CMLS, Cell. Mol. Life Sci. 61 (2004) 930–944
1420-682X/04/080930-15
DOI 10.1007/s00018-003-3387-y

CMLS Cellular and Molecular Life Sciences

Review

Prediction of protein function and pathways in the genome era

T. Gabaldón and M. A. Huynen*

NCMLS, Nijmegen Center for Molecular Life Sciences, P/O: CMBI, Center for Molecular and Biomolecular Informatics, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen (The Netherlands), Fax: +31 24 3652977, e-mail: huynen@cmbi.kun.nl

Received 16 October 2003; received after revision 25 November 2003; accepted 26 November 2003

Abstract. The growing number of completely sequenced genomes adds new dimensions to the use of sequence analysis to predict protein function. Compared with the classical knowledge transfer from one protein to a similar sequence (homology-based function prediction), knowledge about the corresponding genes in other genomes (orthology-based function prediction) provides more specific information about the protein's function, while the analysis of the sequence in its genomic context (context-based function prediction) provides information about its functional context. Whereas homology-based methods predict the molecular function of a protein, genomic context methods predict the biological process in which it plays a role. These complementary approaches can be combined to elucidate complete functional networks and biochemical pathways from the genome sequence of an organism. Here we review recent advances in the field of genomic-context based methods of protein function prediction. Techniques are highlighted with examples, including an analysis that combines information from genomic-context with homology to predict a role of the RNase L inhibitor in the maturation of ribosomal RNA.

Key words. Comparative genomics; function prediction; pathways; genomic context; orthology; RNase L inhibitor.

Introduction

Since the completion of the first bacterial genome, that of *Haemophilus influenzae* in 1995 [1], the number published genome sequences has been growing exponentially [2] (fig. 1). At the time of this writing there are 144 fully sequenced genomes (excluding viral and organellar ones), of which 18 are of eukaryotic species, with at least an additional 134 prokaryotes and 33 eukaryotes in the pipeline [3]. The completion of a new genome sequence is followed by a process known as genome annotation to predict, among others, its protein coding regions and, to the extent that that is possible, their functions. This assigning of functions to predicted genes constitutes a major goal in the genomic era.

Despite the development of new advances in experimental techniques such as DNA microarrays [4, 5], yeast two-hybrid system [6], RNA interference (RNAi) [7, 8] or large-scale systematic deletions [9], experimental characterization of proteins to elucidate their function lags far behind the availability of new sequences, and the annotation of newly sequenced genomes relies mostly on computational methods. The most ancient and straightforward computational method for assigning function to a protein is based on the detection of homologs with known function (homology-based function prediction). 70–90% of the genes of a genome have homologs in other species [10], and these fractions will likely increase as new genomes are sequenced. However there is no clear functional prediction for ~40% of genes in most genomes [11], and for many of the rest, the predictions that can be made are only very general.

* Corresponding author.

