) are shown.

As mentioned, browsing can either begin by searching for a protein or a profile. The profile search routine is shown in Figure 12.

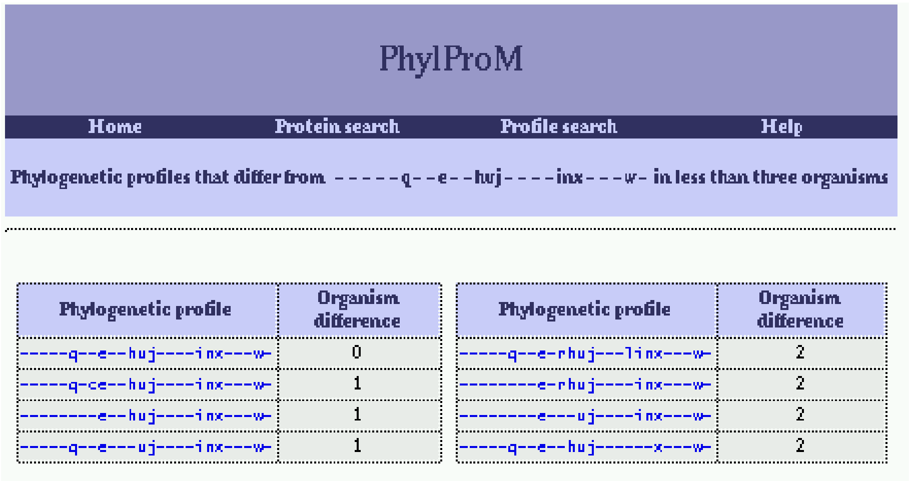


Phylogenetic profiles evolutionary most related to -----q--e--huj----inx---w-

Phylogenetic profile	Parsimony difference
-----q--e--huj----inx---w-	0
--------e--huj----inx---w-	1
-----q-ce--huj----inx---w-	1
-----q--e---uj----inx---w-	1
-----q--e-rhuj---linx---w-	2

Phylogenetic profile	Parsimony difference
--------e-rhuj----inx---w-	2
--------e---uj----inx---w-	2
-----q--e--huj------x---w-	2
a---yq--e--huj----inx---wC	2

Figure 10: Phylogenetic profiles with a parsimony difference less than three compared to (qehujinxw).

