

Evolution of proteins and proteomes: a phylogenetics approach

Toni Gabaldón

Bioinformatics Department, Centro de Investigación Príncipe Felipe

Abstract: The study of evolutionary relationships among protein sequences was one of the first applications of bioinformatics. Since then, and accompanying the wealth of biological data produced by genome sequencing and other high-throughput techniques, the use of bioinformatics in general and phylogenetics in particular has been gaining ground in the study of protein and proteome evolution. Nowadays, the use of phylogenetics is instrumental not only to infer the evolutionary relationships among species and their genome sequences, but also to reconstruct ancestral states of proteins and proteomes and hence trace the paths followed by evolution. Here I survey recent progress in the elucidation of mechanisms of protein and proteome evolution in which phylogenetics has played a determinant role.

Keywords: protein evolution, proteome evolution, protein domains, phylogenetics, ancestral sequence reconstruction, function prediction

Introduction

During the 1960s, a decade before DNA sequencing become feasible, biochemists and molecular biologists were increasingly attracted by questions regarding the evolution of proteins. By that time, protein sequencing techniques were producing a growing number of sequences (Eck and Dayhoff 1966), and soon it was realized that proteins and nucleic acids could be used to document the history of past evolutionary events (Zuckerkindl and Pauling 1965). Computers, which had already been used in the sequence determination of proteins (Dayhoff 1965), were also recruited to the task of comparing sequences from different organisms. In their pioneering work, Fitch and Margoliash (1967) computationally compared sequences of the respiratory pigment cytochrome c from different organisms to assess their phylogenetic relationships. Since then, bioinformatics has been increasingly involved in protein evolution (Hagen 2000). Bioinformatics applications to the study of protein evolution include, among many others, algorithms to efficiently align similar sequences, to detect homologous sequences in large databases or to reconstruct phylogenetic trees from a given set of sequences.

The relationship between bioinformatics and the study of protein evolution was further strengthened with the advent of large-scale sequencing projects. The growing number of sequences stored in the databases, including those of complete genomes, provided a completely new dimension to the study of protein evolution: that of the evolution of complete proteomes. New bioinformatics tools were developed that allowed the comparison of complete genomes, the efficient detection of orthology relationships and the reconstruction of the evolution of complete proteomes. Almost four decades after the first computer-aided evolutionary analysis of proteins, there is a vast scientific literature reporting bioinformatics analyses that focus on protein or proteome evolution. The aim of this review is to provide a general overview of recent bioinformatics approaches to the study of the evolution of proteins and proteomes that involve the use of phylogenetics. Due to space limitations, I will focus on the evolution of proteins at the sequence level and of proteomes in terms of their protein content. Therefore, I purposely excluded considerations regarding evolution of protein structures, protein interaction or regulatory networks. I start providing an overview of the different bioinformatics applications that can be used to reconstruct the phylogenetic relationships of a protein family to then describe its applications. These include the reconstruction of ancestral protein sequences, the determination of orthology and paralogy relationships and the use of phylogenetic profiles and co-evolution to predict protein function. Subsequently, the evolution of proteins is considered within the context of the pathways and the complete proteomes in which they function. To illustrate the different sections, some specific examples from the literature have been selected.

Correspondence: Autopista del Saler 16, 46013 Valencia, Spain. E-mail: tgabaldon@cipf.es, Telephone: +34 96 328 96 80, Fax: +34 96 328 97 01

Phylogenetics and the study of protein evolution

The phylogenetic analysis of a protein (Figure 1) starts with the detection of other members of its family. This is usually done by comparing the sequence of the protein of interest with other sequences stored in the databases and, subsequently, selecting the hits that are significantly similar. The assumption is that proteins with similar sequences are derived from a common ancestral protein. In other words, they are considered to be homologous proteins (Fitch 1970, 2000). Several algorithms have been developed that allow efficient automatic detection of homologous proteins in large databases. These include pair-wise comparison algorithms like

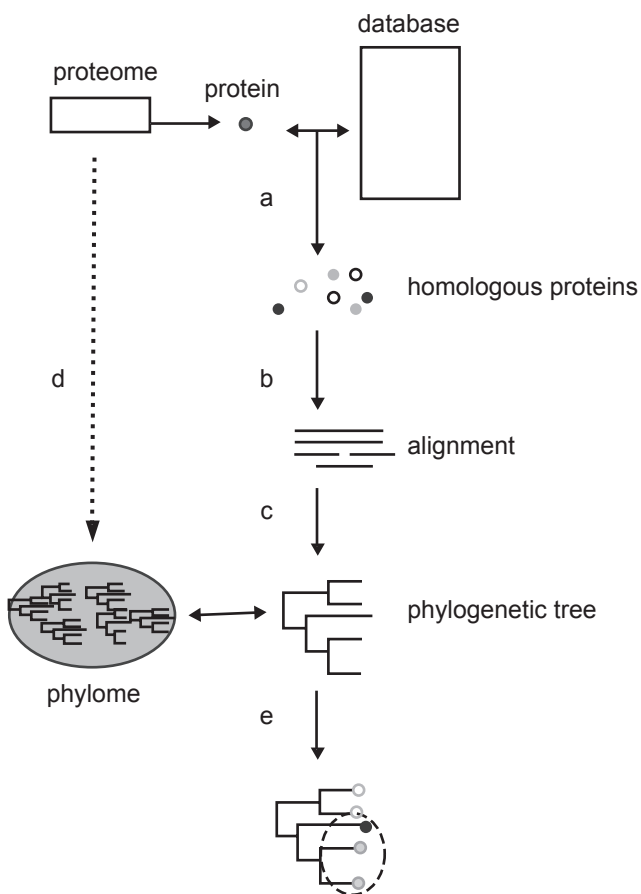


Figure 1: Schematic representation of the phylogenetic analysis process of a protein family. The protein sequence of interest is compared to a sequence database to retrieve significantly similar proteins (a); homologous proteins are aligned to place homologous residues on top of each other (b); under the assumption of an evolutionary model, a phylogenetic tree representing the evolutionary relationships among the protein sequences is reconstructed (c); if this process (a to c) is repeated over all proteins encoded by a genome, the total set of phylogenetic trees or phylome is reconstructed (d); the topology of the phylogenetic tree can be subsequently analyzed for different purposes, e.g. to determine orthology and paralogy relationships (e).

