

# The Use of Phylogenetic Profiles for Gene Predictions

David A. Liberles, Anna Thorén, Gunnar von Heijne and Arne Elofsson\*

*Stockholm Bioinformatics Center, Stockholm University, SE-10691 Stockholm, Sweden*

**Abstract:** Determining gene functions from genomic sequences is a central goal of bioinformatics. Most purely computational approaches to this problem are based on the detection of genes with similar sequences. With the completion of fully sequenced genomes alternative approaches have become feasible. One such method is that of phylogenetic profiles. In this method a gene is described by its phylogenetic profile, i.e. a string that encodes the presence or absence of a homologous gene in other genomes. This string is then used to search for other genes with similar profiles. In this paper we briefly review the field as well as present an analysis on the performance of the method. We also discuss variations on this theme including inverse phylogenetic profiles and non-exact profiles using phylogenetic trees. In conclusion this indicates that phylogenetic profiles might be useful for some, but not all functional annotations. Functional annotation of genomes remains an important problem in genomics when no close homologs exist.

## INTRODUCTION

In the emerging era of functional genomics, a major goal is to attach 'functions' - taken in a wide sense- to as many genes as possible. Function can be defined in several different ways stemming from biochemical and genetic analysis of gene function. It might describe the actual chemical reaction catalyzed by an enzyme or the involvement of a gene in a particular cellular process. In the pre-genomic era, the only non-experimental method to achieve functional knowledge about a gene was by inference from the known function of a homologous gene.

One new approach, that of phylogenetic profiles, originated from a related line of biological research that seeks to understand the divergent phenotypes of organisms through a comparative genomic approach. The molecular basis of this is the set of genes and pathways present in different complete genomes. The intersection of the gene set of all possible genomes may define the minimum complement of genes necessary for cellular life [1]. Beyond that, additional genes allow for more complex lifestyles beyond the absolute minimum. The set of genes in any given genome reflects the lifestyle of the organism and differences in gene sets can drive differences in lifestyle mediated by differences in metabolism. Clusters of orthologous groups (COGs) of genes were established to trace these differences in lifestyle through genomic content. A similar idea was also proposed by Gaasterland *et al* [2]. From this, Eisenberg and coworkers [3] introduced phylogenetic profiles as a non-homology based method for functional predictions.

Another recent method that does not use homology for gene annotations is the "Rosetta Stone" method introduced by Eisenberg and coworkers [3], and Ouzounis and coworkers [4]. This method is based upon searching for gene

fusion or fission events when comparing genomes. Gene fission appears to be common, driven at least partially by the process of subfunctionalization [5,6]. The Rosetta Stone method is one of several methods introduced to go beyond sequence similarity when predicting the function of a gene [3,4,7,8]. However, the Phylogenetic Profile method seems to perform best as a standalone method [8].

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. Further how much information is obtained from the use of phylogenetic profiles has not been carefully studied. Here we try to review the current status of phylogenetic profiles and answer these questions. To facilitate the to phylogenetic profiles, we have created an easily search- and browse-able database (<http://www.sbc.su.se/PhylProM/>), that includes phylogenetic profiles based on 24 prokaryotic and 2 eukaryotic genomes.

In addition to identical phylogenetic profiles, we have also considered inverse profiles and closely related profiles. We have used two different methods to define closely related phylogenetic profiles. One is simply based on the number of organisms that differ between profiles, and the second is based on the differential parsimony between profiles. The PhylProM database has also been set up to allow searches using 'degenerate' phylogenetic profiles.

It is not likely that all phylogenetic profiles contain the same amount of information. For instance there are many genes that only exist in one organism or exist in all organisms. Genes with these profiles are not very likely to be functionally related. On the other hand, if a gene pair only exists in a few distantly related organisms with similar ecological and metabolic constraints, intuitively it is more likely that they are functionally linked. To explore this question, we have grouped the profiles based on the number of organisms in the profile and according to their total parsimony. In general, we find that profiles that contain more than a few organisms or that are of intermediate to high

\*Address correspondence to this author at the Stockholm Bioinformatics Center, Stockholm University, SE-10691 Stockholm, Sweden; Tel: +46-8-553-78568; Fax: +46-8-553-78214; E-mail: arne@sbcsu.se

total parsimony provide 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.