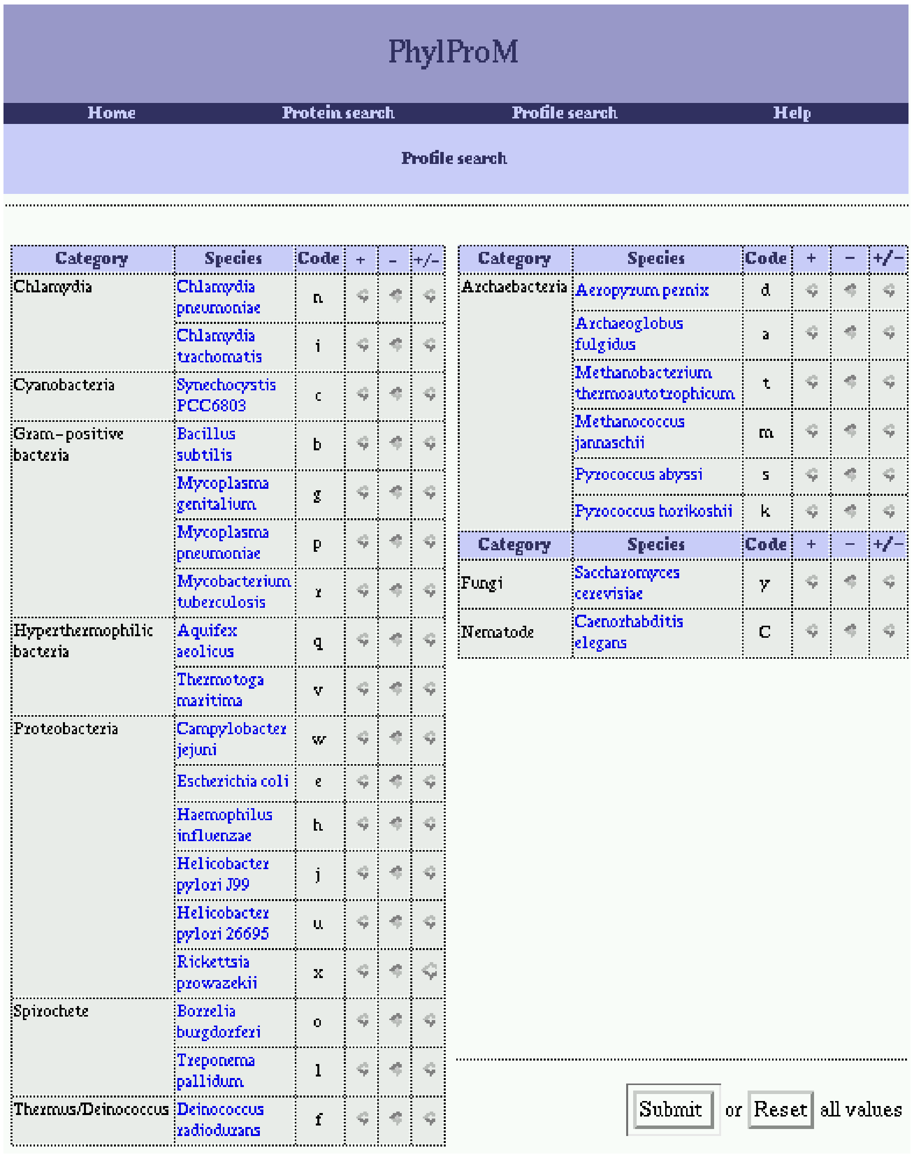


Phylogenetic profiles that differ from -----q--e--huj----inx---w- in less than three organisms

Phylogenetic profile	Organism difference
-----q--e--huj----inx---w-	0
-----q-ce--huj----inx---w-	1
--------e--huj----inx---w-	1
-----q--e---uj----inx---w-	1

Phylogenetic profile	Organism difference
-----q--e-rhuj---linx---w-	2
--------e-rhuj----inx---w-	2
--------e---uj----inx---w-	2
-----q--e--huj------x---w-	2

Figure 11: Phylogenetic profiles that differ in less than three organisms when compared to (qehujinxw).



Category	Species	Code	+	-	+/-
Chlamydia	Chlamydia pneumoniae	n	○	○	○
	Chlamydia trachomatis	i	○	○	○
Cyanobacteria	Synechocystis PCC6803	c	○	○	○
Gram-positive bacteria	Bacillus subtilis	b	○	○	○
	Mycoplasma genitalium	g	○	○	○
	Mycoplasma pneumoniae	p	○	○	○
	Mycobacterium tuberculosis	r	○	○	○
Hyperthermophilic bacteria	Aquifex aeolicus	q	○	○	○
	Thermotoga maritima	v	○	○	○
Proteobacteria	Campylobacter jejuni	w	○	○	○
	Escherichia coli	e	○	○	○
	Haemophilus influenzae	h	○	○	○
	Helicobacter pylori J99	j	○	○	○
	Helicobacter pylori 26695	u	○	○	○
	Rickettsia prowazekii	x	○	○	○
Spirochete	Borrelia burgdorferi	o	○	○	○
	Treponema pallidum	l	○	○	○
Thermus/Deinococcus	Deinococcus radiodurans	f	○	○	○

Category	Species	Code	+	−	+/−
Archaebacteria	Aeropyrum pernix	d	○	○	○
	Archaeoglobus fulgidus	a	○	○	○
	Methanobacterium thermoautotrophicum	t	○	○	○
	Methanococcus jannaschii	m	○	○	○
	Pyrococcus abyssi	s	○	○	○
	Pyrococcus horikoshii	k	○	○	○
Category	**Species**	**Code**	**+**	**−**	**+/−**
Fungi	Saccharomyces cerevisiae	y	○	○	○
Nematode	Caenorhabditis elegans	C	○	○	○

Figure 12: The Profile search routine.

In this it can be seen that all genomes included in PhylProM are listed, so that closely related species are grouped into similar categories. Each genome has three different radio buttons. A search for a specific phylogenetic profile or a number of different ones, is specified by the selection of one of these radio buttons per genome. The plus button is selected for all genomes that correspond to letters that are present in the profile/profiles to search for, the minus button for genomes that correspond to letters that are not present and the plus/minus button for all genomes that correspond to letters that are either present or not in the profiles to search for. The minus button is default for each genome.