Smith-Waterman (Smith and Waterman 1981) and its faster approximation BLAST (Altschul, et al 1997). The sensitivity of such homology searches has been more than doubled (Park, et al 1998) by profile-based methods such as PSI-BLAST (Altschul and Koonin 1998, Altschul, et al 1997) or hidden markov models (Baldi, et al 1994, Eddy 1998). Once the sequences of a protein family are retrieved, they can be aligned. The alignment of multiple sequences basically aims to place ‘homologous’ residues of different proteins on top of each other. This constitutes a crucial step for the study of the evolution of proteins because it is assumed that all positions in a column of a multiple sequence alignment derive from a common ancestral residue. Several multiple-sequence alignment algorithms exist that combine speed with reasonable accuracy, these include programs such as ClustalW (Thompson, et al 1994), T-Coffee (Notredame, et al 2000) or MUSCLE (Edgar 2004).

By applying a specific evolutionary model to explain the amino acid substitutions observed in the multiple sequence alignment, the evolutionary distances between all pairs of proteins can be computed. This evolutionary distance, which reflects the expected mean number of changes per site that have occurred since two sequences diverged from their common ancestor, is used by the so-called distance-methods for phylogenetic inference. One such method is Neighbor Joining (NJ). NJ constitutes a good and fast heuristic algorithm that estimates the “minimal evolution” tree, a phylogenetic tree which minimizes the sum of the lengths (evolutionary distances) of all its branches (Saitou 1987). NJ and variations of it have long been proven to be quite efficient in finding the “right” tree topologies given a set of homologous sequences (Kuhner and Felsenstein 1994, Takahashi and Nei 2000), although its accuracy may suffer in large datasets (Nakhleh, et al 2002). Compared to other methods, NJ has the advantage of being very fast, which allows the construction of large trees including hundreds of sequences. Therefore, it is usually the method of choice when doing large-scale phylogenetic approaches.

A different approach for phylogenetic inference is that of Maximum Likelihood (ML) (Schadt, et al 1998). Here, the concept of likelihood refers to the probability that a certain tree with a set of parameters (topology, branch-lengths, etc) produces,

assuming a specific evolutionary model, a given set of data (sequences). ML-methods try to find the tree with the maximal likelihood to produce the variation observed in the given set of data. However, computing the likelihood of all possible trees for a decent number of proteins is a very computationally intense task and becomes unfeasible for large sets of sequences. Therefore, all practical methods rely on heuristics that are able to reduce the search-space and find good sub-optimal trees in a reasonable time. For instance PhyML (Guindon and Gascuel 2003) uses a simple hillclimbing algorithm to optimize a seed NJ-tree whereas MrBayes (Ronquist and Huelsenbeck 2003) uses bayesian inference with a Markov Chain Monte Carlo (MCMC) algorithm. Recently, some ML methods have been developed that allow the joint iterative reconstruction of the protein alignment and the corresponding phylogenetic tree (Redelings and Suchard 2005). The phylogenetic reconstruction with such methods is usually improved by the implementation of models of sequence evolution that allow the substitution rates to vary among the different positions (Felsenstein 2001). These models, which generally approximate a Gamma distribution for the variation of rates across positions, better reflect the real situation in which functional constraints are not uniform over the entire protein sequence. Even the substitution rate for a given residue may vary over time, a process known as heterotachy, which is taken into account by recent implementations (López et al, 2002).

Yet another type of phylogenetic inference is that of Maximum Parsimony (MP) (Felsenstein 1996), which selects the tree that requires the minimum number of character changes (mutations) to explain the given set of sequences. MP approach does not allow the correction for multiple mutations per site and it is more prone to the so-called long-branch attraction effect (placing long branches preferentially together towards the root of the tree).

Perhaps, the development of algorithms for phylogenetic reconstruction, database search and multiple sequence alignment represents the most visible contribution of bioinformatics to the study of protein evolution. This is not surprising because the applications of phylogenetic trees are many and they are used in diverse fields. Besides their traditional taxonomic use

for inferring the evolutionary relationship between organisms, phylogenetic trees can be used to establish orthology and paralogy relationships among proteins (Fitch 2000), to detect horizontal gene transfers (HGT) (Bapteste, et al 2004), gene and genome duplications (Van de Peer 2004), positively selected residues (Bielawski and Yang 2004), to define strains in epidemiologic studies, to predict functional interactions among coevolving genes (see below) or to estimate model parameters and substitution rates (Lio and Goldman 1998). Phylogenetics is thus central for many evolutionary analyses, some applications which include a phylogenetic reconstruction at some stage will be shown in more detail in the following sections.

Orthology and paralogy considerations

Orthology and paralogy are key concepts in the field of protein and proteome evolution. The use of these terms has been extended as comparative and evolutionary genomics have penetrated other fields. However, there is still some confusion about their exact meanings. Many researchers believe that orthologous proteins are simply proteins with the same function in different organisms, whereas paralogs are simply homologs within one organism. These definitions do not agree with the original given by Walter Fitch (1970), in which orthologs are homologous genes (proteins) derived by speciation from a common ancestor whereas paralogs are homologs derived by duplication. Therefore the definition of orthology and paralogy is strictly phylogenetic and do not include any functional consideration. The correct detection of orthology relationships allows the comparison of genomes in terms of their gene content, an essential step for studying the co-evolution of proteins and the evolution of complete proteomes (see below).