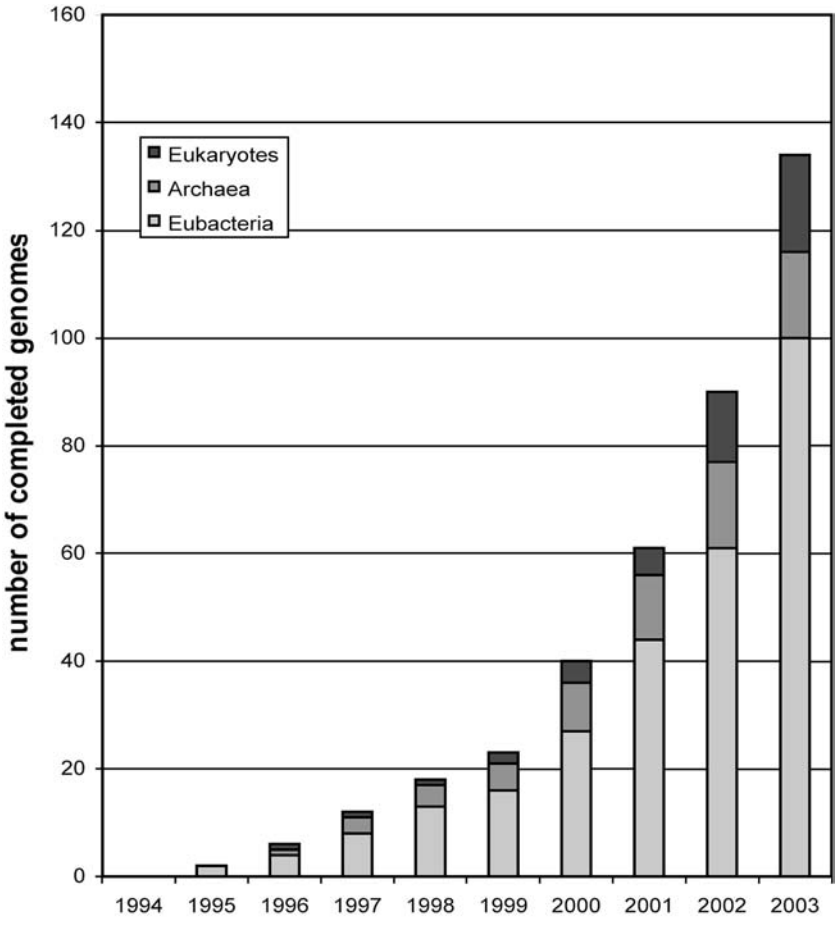


Figure 1. Cumulative number of fully sequenced genomes deposited in the public databases. Fraction of bacterial, archaeal and eukaryotic species are indicated with different grey scales. Data for year 2003 only represents fully sequenced genomes by August.

As the number and variability of sequenced genomes grew, so did ways of exploiting genomics data to predict protein function. First they have led to development of methods for large-scale orthology detection, allowing more specific function prediction than 'just' homology. Second they allowed development of methods, known as genomic context [12], or nonhomology [13] methods, that exploit information about the relations between genes on the genome, such as gene fusion, chromosomal proximity, distribution across species or conserved gene coexpression to predict functional interactions between their proteins (fig. 2). Here we review the conceptual and technical advances in function prediction based on the above methods.

From homology-based to orthology-based function prediction

Homologous proteins are proteins derived from a common ancestral sequence [14]. They have a similar three-dimensional (3D) structure and are likely to perform a similar function [15], at least at the molecular level. This is the basis of homology-based function prediction, in which one infers the function of a protein by extrapolating the knowledge from its experimentally characterized homologs. Initial characterization of new protein sequences starts by searching a protein database such as SWISS-PROT [16] with algorithms such as Smith-Waterman [17] and its faster approximation, BLAST [18], to detect experimentally characterized homologous sequences and obtain their functions. The sensitivity of such homology searches has more than doubled [19] thanks to profile-based methods such as PSI-BLAST [18] and Hidden Markov Models [20]. The genome era has had two main effects on homology-based function prediction. First of all, more sequences allow us to make better sequence profiles and have led to the development of domain databases such as SMART [21] and PFAM [22], and collections of those, such as InterPro [23]. For a recent review of the challenges and opportunities of protein domain analysis in the genome era see [24]. Second, and more important from a conceptual point of view, we can now do function prediction at a higher level of resolution,

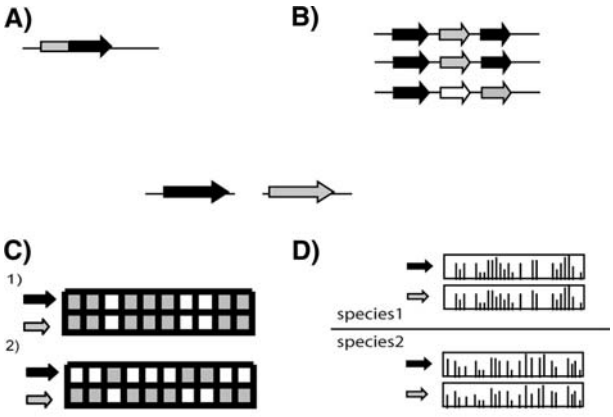


Figure 2. Main types of genomic context associations between two genes of a certain genome (centre of the figure). (*A*) Gene fusion: both sequences are encoded in a single gene in another genome. (*B*) Genomic neighborhood: both genes are close in the chromosome in several distantly related species, generally prokaryotes. (*C*) 1. Similar phylogenetic pattern: both genes have a similar pattern of presence/absence across species or 2. complementary phylogenetic pattern (*D*) Conserved co-expression: both genes have a similar pattern of expression under different conditions and this co-expression is conserved between species.