5 Conclusions

Recently a number of new non-homology based methods for protein functional predictions have been introduced. In this project one of these, the method of phylogenetic profiles, have been utilized. This have resulted in the finding that the method of phylogenetic profiles is useful for detecting some types of functional relationships, and best applied on functions related to biosynthesis and specific functions unique for some organisms. Furthermore, phylogenetic profiles that contain approximately half the number of organisms or that are of intermediate to high total parsimony have been shown to contain most significant functional information.

Phylogenetic profiles have also been found to be useful in the detection of functionally complementary proteins, that have evolved through one of several mechanisms of convergent evolution, since these proteins in some cases show inverse phylogenetic profiles.

A prediction for the next years to come, is that even more new techniques will be introduced, and that these along with the other non-homology based methods will be integrated with homology- and structure derived protein functions, as well as with known experimental data.

6 Future improvements

There are some limitations in the current implementation of the PhylProM database. One major problem is that the all-against-all search using PSI-BLAST makes updating very time-consuming. One solution to this problem might be to match updated sets of fully sequenced genomes against a stable set of "seed" proteins, as is done in the database Pfam [Sonnhammer *et al.*, 1997, Sonnhammer *et al.*, 1998a, Bateman *et al.*, 1999]. Protein domains that do not match Pfam will still need to be matched against each other, but in this similarity search the largely reduced number of sequences makes updating less time-consuming.

In the matching of protein domains that do not match Pfam, an improvement would be to optimize the search parameter values for proteins, so that different parameter settings are used for membrane proteins and globular proteins. This would remove the problem of missing weak but real sequence similarities between globular proteins.

7 Acknowledgment

This work was supported by grants form the Swedish Foundation for Strategic Research, the Swedish Natural Sciences Research Council, and the Swedish Research Council for Engineering Sciences. Stockholm Bioinformatics Center is supported from the Swedish Foundation for Strategic Research.

References

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

[Bairoch & Apweiler, 1996] Bairoch, A. & Apweiler, R. (1996). The swiss-prot protein sequence data bank and its new supplement trembl. *Nucleic Acids Res.* **24**, 17–21.

[Bateman *et al.*, 1999] Bateman, A., Birney, E., Durbin, R., Eddy, S., Finn, R., & Sonnhammer, E. (1999). Pfam 3.1: 1313 multiple alignments and profile hmms match the majority of proteins. *Nucleic Acids Res,* **27** (1), 260–262.

[Benner *et al.*, 2000] Benner, S., Chamberlin, S., Liberles, D., Govindarajan, S., , & Knecht, L. (2000). Functional inferences from reconstructed evolutionary biology involving rectified databases- an evolutionarily grounded approach to functional genomics. *Res. Microbiology,* **151**, 97–106.

[Bork *et al.*, 1998] Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., & Yan, Y. (1998). Predicting function: From genes to genomes and back. *J.Mol.Biol.* **283**, 707–725.

[Bork & Eisenberg, 2000] Bork, P. & Eisenberg, D. (2000). Sequences and topology genome and proteome informatics. *Current Opinion in Structural Biology,* **10**, 341–342.

[Castresana & Moreira, 1999] Castresana, J. & Moreira, D. (1999). Respiratory chains in the last common ancestor of living organisms. *J. Mol. Evol,* **49**, 453–460.

[Doolittle, 1999] Doolittle, W. (1999). Phylogenetic classification and the universal tree. *Science,* **284**, 2124–2128.

[Enright *et al.*, 1999] Enright, A., Iliopoulos, I., Kyrpides, N., & Ouzounis, C. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature,* **402** (6757), 86–90.

[Fitch, 1971] Fitch, W. (1971). Defining the course of evolution: Miminum change for a specific tree topology. *Systematic Zoology,* **20**, 406–416.

[Galperin *et al.*, 1998] Galperin, M., Walker, R., & Koonin, E. (1998). Analogous enzymes: Independent inventions in enzyme evolution. *Genome Research,* **8**, 779–790.

[Heider *et al.*, 1999] Heider, J., Spormann, A., Beller, H., & Widdel, F. (1999). Anaerobic bacterial metabolism of hydrocarbons. *FEMS Microbiology Reviews,* **22**, 459–473.

[Hobohm *et al.*, 1992] Hobohm, U., Scharf, M., Schneider, R., & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.