The detection of orthology relationships is ideally performed by detecting speciation and duplication events through phylogenetic analysis (Figure 2). Alternative methods which only rely on sequence similarity levels like “best bi-directional hits” (Huynen and Bork 1998) and its multiple-genome extensions (Tatusov, et al 2001), are more prone to errors, especially when there is variation in the rate of sequence evolution

within an orthologous group (Eisen and Fraser 2003). Recent developments in the automatic detection of orthology from phylogenetic trees (Dufayard, et al 2005, Gabaldón and Huynen 2005) are promising and allow its application over large datasets.

Protein domains as evolutionary units

We have seen how phylogenetic reconstruction can ascertain the evolutionary relationships within members of a protein family, which evolved independently after speciation and duplication events. A complication may arise when different protein families combine through recombination events. In this case, the evolutionary unit is not the full protein anymore but a smaller, discrete molecular entity called protein domain (Doolittle 1995).

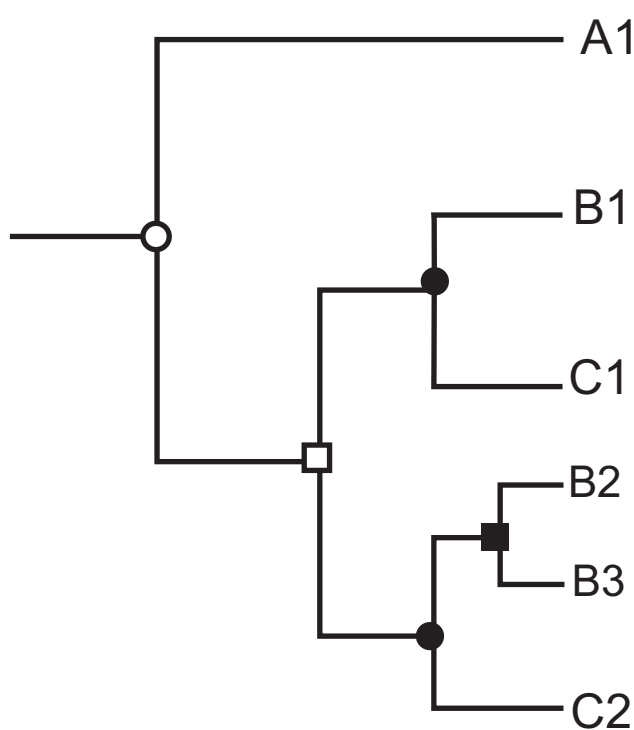


Figure 2: Orthology and paralogy relationships within a protein family. The phylogenetic tree of a hypothetical protein family comprising a total of six members: one in species A (A1), three in species B (B1, B2 and B3) and two in species C (C1, C2). Speciation and duplication events are represented as circles and squares, respectively. A first duplication event (white square) occurred before the B-C speciation (black circle) while a later one (black square) occurred within the B lineage. In this scenario the only protein present in A (A1) is orthologous to all the others and vice versa. When comparing members from B and C: C1 is orthologous to B1, while C2 has two orthologs in B (B2 and B3).

One of the first multi-domain proteins to be studied in detail was the tissue plasminogen activator (TPA), which contains four different domains that are also present in other protein families (figure 3). The fact that different parts of TPA presented homology with different protein families is an indication that TPA is actually an evolutionary chimera. Thus, protein domains of TPA are the evolutionary units that underwent duplication, followed by recombination with other domains, and in present-day proteins are found as units of various multi-domain arrangements. Initially, multi-domain proteins were considered an exception to the rule but, as more sequences became available in the databases more cases of mosaic proteins and different domain combinations were identified. With the current data it is estimated that multi-domain proteins comprise more than two-thirds of the proteins encoded by prokaryotic genomes (Teichmann, et al 1998), and even a greater fraction in eukaryotic genomes (Gerstein 1998).

Multi-domain proteins are created through gene duplication and recombination events. Since the role of protein domains is dependent on the context in which they are found, the emergence of new domain combinations may involve the creation of completely new functions. Not surprisingly, this mechanism of domain shuffling has been extensively exploited during evolution. This constitutes another example of how completely new functions can emerge from the tinkering of preexisting components.

Several large-scale studies have recently focused on the evolution of proteomes in terms of the domain combinations that they contain. One of such studies shows that the repertoire of domain combinations observed in nature is just a small fraction of all possible combinations (Apic, et al 2001). This suggests that domain combinations are subjected to strong selection during evolution. While most of the protein domains can be found in combination with just one or two different domains, some others, the so called promiscuous domains, can be combined with many different domains (Wuchty 2001). Interestingly, most protein domains are present in eukaryotes, bacteria and archaea. This suggests a very ancient origin for all of them and a last common universal ancestor possessing an almost complete protein domain repertoire. In contrast, most domain

combinations are kingdom- or lineage-specific and have, therefore, appeared in later stages during evolution.

Ancestral sequence reconstruction

In the phylogenetic tree, the internal nodes represent ancestral sequences from which proteins at the leaves have evolved. The properties of these sequences are often relevant to ascertain how modern functions come about or what mutations were crucial in the development of functional specificities within a protein family. Infer the properties of such ancestral sequences or even to reconstruct them is the aim of the emerging field of “ancestral sequence reconstruction”. Ancestral sequence reconstruction uses extant sequences and the phylogenetic relationships among them to infer the most plausible ancestral sequences (Cai, et al 2004, Yang, et al 1995). Attending their scope, ancestral reconstruction can be divided into joint reconstruction (Pupko, et al 2000), when it finds the most likely set of amino acids for all internal nodes at a site, or marginal reconstruction (Koshi and Goldstein 1996), when it

limits the inference to a particular node or sets of nodes.