## METHODS

### Pairwise Comparisons

In order to detect sequence similarities, we performed an all-against-all search on these genes using PSI-BLAST [9]. A PSI-BLAST search result is not necessarily symmetric. Two genes might be identified when the search is performed starting with one gene, but not with the other gene as a query. Furthermore, even if two genes 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 we therefore replaced each case where the matches were not symmetric with the most significant (i.e. lowest) E-value.

The constructed phylogenetic profile represents for each gene the presence or absence of homologs in different organisms. This pattern of inheritance is based on the idea that phylogenetic profiles correlate between genes that are functionally related. In order to find genes with related functions we therefore grouped genes with identical profiles. A set of proteins sharing each phylogenetic profile is reported.

### Phylogenetic Reconstructions

The phylogenetic relationship of each species with a complete genome sequence can be downloaded from the NCBI web-site [10]. Then ancestral nodes can be reconstructed probabilistically using the Fitch algorithm [11] and tree similarity assessed subtractively along each branch. Total parsimony is defined by summing the total number of changes along each branch of the tree.

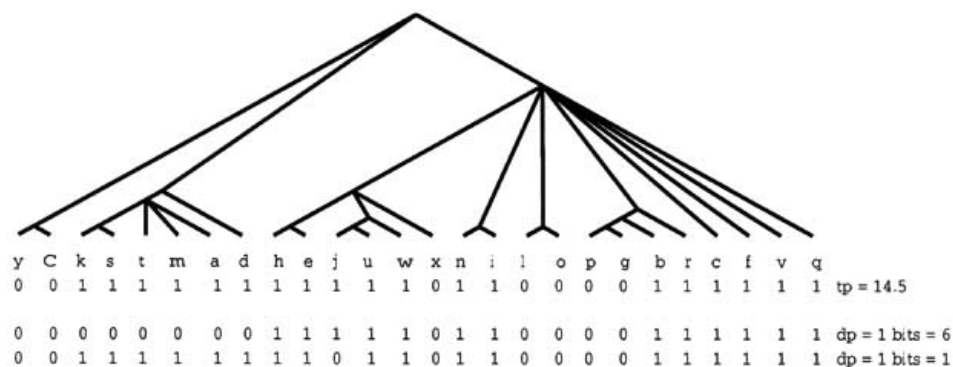
It should be noted that while a maximum likelihood approach can be envisioned for reconstructing gene loss and lateral transfer events, the low probability of these events for any specific gene along a given branch is the condition under which parsimony works best [12].

### Keyword Analysis

All studies of the performance of the phylogenetic profiles are based on the *E. coli* genome as it is one of the genomes with the largest fraction of the genes classified in SWISSPROT. We define two genes as being "linked" if they are predicted to be related. A link can occur either by two genes being detected by the PSI-BLAST search or by having similar phylogenetic profiles. To examine if two linked genes are functionally related it is necessary to have some knowledge of the function of the genes. We have used a simple approach to infer functional similarity by searching for similar keywords in SWISSPROT [13]. This measure is similar to the keyword recover rate as used in [8]. While keyword matches are not at all perfect for measuring performance, it is a commonly used method that at least can give an indication about the performance.

### Comparison with the Cellular Roles of Human Transcripts

In addition to analyzing the keyword recovery rate, we also compared genes found in the *C. elegans* genome with their neighbors. This was assessed using the Expressed Gene Anatomy Database (EGAD) (<http://www.tigr.org/tdb/egad/egad.html>), which contains cellular functions for human genes. To determine hits with human genes with known functions genes were searched against the EGAD database with BLAST using a threshold of  $10^{-15}$ . Analysis was performed as above when analyzing SWISSPROT keywords.



**Fig. (1).** Descriptive figure of bitwise difference and differential parsimony (dp). Here, the phylogenetic tree utilized is depicted with organism legend found on the webpage. A sample phylogenetic profile with total parsimony (tp) score of 14.5 is shown. Two sample profiles with differential parsimony of 1 are shown to illustrate the difference between the nonexact match methodologies. The first profile has a bitwise difference of 6 and would not be considered close using this method, while the second has a bitwise difference of 1 and would be considered close using both methods.