[Huynen *et al.*, 2000] Huynen, M., Snel, B., Lathe III, W., & Bork, P. (2000). Exploitation of gene context. *Current Opinion in Structural Biology,* **10**, 366–370.

[Ito & Braithwaite, 1991] Ito, J. & Braithwaite, D. (1991). Compilation and alignment of DNA polymerase sequences. *Nucleic Acids Res.* **19** (15), 4045–4057.

[Ito & Braithwaite, 1993] Ito, J. & Braithwaite, D. (1993). Compilation, alignment and phylogenetic relationships of DNA polymerase sequences. *Nucleic Acids Res.* **21**, 787–802.

[Jain *et al.*, 1999] Jain, R., Rivera, M., & Lake, J. (1999). Horizontal gene transfer among genomes: The complexity hypothesis. *Proc Natl Acad Sci.* **96**, 3801–3806.

[Jones *et al.*, 1994] Jones, D., Taylor, W., & Thornton, J. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry,* **33**, 3038–3049.

[Kodra, 1998] Kodra, J. (1998). *Chemistry and Enzymology of an Expanded Genetic Alphabet.* PhD thesis Swiss Federal Institute of Technology.

[Koonin *et al.*, 1998] Koonin, E., Tatusov, R., & Galperin, M. (1998). Beyond complete genomes: from sequence to structure and function. *Current Opinions in Structural Biology,* **8**, 355–363.

[Leipe *et al.*, 1999] Leipe, D., Aravind, L., & Koonin, E. (1999). Did DNA replication evolve twice independently? *Nucleic Acids Res.* **27**, 3389–3401.

[Marcotte, 2000] Marcotte, E. (2000). Computational genetics: finding protein function by nonhomology methods. *Current Opinion in Structural Biology,* **10**, 359–365.

[Marcotte *et al.*, 1999a] Marcotte, E., Pellegrini, M., Ng, H.-L., Rice, D., Yeates, T., & D, E. (1999a). Detecting protein function and protein-protein interactions from genome sequences. *Science,* **285** (5428), 751–753.

[Marcotte *et al.*, 1999b] Marcotte, E., Pellegrini, M., Thompson, M., Yeates, T., & Eisenberg, D. (1999b). A combined algorithm for genome-wide prediction of protein function. *Nature,* **402** (6757), 83–86.

[Pellegrini *et al.*, 1999] Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D., & T.O., Y. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A,* **96** (8), 4285–4288.

[Sargent *et al.*, 1998] Sargent, F., Bogsch, E., Stanley, N., Wexler, M., Robinson, C., & Berks, B.C. amd Palmer, T. (1998). Overlapping functions of components of a bacterial sec-independent protein export pathway. *The EMBO Journal,* **17**, 3640–3650.

[Sonnhammer *et al.*, 1997] Sonnhammer, E., Eddy, S., & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins, Structure function and genetics,* **28**, 405–420.

[Sonnhammer *et al.*, 1998a] Sonnhammer, E., Eddy, S., E., B., Bateman, A., & Durbin, R. (1998a). Pfam: multiple sequence alignments and hmm-profiles of protein domains. *Nucleic Acids Res.* **26**, 322–325.

[Sonnhammer *et al.*, 1998b] Sonnhammer, E., G. von Heijne, G., & Krogh, A. (1998b). A hidden markow model for predicting transmembrane helices in protein sequences. *ISMB,* **6**, 175–182.

[Strimmer & Moulton, 2000] Strimmer, K. & Moulton, V. (2000). Likelihood analysis of phylogenetic networks using directed graphical models. *Molecular Biology and Evolution,* **17**, 875–881.

[Tatusov *et al.*, 1997] Tatusov, R., Koonin, E., & Lipman, D. (1997). A genomic perspective on protein families. *Science,* **5338**, 631–637.

[Tauer & Benner, 1997] Tauer, A. & Benner, S. (1997). The b12-dependent ribonucleotide reductase from the archaebacerium thermoplasma acidophila: An evolutionary solution to the ribonucleotide reductase conundrum. *Proc. Natl. Acad. Sci,* **94**, 53–58.

[Wallin & von Heijne, 1998] Wallin, E. & von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Science,* **7**, 1029–1038.

[Yona *et al.*, 1999] Yona, G., Linial, N., & Linial, M. (1999). Protmap:automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *PROTEINS: Structure, Function, and Genetics,* **37**, 360–378.