Once the sequence of interest is computationally reconstructed, it is possible to infer its functional properties based on the presence of specific residues at key positions, e.g. the active site. Additionally, one can take one step further and synthesize the ancestral protein to directly test its biochemical properties, an experiment that is often referred to as “ancestral sequence resurrection” (Thornton 2004). An interesting application of ancestral sequence reconstruction is the testing of specific scenarios that involve extinct organisms. For instance, knowledge about the light-response properties of the visual protein rhodopsin from early dinosaurs is useful in the assessment of their day and night habits. Using sequences from living vertebrates, Chang and colleagues reconstructed, and subsequently synthesized, the ancestral archosaurs rhodopsin (Chang, et al 2002). Its biochemical characterization suggested that these early dinosaurs could have seen well under dim lighting conditions. A few more examples include the

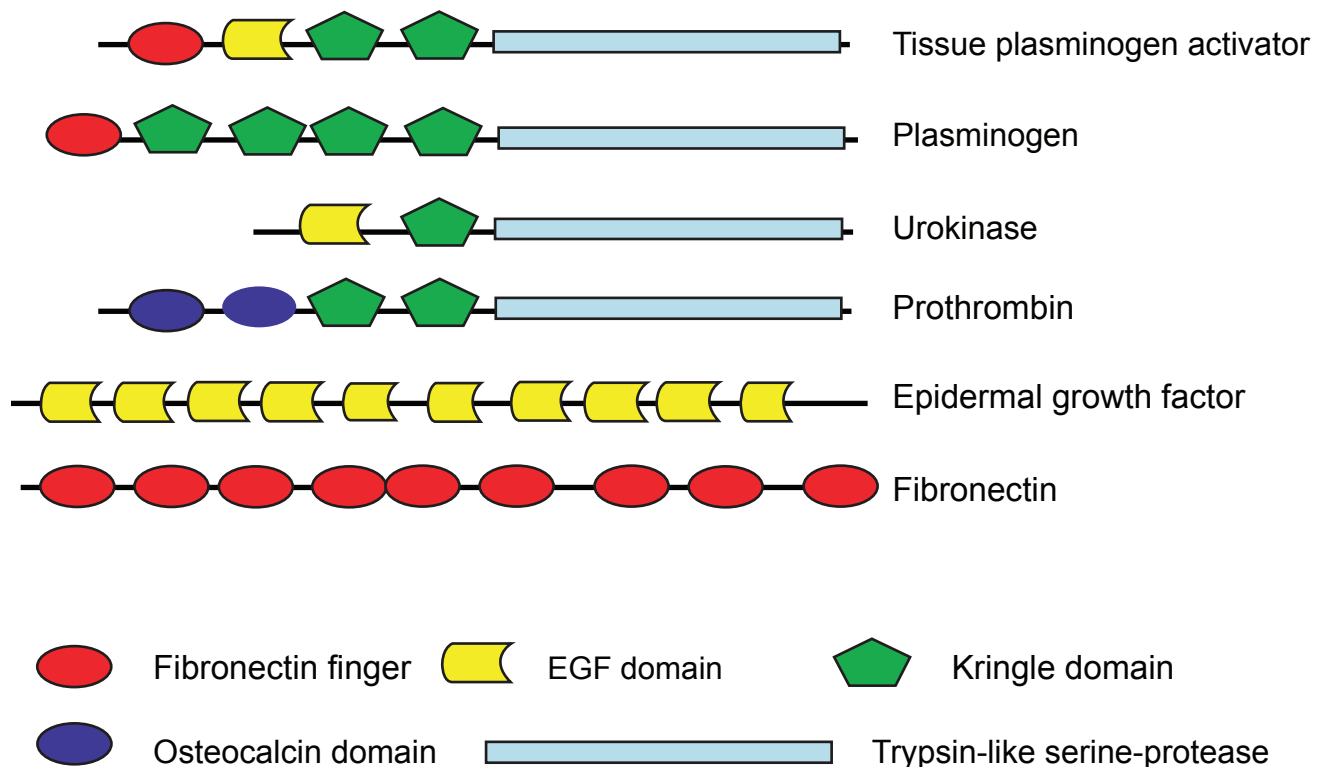


Figure 3: Schematic domain organization of tissue plasminogen activator protein (TPA) and several other proteins that share, at least, one of the domains present in TPA. The specific function of each protein emerges from a particular domain combination.

reconstruction of ancestral hormone receptors (Thornton, et al 2003) and ancestral bacterial translational elongation (EF-Tu) proteins (Gaucher, et al 2003).

But how reliable are these reconstructions and therefore the functional inferences? This is a hotly debated issue and some researchers remain sceptic over conclusions on extinct sequences. In general, ancestral sequence reconstruction suffers from the same weaknesses as other evolutionary methods: its correctness depends on the quality of the data and the adequateness of the model. Moreover, these methods are based on probabilistic approaches and thus the reconstructed sequences are not free of ambiguities. With no real ancestral sequences at hand it is hard to judge the correctness of the reconstructions. Recently, some analyses have suggested that ancestral reconstruction may indeed have a

sequence-compositional bias that can affect inference of ancestral function (Krishnan, et al 2004). A reasonable solution to these caveats is to base the functional inference not on a single reconstructed protein but on a sample of possible ancestral proteins, which takes into account ambiguously reconstructed positions. Nevertheless, the question of how many ancestral sequence samples are necessary to reliably estimate ancestral function remains open.