## RESULTS AND DISCUSSION

### The PhylProM Database

Based on an all-against-all PSI-BLAST search of all genes encoded in 24 prokaryotic and 2 eukaryotic genomes and topology prediction of the potential membrane proteins in the collection, the PhylProM database has been developed. PhylProM is available for review at <http://www.sbc.su.se/PhylProM/>. The database can be searched starting with a keyword, a protein or gene name, a protein sequence, or by specifying a phylogenetic profile (both exact and nonexact profiles are permissible). Profiles representing changes along a specific branch of the evolutionary tree can also be viewed from the web-site, enabling one to see which functions were co-evolving in specific species. This type of approach has been utilized by [14] to build whole genome trees based upon the gain or loss of orthologs along specific branches.

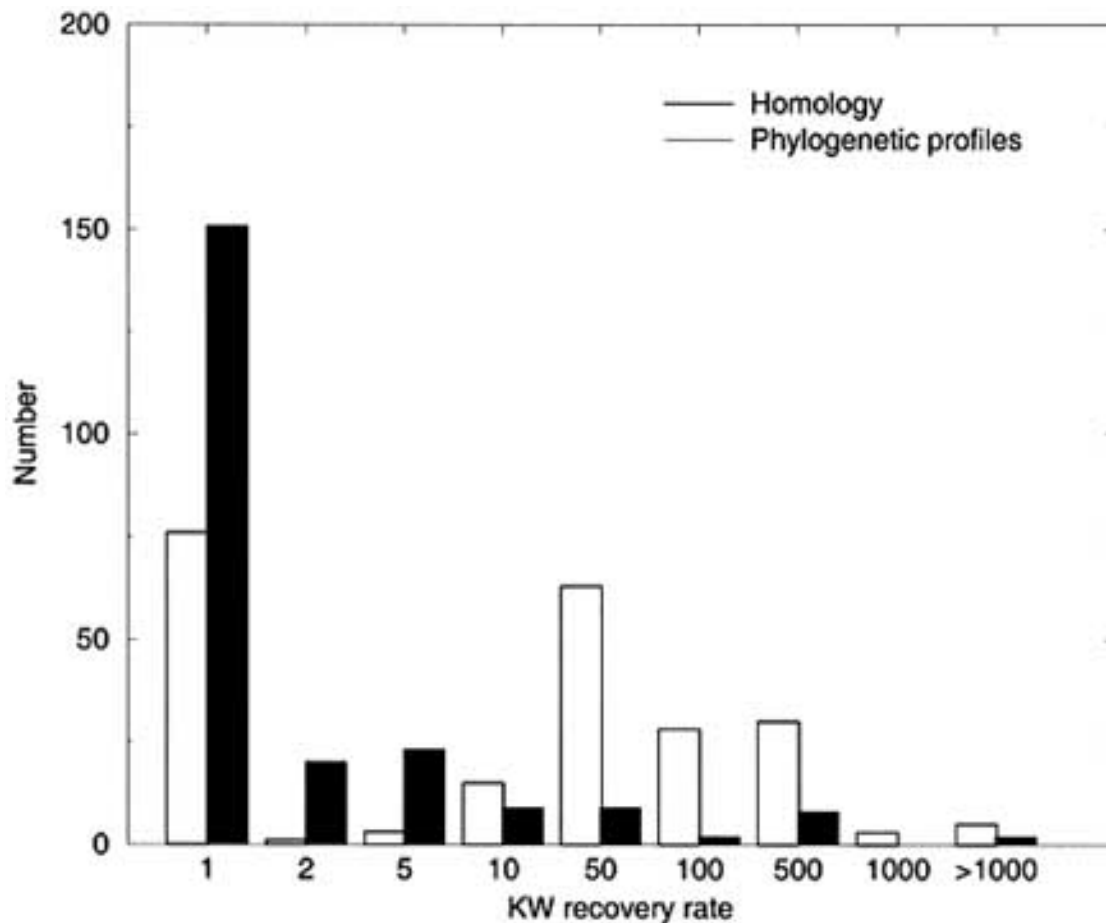
### Do Phylogenetic Profiles Contain Functional Information?