that of orthologous relations between genes [14]. Orthology, like homology, is primarily an evolutionary concept, and not a functional one. Sequences are orthologous when their independent evolution reflects a speciation event rather than a gene duplication event: i.e. they were one gene at the moment of speciation (fig. 3). Orthology is relevant for function prediction as orthologs are, relative to paralogs, more likely to perform the same function. Orthology is essential to genome comparison because it allows us to compare genomes in terms of their gene content. It is therewith also essential for methods that predict functional interactions between proteins based on the comparison of genomes' gene content (see below). Aside from being imperative for comparing genomes, orthology also relies on having complete genomes: techniques for large-scale orthology prediction such as 'best bi-directional hits' [25], and multiple-genome extensions thereof, such as the COG database, [26] depend on knowing the similarity levels between all genes in the genomes that are compared. The large-scale prediction of orthologous groups of proteins using such best-hit approaches is far from trivial. Aside from the technical issues such as homology detection, gene fusion and fission, and highly variable rates of evolution, there is the conceptual issue how to handle gene duplication [27] (fig. 3), which is rampant in eukaryotes [28]. Orthology databases that implicitly or explicitly 'trace-back' to the last common ancestor of life, in which all genes that were one gene in the last common ancestor are considered orthologous to each other, necessarily have a low level of evolutionary and functional resolution. The recent update of the COG database has a separate set of eukaryotic orthologous groups (KOGs), and has a much higher level of resolution for the eukaryotes than the original COG database [29]. Nevertheless, orthology determination by best-hit approaches is more prone to errors than the classical method of inferring orthology from phylogenetic trees, especially when

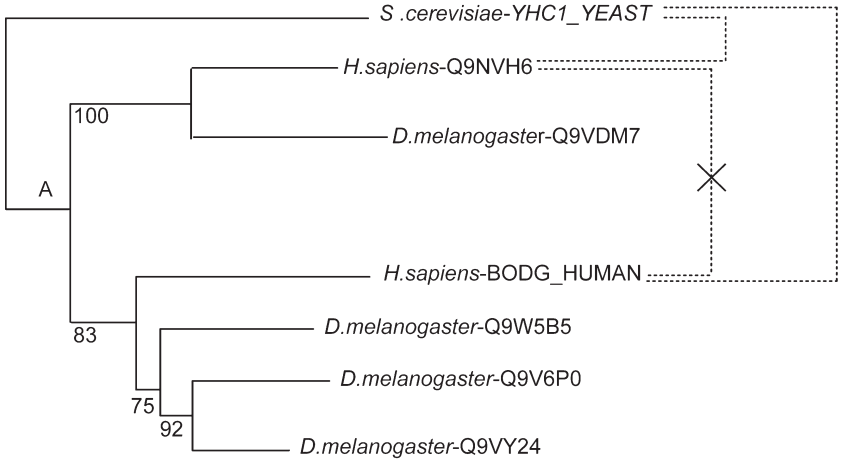


Figure 3. Orthology, paralogy and the conceptual issues when more than two species are compared. A parsimonious explanation of the tree of gamma-butyrobetaine, 2-oxoglutarate dioxygenase (BODG) and its close homologs *Homo sapiens, D. melanogaster* and *S. cerevisiae* is that there has been one gene duplication in the metazoa, and further duplications in *Drosophila*. The yeast gene can be considered orthologous to all the other genes in this tree, yet the human genes BODG (the last step of carnitine biosynthesis) and Q9NVH6 (the first step of carnitine biosynthesis) are not orthologous to each other (marked with a cross) because their independent evolution reflects a gene-duplication event at the root of the metazoan. Strictly speaking, orthology, in contrast to homology, is therefore nontransitive (if A is orthologous to B and B is orthologous to C, A and C are not necessarily orthologous to each other). A phylogeny-based orthology database that includes all eukaryotes would consider all these genes to be part of one orthologous group as they were all one gene in the last common ancestor of the fungi and metazoa. A taxon-specific orthology database, e.g one that is specific for the metazoa, would classify QNVH6 and BODG in different orthologous groups, providing a higher resolution for function prediction. Example taken from [27].

there is variation in the rate of sequence evolution within an orthologous group [30]. Although the evidence that it leads to better function prediction is still anecdotal, phylogenetic analyses for orthology prediction should in principle improve function prediction [30–32], and there are promising steps to implement these methods on a large scale [33–36].