Use of phylogenetic profiles to infer function

One of the most powerful techniques for assigning a biological function to a protein sequence is the detection of homologous sequences with known function in a sequence-similarity search. When performing this transfer of knowledge between similar sequences we are using an evolutionary

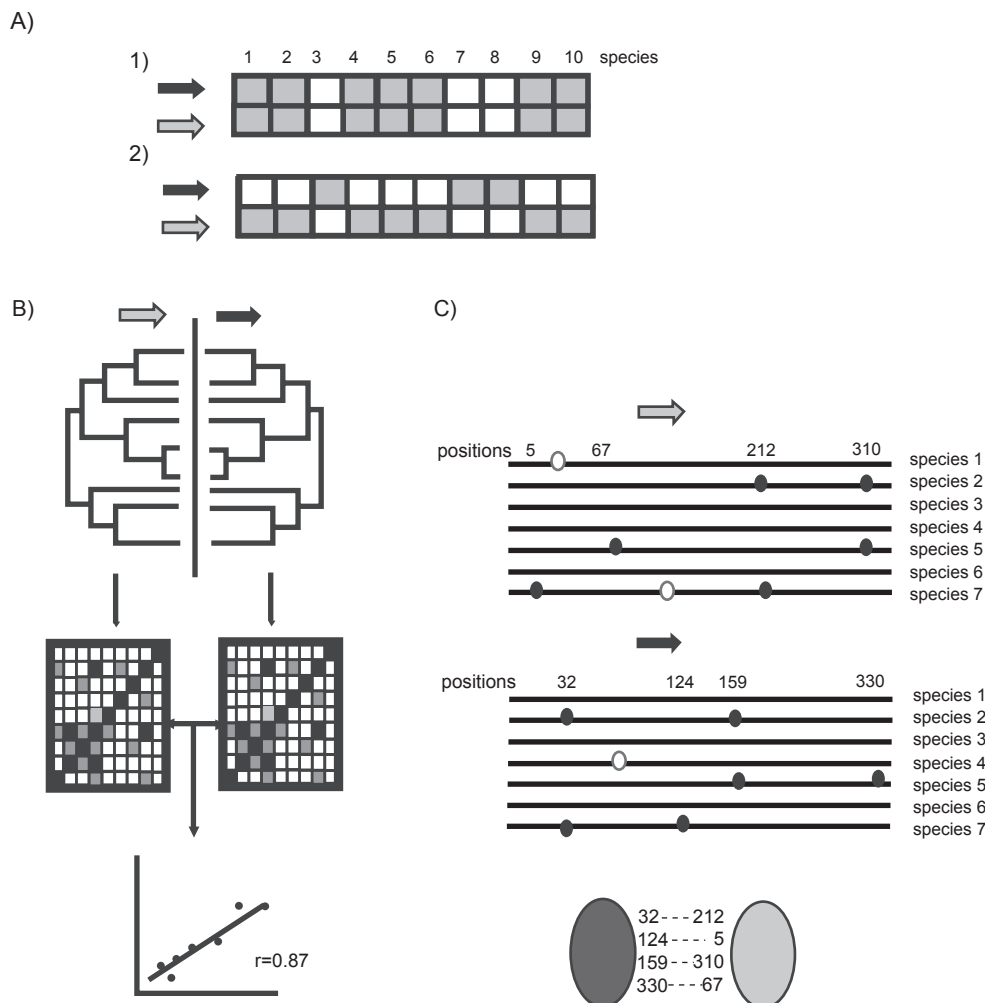


Figure 4: Co-evolution as a tool for functional inference. Several methods that use co-evolution to predict functional interactions among two protein families (grey and black arrows) are illustrated. A) A similar pattern (1) of presence (grey boxes) and absence (white boxes) in a set of genomes might indicate that both proteins function in the same biological process or pathway. Complementary phylogenetic profiles (2) suggest a nonorthologous gene displacement event, i.e. both proteins perform the same function and can functionally replace each other. B) Similarity of phylogenetic trees, which can be measured by the correlation of the distance matrices as shown in the figure, are indicative of similar evolutionary constraints for both protein families. This, in turn, suggests that the two proteins interact physically. C) Correlated mutations (black circles), those that occur in both proteins in the same set of species, can be detected by the comparison of the protein alignments from the two protein families. The method predicts not only that the two proteins interact (ovals) but also which are the interacting residues (numbered residues).

approach: we assume that the sequences share a common ancestor and that their function has been maintained during evolution. However, homology-search is not the only way in which the evolutionary analysis of a protein can serve to infer its functional role.

In recent years, as part of the so-called genome-context analyses (Gabaldón and Huynen 2004), several methods have been developed that exploit the co-evolution of protein families to infer a functional interaction between them. One such technique, called gene co-occurrence or phylogenetic profiles (Huynen, et al 1998, Pellegrini, et al 1999), compares the patterns of presence/absence of proteins in a set of complete genomes and predicts functional interactions between proteins with similar profiles (figure 4-A). This method is based on the observation that proteins with a similar distribution across species have a high tendency to functionally interact (Gaasterland and Ragan 1998, Huynen and Bork 1998, Huynen and Snel 2000, Pellegrini, et al 1999). It must be noted, however, that the reverse assumption, that functionally interacting proteins have a similar distribution across species, is not necessarily true as we will see in the following section. Moreover, the detection of proteins with complementary phylogenetic patterns might indicate a non-orthologous gene displacement and thus a similar function for both protein families (Galperin and Koonin 2000).

Another variant of the use of co-evolution to predict protein function uses the evolutionary information that is contained in the sequences (figure 4-B). For specific protein families that are known to physically interact, such as the chemokine receptor system (Goh, et al 2000, Hughes and Yeager 1999), it was shown that their phylogenies are more similar to each other than expected considering the evolutionary divergence between the species. This suggested the existence of correlated evolution reflecting similar evolutionary constraints. Some authors (Pazos and Valencia 2001, Valencia and Pazos 2003) have applied this property to predict interaction partners in *E. coli* by detecting significantly correlated positions between the distance matrices used to build the phylogenetic trees. Others (Ramani and Marcotte 2003) used a similar approach to predict the binding specificities among members of 18 ligand and receptor families with many members in the

human genome. Finally, the co-evolution of interacting proteins can be analyzed more closely by searching for amino acid substitutions that are correlated in both protein families (they occur in the same species) (figure 4-C). These positions may correspond to residues on the interface that undergo mutations in one protein to compensate the effects of mutations in the other (Dimmic et al 2005, Fukami-Kobayashi et al 2002, Pazos, et al 1997, Pazos and Valencia 2002). This method has the advantage of predicting not only the interacting proteins, but also the residues potentially involved in the interaction. With more fully sequenced genomes to come, the accuracy and coverage of these genome-context techniques can only improve. It is also expected that new discoveries of mechanisms of protein and proteome evolution will fuel the development of new techniques that exploit them to predict protein function.