Using PhylProM, we have asked first, how the phylogenetic profiles compare to standard sequence-based searches, and, second, whether the predictive power of

phylogenetic profiles can be improved by basing the analysis not only on identical but also on 'similar' profiles.

The fraction of the neighbors varies greatly depending on the keyword used. The most successful example is found by using the keyword Flagell, where 66% of all genes with the phylogenetic profile "qvebujolw" share this keyword. When the profiles are not identical this number is significantly decreased.

In a recent study it was pointed out that the information in phylogenetic profiles depends on the function that is examined [3]. To identify the type of functions phylogenetic profiles can detect we studied the keyword recovery rate for all 291 keywords. In Fig. (2) the number of keywords with a particular keyword recovery rate is plotted against the recovery rate. For a few very specific functions, such as the existence of a flagella or not, a majority of the genes found with the same phylogenetic profiles are related and the recovery rate is very large. For a larger group of functional keywords the phylogenetic profiles increase the recovery rate, but still only a small fraction of all genes that have the same profile actually share the same function. For another large group, such as iron, there is no significant increase in the recovery rate. Using traditional homology searches, empty bars in Fig. (2), a similar trend is seen. However a



**Fig. (2).** The number of keywords is 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.

significantly larger proportion of the keywords have high kw-recovery rates.

The 30 keywords with a keyword recovery rate higher than five for the exact profiles are shown in Table 1. Interestingly many of these are involved in biosynthesis of small molecules. Several of the others are involved in

specific functions unique for some organisms, such as chemotaxis, cell wall or flagella. The best example from Table 1 is that of "Organic Radical". This refers to enzymes that enable bacteria to metabolize hydrocarbons, including alkylbenzenes, alkanes, and alkenes, using a radical intermediate [15]. The organisms that share this keyword are all capable of metabolizing hydrocarbons anaerobically, a

**Table 1. Keyword Recovery Rate for the Keywords with Most Information from Exact Phylogenetic Profiles. In the Exact PP Only Genes that Share Exactly the Same Phylogenetic Profiles are Included, While in the Non-Exact PP the Profiles Might Differ in up to Two Organisms. The Parsimony-PP Include all Genes with a Differential Parsimony of Less than 2**

Keyword	Family	Exact-PP	Non-Exact-PP	Parsimony-PP
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	1	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

trait that likely arose convergently under selective pressure [16].

A second comparison was performed with genes found in *C. elegans* and their neighbors using the Expressed Gene Anatomy Database (EGAD) for comparison. This database contains human sequences linked to known cellular functions. Some specialized functions showed increases in keyword identification when compared with random, but there is presently too little data to obtain any firm conclusions. The current dearth of metazoan and especially chordate genome sequences limits the utility of this database for performing this analysis. Given the greater degree of recombination of functional domains in polyploid eukaryotes, it will be interesting to test the utility of phylogenetic profiles in these species using a database like EGAD.

### Inverse Profiles

A second type of evolution, convergent evolution, is expected to be found in genes that perform redundant functions but lack a common evolutionary origin [17,18]. This will include genes that perform the same biological function, but are not homologs and may not even perform precisely the same chemistry. If such a function is present in all organisms, then these genes could possibly be detected using an inverse phylogenetic profile (i.e. a profile where ones have been replaced by zeros and vice versa).

Using inverse profiles, we find only three keywords - DNA-directed DNA polymerase (recovery rate=49), DNA repair (9), Isomerase (8) - with a recovery rates higher than two. 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 [19,20,21]. It has been suggested [22] that DNA replication has evolved at least twice independently and that the last common ancestor of Bacteria and Eukaryota/Archea had a genetic system that contained RNA and DNA, with the latter produced by reverse transcription only.

In a search for inverse profiles among the minimal gene set for cellular life, Koonin found 55 such examples [1]. These analyses demonstrate that examining inverse phylogenetic profiles can be useful for finding functionally complementary genes 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 in some cases here, or alternatively mutually exclusive functionalities.