Genomic-context based function prediction

Apart from symbiotic or parasitic cases, being encoded in the same genome is a prerequisite for two proteins to interact. It can therewith in principle also be used to predict interactions, but the information that two genes are encoded in the same genome provides of course only a very weak signal that they interact. A number of methods that use genomics comparison to predict functional interactions between proteins increase that signal by (a combination of) three strategies. (i) detecting a more direct association of the genes on the genome, e.g. a close physical association of the genes on the genome or the similar performance of two genes in genome-wide experiments, (ii) detecting evolutionary conservation of that association between species 'horizontal comparative genomics', and (iii) detecting the same association in different types of genomic context 'vertical comparative genomics'.

The type of information that such 'genomic context'-based prediction provides is qualitatively different from that of the homology-based one, because proteins that are part of the same biological process tend to occur in each other's genomic context, regardless of their sequence similarity. The type of information it provides is also very different from homology searches. While the identification of domains or homologs of a protein can give us a clue about its molecular function, the analyses of its genomic context instead allow us to identify interacting partners and the biological process in which is playing a role.

Gene fusion

The finding of two or more proteins encoded by separate genes of which orthologs in a different species are encoded in a single gene reveals a gene fusion or gene fission event [37]. This is the most direct form of genomic context and, from a functional point of view, the fusion of two proteins can result in an enhancement of the interaction between their respective biochemical activities to facilitate, for example, the channelling of a substrate [38]. This process has already been observed for several enzymes, and the inference of a functional link between two fused genes has been intuitively used on a small scale for many years, the most widely known example being the fusion of alpha and beta subunits of tryptophan synthetase in fungi [39]. An extension of this intuitive approach to the analyses of complete genomes was introduced in 1999 by Marcotte et al. [13] and Enright et al. [40]. By showing that many of the observed gene fusions events involved genes known to functionally interact, they proposed detecting gene fusions to predict interactions on a large scale. In concordance with the above-mentioned substrate-channelling effect, most of the observed fusion events involve metabolic enzymes. Although the fusions do not always involve subsequent steps in the pathway [40, 41], in *Escherichia coli* three quarters of the total of gene fusions affect metabolic genes [42]. More recently, a comparative study of 30 microbial genomes revealed that on average as much as 72% of annotated genes linked by fusion events belong to the same functional category [41]. Although gene fusion can be considered a rare event when comparing few genomes, the number of genes linked by fusions is likely to increase with the number of genomes that are compared: in the studies mentioned above we observe an ~114-fold increase from the 88 gene fusion events detected by comparing three bacterial genomes [40] to the 10,073 events when comparing 30 [41]. Not all of the detected gene fusions are equally informative, and one should be aware of the existence of promiscuous domains, such as the ones involved in signal transduction, that tend to appear in a variety of functional and genomic contexts [13].

One interesting property of the gene fusion approach is its transitiveness, which allows expanding the functional association to larger groups of genes that are interconnected. In other words, if gene B is fused with gene A in one genome and with gene C in another genome, then A, B and C form a functional network. Once again, the tryptophan synthesis pathway can serve to illustrate this example [43]. Whereas in *E. coli* we find a fusion between *trpG* and *trpD* and between *trpC* and *trpF*, in *Saccharomyces cerevisiae* we can detect, apart from the above-mentioned *trpA-trpB* fusion in tryptophan synthetase, a fusion involving *trpG* and *trpC*. From these pairwise relationships we can infer that *trpD, trpG, trpC* and *trpF* are part of the same functional network, in this case the tryptophan synthesis pathway. Extending the analyses to more genomes, all the proteins in the tryptophan synthesis pathway can be related by fusion events (fig. 4).

Chromosomal proximity in prokaryotes

The first pairwise genome-wide sequence comparisons revealed that even closely related species lack large-scale conservation of gene order [25, 44–46], indicating that in the course of evolution genomes are rapidly rearranged and shuffled. Yet in prokaryotes some clusters of genes appear conserved in evolution, including the relative location of the genes within them, over large evolutionary distances. Further inspection of these genes revealed that