Phylogenetic diversity of protein complexes and pathways

Proteins do not work as isolated entities. Instead, they perform their function through interactions with other proteins, as part of pathways, complexes and other types of functional modules. Therefore, to fully understand the evolution of a protein it is necessary to consider it in the context of the evolution of its functional partners. Recently, the availability of fully sequenced genomes has enabled the comparison across species of the composition of pathways, protein complexes and other functional modules (Huynen, et al 2005). Such comparative analyses are usually based in the detection of orthology relationships between the components of a functional system in a certain species and proteins encoded in other genomes. Perhaps the most unexpected result from the first analyses of this kind was the finding of a relatively large degree of variation across species in the composition of metabolic pathways and complexes. In the case of large complexes, comparative genomics analyses have revealed significant variations of the subunit content in, among others, the proteasome (Gille, et al 2003), the nuclear pore complex (Mans, et al 2004) and the eukaryotic NADH:ubiquinone oxidoreductase (Complex I) (Gabaldón, et al 2005). In the latter, the differences were mapped onto the species phylogeny

and, assuming a parsimonious scenario, the history of gain and loss of subunits was reconstructed. The results showed a non-modular evolution of Complex I in the eukaryotes that contrasts with the modular pattern of evolution observed for this complex in the prokaryotes (Friedrich and Weiss 1997).

In the case of pathways, the evolutionary analyses of glycolysis (Dandekar, et al 1999), citric acid cycle (Huynen, et al 1999) or tryptophan synthesis (Xie, et al 2003), also revealed large divergences from the canonical pathway topology described in the textbooks. These results are indicative of a lack of modularity in the evolution of biochemical pathways. In other words, biochemical pathways do not seem to constitute evolutionary units that are either completely present or completely absent from a certain organism. Instead, a wide range of intermediate incomplete states of the pathway can be found. One might argue, however, that the lack of observed modularity might be related to the fact that the splitting of the cell metabolism into different pathways is rather artificial. For instance, glycolysis has many entry and exit points that are connected to other pathways such as the pentose phosphate pathway, glycerolipid metabolism or fructose and mannose metabolism. Therefore, the presence of pathways connected to glycolysis might involve the presence of glycolytic enzymes even in the absence of glycolysis itself.

To overcome such conceptual issues, Snel and Huynen (Snel and Huynen 2004) performed a large-scale analysis of the variation of functional modules that were defined using various criteria, including the automatic inference from high-throughput experiments results. Their observations are consistent with a general low degree of modularity in the evolution of functional modules. Nevertheless, half of the functional modules do tend to evolve more cohesively than random, indicating that a certain level of evolutionary coherence exists between functionally interacting proteins.

Phylogenetic reconstruction of ancestral proteomes

In the genomic era it has been possible to move from the evolutionary analysis of single protein families to that of complete genomes and

proteomes. Large-scale comparative genomics analyses have shown that, during evolution, the protein repertoire encoded in a species genome is continuously shaped by processes such as gene loss, gene gain and gene duplication. These processes can be studied and quantified by reconstructing ancestral proteomic states along the species tree (Snel, et al 2002). Moreover, as in the case of ancestral sequence reconstruction, reconstruction of ancestral proteomes allows us to test specific evolutionary scenarios.

One of the first such scenarios to be studied with the help of comparative genomics was the origin of the first cell and the properties of the so-called Last Universal Common Ancestor (Kyrpides, et al 1999, Mushegian and Koonin 1996). The reconstruction of this ancestral proteome involved the comparison of fully-sequenced genomes in terms of their content in protein-coding genes and a parsimonious reconstruction of the ancestral protein content. Recent estimates, that correct for horizontal gene transfers and non-orthologous gene displacements, suggest a simple last universal common ancestor with only 500-600 proteins (Koonin 2003). Although that amount of proteins might seem very small, it represents a substantial complexity if we consider that the minimal proteomic set to sustain cellular life in a rich medium could comprise as few as 206 proteins (Gil, et al 2004).

Another evolutionary scenario that has been investigated through ancestral proteome reconstruction is that of the origin of mitochondria. Mitochondria are eukaryotic organelles that originated from the endosymbiosis of an α proteobacterium and a proto-eukaryotic cell (Gray, et al 1999). Several hypotheses have been proposed that explain the initial endosymbiosis in terms of different metabolic properties of the host and the endosymbiont (Martin, et al 2001). To address this issue Gabaldón and Huynen reconstructed the mitochondrial ancestor proteome and inferred its metabolism (Gabaldón and Huynen 2003). In this case a phylogenomic approach was used to reconstruct the ancestral proteome. First, thousands of phylogenetic trees were reconstructed to subsequently select those whose topology indicated the presence of a member of that protein family in the mitochondrial ancestor. In addition, by mapping onto metabolic maps the functions of those protein

families, the proto-mitochondrial metabolism was inferred. The emerging picture is that of a (facultatively) aerobic endosymbiont catabolizing compounds provided by the host. In the absence of a reconstructed proteome for the host, it is difficult to define a specific symbiotic scenario. Nevertheless, the conservation of a diverse set of protomitochondrial pathways in the modern eukaryotes suggests a multifaceted benefit for the host cell.