### What Profiles are Most Informative?

It is likely that different phylogenetic profiles contain different amount of functional information. To detect the profiles that are most likely to contain functional information we have studied the keyword recovery rate for groups of profiles. In Fig. (3) the total parsimony for a certain profile is

plotted against the average increase in keyword recovery for all genes of this total parsimony. We also compare the recovery rate with the number of organisms in the profile.

As predicted, profiles with a low total parsimony or few organisms have a lower recovery rate, Fig. (3). It should be noted that the peak at 9 organism is largely due to a single profile, therefore it is not possible to examine if profiles with an intermediate number of organisms contain more information than profiles containing almost all organisms. However, it seems as though phylogenetic profiles that contain only a few organisms have a lower keyword recovery rate than other profiles.

### Examining Non-Identical Profiles

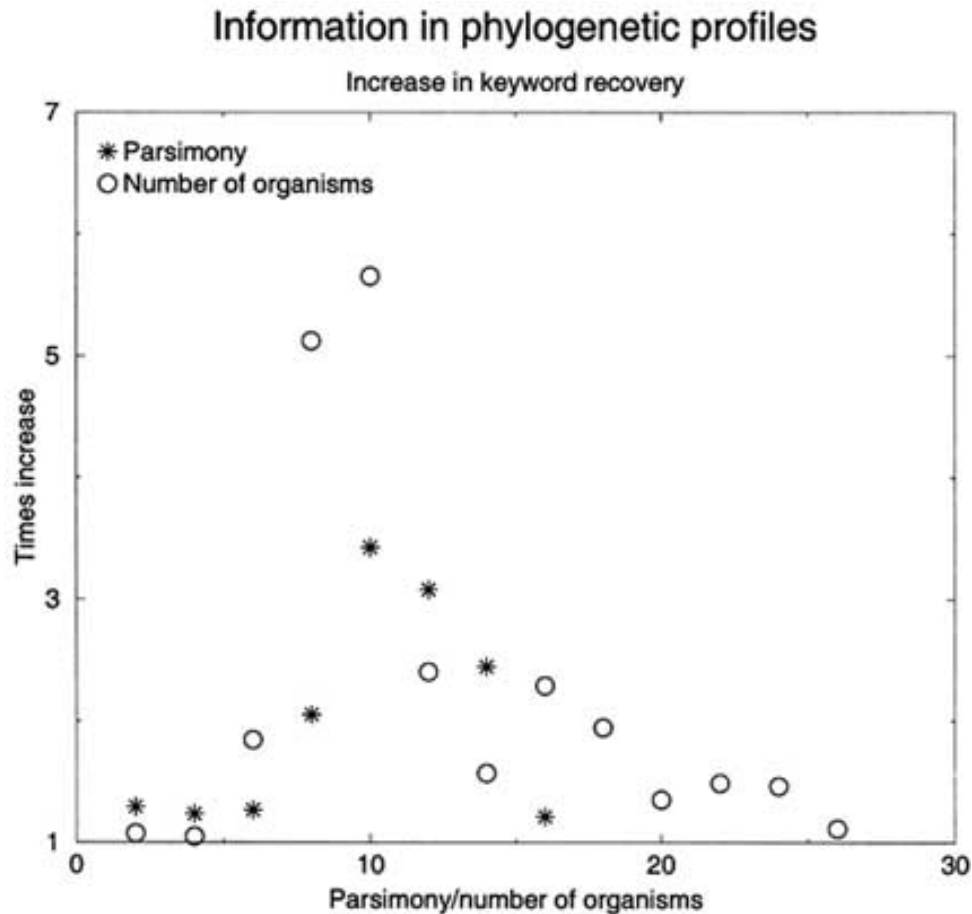
As discussed in the introduction, it is of interest to not only detect identical profiles, but also "almost" identical profiles. The similarity between two genes can be measured by counting the number of organisms that are present in one profile but not in the other. However, because all genomes evolved from common ancestral genomes at the root of the tree of life, genome deletion, insertion, or duplication events can be analyzed to take advantage of this information using phylogenetic trees. In addition to various mechanisms of convergent complementation (described above), some homologous genes will have evolved too far to be recognized by sequence comparison [23] and others that interact will likely be horizontally transferred together. Because these events occur at specific points in evolutionary history, it might be easiest to accurately detect them using a phylogenetic model that takes this history into account.

The differential parsimony values reported in Table 1 are higher in some but not all cases when compared with simple non-exact profiles. This might indicate that the most important factor for non-identical phylogenetic profiles are simple search misses.

Another potential problem for the phylogenetic profiles might be paralogs. However, many paralogs perform related functions that to the resolution of function described here are the same [24,25]. Another potential problem with differential parsimony is that of lateral gene transfer. While lateral gene transfer is extremely common in eubacterial and archaeobacterial species, especially among housekeeping genes [26,27], this will make individual trees less parsimonious, but should not affect the differential parsimony values if interacting partners are transferred together because lateral transfer events themselves can be reconstructed parsimoniously, like indel events. In the future, the development of branch length weighted parsimony approaches based upon the network of life [28] may become a better model for treating prokaryotic evolution and genome annotation.

### CONCLUSIONS

The method of phylogenetic profiling is useful for detecting some types of functional relationships. We show that the method is best applied to functions related to



**Fig. (3).** Plot of average increase in keyword recovery versus average total parsimony (stars) and number of organisms in the profile (circles). The entries are smoothed by averaging over three bins of size 2.

biosynthesis and organism specific functions. Profiles that contain more than a few organisms contain most information. The method does not appear to be useful for blind annotation of genomes, but is a valuable starting point for applying knowledge of mechanisms of genome evolution to the understanding of genome functional content and to genome annotation. In conjunction with this, a more rigorous scheme to annotate genomes, such as the gene ontology project, will help with the future evaluation of genome annotation methodologies [29]. Ultimately, new approaches based upon what we have learned coupled with homology dependent methods will allow us to better understand genome contents in the future.

A PhylProM is publicly available at <http://www.sbc.su.se/PhylProM/>

#### ACKNOWLEDGEMENT

This work was supported by grants from 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. We thank Dr. Jens Lagergren for the discussions about phylogeny. We also thank Malin Almgren for assisting with some BLAST searches.

#### REFERENCES

- [1] Koonin, E.V. How many genes can make a cell, The minimal-gene-set concept. (2000) *Annu. Rev. Genomics Hum. Genet.*, **1**, 99-116.
- [2] Gaasterland, T. and Ragan, M.A. Microbial genescapes, phyletic and functional patterns of orf distribution among prokaryotes. (1998) *Microb. Comp. Genomics*, **3**(4), 199-217.
- [3] Pellegrini, M.; Marcotte, E.M.; Thompson, M.J.; Eisenberg, D. and Yeates, T.O. Assigning protein functions by comparative genome analysis, protein phylogenetic profiles. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 4285-4288.
- [4] Enright, A.J.; Iliopoulos, I.; Kyrpides, N.C. and Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. (1999) *Nature*, **402**(6757), 86-90.