Concluding remarks

The advent of bioinformatics, in combination with the availability of data obtained at a genome scale, has radically changed the way in which we study protein and proteome evolution. First, sophisticated tools for the comparison of protein sequences and the reconstruction of phylogenetic trees have allowed a better understanding of protein evolution at the molecular level. Moreover, the large variation observed in the composition of functional modules and proteomes from different species shows the great plasticity of living systems to adapt to different environments. Finally, through reconstructions of ancestral states it is now possible to trace the series of events that have shaped proteins and proteomes. Even the resurrection of extinct molecules is now a possibility. Will we see one day the experimental resurrection of an extinct cellular organism based on an ancestral proteome reconstruction? We will probably have to wait a long while. What is certain, however, is that future developments in bioinformatics will continue to shed light on the underlying mechanisms that govern the evolution of proteins and proteomes.

Acknowledgements

The author wishes to thank Martijn A. Huynen and members of his group for inspiring discussions about protein and proteome evolution. T.G. is recipient of a fellowship from the European Molecular Biology Organization (EMBO LTF 402- 2005)

References

Altschul, S. F. and Koonin, E. V. 1998. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem Sci*, 11 (23):444-7

Altschul, S. F., Madden, T. L., Schaffer, A. A., et al 1997. Gapped BLAST and PSIBLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 17 (25):3389-402

Apic, G., Gough, J. and Teichmann, S. A. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol*, 2 (310):311-25.

Baldi, P., Chauvin, Y., Hunkapiller, T., et al 1994. Hidden Markov models of biological primary sequence information. *Proc Natl Acad Sci USA*, 3 (91):1059-63.

Bapteste, E., Boucher, Y., Leigh, J., et al 2004. Phylogenetic reconstruction and lateral gene transfer. *Trends Microbiol*, 9 (12):406-11.

Bielawski, J. P. and Yang, Z. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J Mol Evol*, 1 (59):121-32.

Cai, W., Pei, J. and Grishin, N. V. 2004. Reconstruction of ancestral protein sequences and its applications. *BMC Evol Biol*, 1 (4):33.

Chang, B. S., Jonsson, K., Kazmi, M. A., et al 2002. Recreating a functional ancestral archosaur visual pigment. *Mol Biol Evol*, 9 (19):1483-9

Dandekar, T., Schuster, S., Snel, B., et al 1999. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem J*, (343 Pt 1):115-24.

Dayhoff, M. O. 1965. Computer aids to protein sequence determination. *J Theor Biol*, 1 (8):97-112.

Dimmic, M.W., Hubisz, M.J., Bustamante, C.D., Nielsen, R. 2005. Detecting coevolving amino acid sites using Bayesian mutational mapping. *Bioinformatics*. 21 Suppl 1:i126-i135.

Doolittle, R. F. 1995. The multiplicity of domains in proteins. *Annu Rev Biochem*, (64):287-314.

Dufayard, J.F., Duret, L., Penel, S., et al 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, 21(11): 2596-603.

Eck, R. V. and Dayhoff, M. O. 1966 Atlas of Protein Sequence and Structure. National Biomedical Research Foundation. silver Spring, Maryland.

Eddy, S. R. 1998. Profile hidden Markov models. *Bioinformatics*, 9 (14):755-63.

Edgar, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 1 (5):113.

Eisen, J. A. and Fraser, C. M. 2003. Phylogenomics: intersection of evolution and genomics. *Science*, 5626 (300):1706-7.

Felsenstein, J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol*, (266):418-27.

Felsenstein, J. 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J Mol Evol*. 53(4-5): 447-55.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. *Syst Zool*, 2 (19):99-113.

Fitch, W. M. 2000. Homology: a personal view on some of the problems. *Trends Genet*, 5 (16):227-31.

Fitch, W. M. and Margoliash, E. 1967. Construction of phylogenetic trees. *Science*, 760 (155):279-84.

Friedrich, T. and Weiss, H. 1997. Modular evolution of the respiratory NADH:ubiquinone oxidoreductase and the origin of its modules. *J Theor Biol*, 4 (187):529-40.

Fukami-Kobayashi, K., Schreiber, D.R., Benner, S.A. 2002. Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. *J Mol Biol*. 319(3):729-43.

Gaasterland, T. and Ragan, M. A. 1998. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb Comp Genomics*, 4 (3):199-217.

Gabaldón, T. and Huynen, M. A. 2004. Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci*, 7-8 (61):930-44

Gabaldón, T. and Huynen, M. A. 2003. Reconstruction of the protomitochondrial metabolism. *Science*, 5633 (301):609.

Gabaldón, T., Rainey, D. and Huynen, M. A. 2005. Tracing the Evolution of a Large Protein Complex in the Eukaryotes, NADH:Ubiquinone Oxidoreductase (Complex I). *J Mol Biol*, 4 (348):857-70.