- [5] Force, M.; Lynch, F.B.; Pickett, A.; Amores, Yan, Y. and Postlethwait, J. Preservation of duplicate genes by complementary degenerative mutations. (1999) *Genetics*, **151**, 1531-1545.
- [6] Lynch, M. and Force, A. The probability of duplicate gene preservation by subfunctionalization. (2000) *Genetics*, **154**, 459-473.
- [7] Marcotte, E.M.; Pellegrini, M.; Ng, H.L.; Rice, D.W.; Yeates, T.O. and Eisenberg, D. Detecting protein function and protein-protein interactions from genome sequences. (1999) *Science*, **285**(5428), 751-753.
- [8] Marcotte, E.M.; Pellegrini, M.; Thompson, M.J.; Yeates, T.O. and Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. (1999) *Nature*, **402**(6757), 83-86.
- [9] Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J. Zhang, Z. Miller, W. and Lipman, D. J. Gapped BLAST and PSI-BLAST, a new generation of protein database search programs. (1997) *Nucleic Acids Res.*, **25**, 3389-3402.
- [10] Federhen, S.; Harrington, F.A.; Harrison, I.; CarolHotto, D.L. and Soussov, V. The NCBI taxonomy homepage. (2000) [http, //www.ncbi.nlm.nih.gov/ Taxonomy/taxonomyhome.html](http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html).
- [11] Fitch, W.M. Defining the course of evolution, Minimum change for a specific tree topology. (1971) *Systematic Zoology*, **20**, 406-416.
- [12] Schuler, D.; Price, T.; Mooers, A.O. and Ludwig, D. Likelihood of ancestor states in adaptive radiation. (1997) *Evolution*, **51**, 1699-1711.
- [13] Bairoch, A. and Apweiler, R. The SWISS-PROT protein sequence data bank and its new supplement TrEMBL. (1996) *Nucleic Acids Res.*, **24**, 17-21.
- [14] Lin, J. and Gerstein, M. Whole-genome trees based on the occurrence of folds and orthologs, Implications for comparing genomes on different levels. (2000) *Genome Research*, **10**, 808-818.
- [15] Heider, J.; Spormann, A.M.; Beller, H.R. and Widdel, F. Anaerobic bacterial metabolism of hydrocarbons. (1999) *FEMS Microbiology Reviews*, **22**, 459-473.
- [16] Castresana, J. and Moreiram, D. Respiratory chains in the last common ancestor of living organisms. (1999) *J. Mol. Evol.*, **49**, 453-460.
- [17] Blow, D. Enzymology. more of the catalytic triad. (1990) *Nature*, **343**, 694-695.
- [18] Galperin, M.Y.; Walker, D.R. and Koonin, E.V. Analogous enzymes, Independent inventions in enzyme evolution. (1998) *Genome Research*, **8**, 779-790.
- [19] Ito, J. and Braithwaite, D.K. Compilation and alignment of DNA polymerase sequences. (1991) *Nucleic Acids Res.*, **19**(15), 4045-4057.
- [20] Ito, J. and Braithwaite, D.K. Compilation, alignment and phylogenetic relationships of DNA polymerase sequences. (1993) *Nucleic Acids Res.*, **21**, 787-802.
- [21] Kodra, J.T. (1998) *Chemistry and Enzymology of an Expanded Genetic Alphabet*. PhD thesis, Swiss Federal Institute of Technology.
- [22] Leipe, D.D.; Aravind, L. and Koonin, E.V. Did DNA replication evolve twice independently? (1999) *Nucleic Acids Res.*, **27**, 3389-3401.
- [23] Tauer and Benner, S.A. The B12-dependent ribonucleotide reductase from the archaeobacterium thermoplasma acidophila, An evolutionary solution to the ribonucleotide reductase conundrum. (1997) *Proc. Natl. Acad. Sci.*, **94**, 53-58.
- [24] Benner, S.A.; Chamberlin, S.G.; Liberles, D.A.; Govindarajan, S. and Knecht, L. Functional inferences from reconstructed evolutionary biology involving rectified databases- an evolutionarily grounded approach to functional genomics. (2000) *Res. Microbiology*, **151**, 97-106.
- [25] Liberles, D.A.; Schreiber, D.R.; Govindarajan, S.; Chamberlin, S.G. and Benner, S.A. The adaptive evolution database (TAED). (2001) *Genome Biology*, **2**(8), Research 0028.
- [26] Doolittle, W.F. Phylogenetic classification and the universal tree. *Science*, (1999) **284**, 2124-2128.
- [27] Jain, R.; Rivera, M.C. and Lake, J.A. Horizontal gene transfer among genomes, The complexity hypothesis. (1999) *Proc Natl Acad Sci.*, **96**, 3801-3806.
- [28] Strimmer, K. and Moulton, V. Likelihood analysis of phylogenetic networks using directed graphical models. (2000) *Molecular Biology and Evolution*, **17**, 875-881.
- [29] Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; Harris, M.A.; Hill, D.P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J.C.; Richardson J.E., Ringwald, M.; Rubin, G.M. and Sherlock, G. Gene ontology, tool for the unification of biology. (2000) *Nature Genetics*, **1**, 25-29.