- Gabaldón, T. and Huynen, M.A. 2005. Lineage-specific gene loss following mitochondrial endosymbiosis and its potential for protein function prediction in the eukaryotes. *Bioinformatics*, 21 Suppl 2: ii144-ii150.
- Galperin, M. Y. and Koonin, E. V. 2000. Who's your neighbor? New computational approaches for functional genomics. *Nat Biotechnol*, 6 (18):609-13.
- Gaucher, E. A., Thomson, J. M., Burgan, M. F., et al 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature*, 6955 (425):285-8.
- Gerstein, M. 1998. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold Des*, 6 (3):497-512.
- Gil, R., Silva, F. J., Pereto, J., et al 2004. Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev*, 3 (68):518-37.
- Gille, C., Goede, A., Schloetelburg, C., et al 2003. A comprehensive view on proteasomal sequences: implications for the evolution of the proteasome. *J Mol Biol*, 5 (326):1437-48.
- Goh, C. S., Bogan, A. A., Joachimiak, M., et al 2000. Co-evolution of proteins with their interaction partners. *J Mol Biol*, 2 (299):283-93.
- Gray, M. W., Burger, G. and Lang, B. F. 1999. Mitochondrial evolution. *Science*, 5407 (283):1476-81.
- Guindon, S. and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 5 (52):696-704.
- Hagen, J. B. 2000. The origins of bioinformatics. *Nat Rev Genet*, 3 (1): 231-6.
- Hughes, A. L. and Yeager, M. 1999. Coevolution of the mammalian chemokines and their receptors. *Immunogenetics*, 2 (49):115-24
- Huynen, M. A. and Bork, P. 1998. Measuring genome evolution. *Proc Natl Acad Sci U S A*, 11 (95):5849-56.
- Huynen, M. A., Dandekar, T. and Bork, P. 1999. Variation and evolution of the citric acid cycle: a genomic perspective. *Trends Microbiol*, 7 (7):28191.
- Huynen, M. A., Gabaldón, T. and Snel, B. 2005. Variation and evolution of biomolecular systems: searching for functional relevance. *FEBS Lett*, 8 (579):1839-45.
- Huynen, M. A. and Snel, B. 2000. Gene and context: integrative approaches to genome analysis. *Adv Protein Chem*, (54):345-79.
- Koonin, E. V. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol*, 2 (1):127-36.
- Koshi, J. M. and Goldstein, R. A. 1996. Probabilistic reconstruction of ancestral protein sequences. *J Mol Evol*, 2 (42):313-20.
- Krishnan, N. M., Seligmann, H., Stewart, C. B., et al 2004. Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference. *Mol Biol Evol*, 10 (21):1871-83.
- Kuhner, M. K. and Felsenstein, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol*, 3 (11):459-68.
- Kyrpides, N., Overbeek, R. and Ouzounis, C. 1999. Universal protein families and the functional content of the last universal common ancestor. *J Mol Evol*, 4 (49):413-23.
- Lio, P. and Goldman, N. 1998. Models of molecular evolution and phylogeny. *Genome Res*, 12 (8):1233-44.
- López, P., Casane, D., Philippe, H. 2002. Heterotachy, an important process of protein evolution. *Mol Biol Evol*, 19(1):1-7.
- Mans, B. J., Anantharaman, V., Aravind, L., et al 2004. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. *Cell Cycle*, 12 (3):1612-37.
- Martin, W., Hoffmeister, M., Rotte, C., et al 2001. An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. *Biol Chem*, 11 (382):1521-39.
- Mushegian, A. R. and Koonin, E. V. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci U S A*, 19 (93):10268-73.
- Nakhleh, L., Moret, B.M., Roshan, U. et al 2002. The accuracy of fast phylogenetic methods for large datasets. *Pac Symp Biocomput*, 7:211-22.
- Notredame, C., Higgins, D. G. and Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 1 (302):205-17.
- Park, J., Karplus, K., Barrett, C., et al 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol*, 4 (284):1201-10.
- Pazos, F., Helmer-Citterich, M., Ausiello, G., et al 1997. Correlated mutations contain information about protein-protein interaction. *J Mol Biol*, 4 (271):511-23.
- Pazos, F. and Valencia, A. 2002. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, 2 (47):219-27.
- Pazos, F. and Valencia, A. 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 9 (14):609-14.
- Pellegrini, M., Marcotte, E. M., Thompson, M. J., et al 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 8 (96):4285-8.
- Pupko, T., Pe'er, I., Shamir, R., et al 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol*, 6 (17):890-6.
- Ramani, A. K. and Marcotte, E. M. 2003. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol*, 1 (327):273-84.
- Redelings, B.D. and Suchard, M.A. 2005. Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.* 54(3):401-18.
- Ronquist, F. and Huelsenbeck, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 12 (19):1572-4.
- Saitou, N. N., M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4 (4):406-25.
- Schadt, E. E., Sinsheimer, J. S. and Lange, K. 1998. Computational advances in maximum likelihood methods for molecular phylogeny. *Genome Res*, 3 (8):222-33.
- Smith, T. F. and Waterman, M. S. 1981. Identification of common molecular subsequences. *J Mol Biol*, 1 (147):195-7.
- Snel, B., Bork, P. and Huynen, M. A. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res*, 1 (12):17-25.
- Snel, B. and Huynen, M. A. 2004. Quantifying modularity in the evolution of biomolecular systems. *Genome Res*, 3 (14):391-7.
- Takahashi, K. and Nei, M. 2000. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Mol Biol Evol*, 8(17):1251-8.
- Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., et al 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*, 1 (29):22-8.
- Teichmann, S. A., Park, J. and Chothia, C. 1998. Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc Natl Acad Sci U S A*, 25 (95):14658-63.
- Thompson, J. D., Higgins, D. G. and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22 (22):4673-80.
- Thornton, J. W. 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet*, 5 (5):366-75.
- Thornton, J. W., Need, E. and Crews, D. 2003. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science*, 5640 (301):1714-7.

- Valencia, A. and Pazos, F. 2003. Prediction of protein-protein interactions from evolutionary information. *Methods Biochem Anal*, (44):411-26
- Van de Peer, Y. 2004. Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet*, 10 (5):752-63.
- Wuchty, S. 2001. Scale-free behavior in protein domain networks. *Mol Biol Evol*, 9 (18):1694-702.
- Xie, G., Keyhani, N. O., Bonner, C. A., et al 2003. Ancient origin of the tryptophan operon and the dynamics of evolutionary change. *Microbiol Mol Biol Rev*, 3 (67):303-42.
- Yang, Z., Kumar, S. and Nei, M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 4 (141):1641-50.
- Zuckerandl, E. and Pauling, L. 1965. Molecules as documents of a history. *J Theor Biol*, 2 (8):357-66.