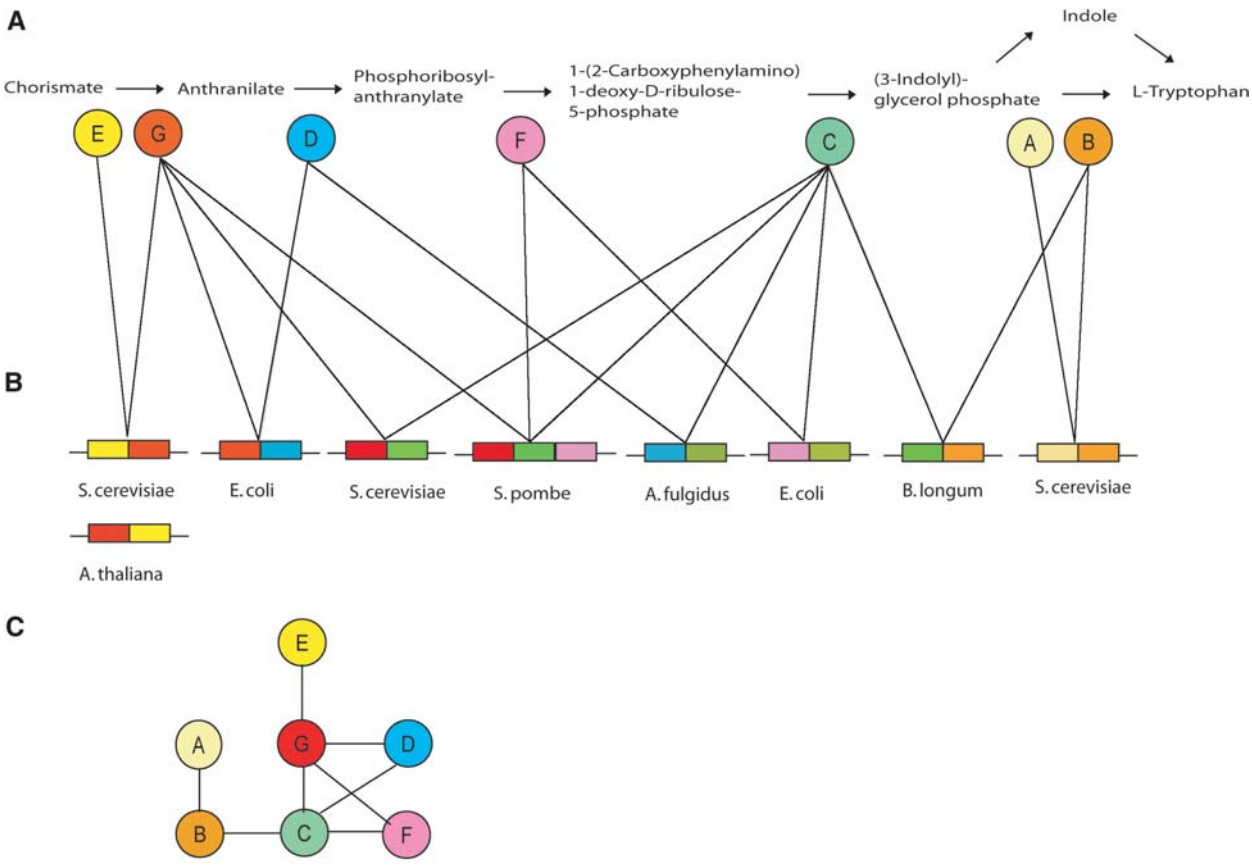


Figure 4. Gene fusions within the tryptophan synthesis pathway. (*A*) L-Tryptophan synthesis pathway and enzymatic activities associated with each step: E (yellow), anthranilate/para-aminobenzoate component I; G (red), anthranilate/para-aminobenzoate component II; D (blue), anthranilate phosphorybosil transferase; F (pink), phosphorybosilanthranilate isomerase; C (green), indole-3-glycerol phosphate synthase; A (cream), tryptophan synthase alpha chain; B (brown), tryptophan synthase beta chain. (*B*) Gene fusions observed in several fully sequenced genomes as provided by STRING database [146]; only one species for each fusion is chosen as an example. (*C*) Functional network derived from the fusion events within genes of the tryptophan synthesis pathways; enzymatic functions are linked if they are observed in the same polypeptidic chain in at least one genome.

they tend to encode proteins that functionally interact [47, 48], and that they tend to be part of the same operon [49]. As in the case of gene fusion, since conservation of chromosomal proximity has functional meaning, it can be used to predict functional interaction between the components of conserved gene clusters. This was proposed in 1998 by Overbeek et al. [48, 50] and Dandekar et al. [47], by measuring conservation of genes in runs (sets of genes encoded in the same strand and separated by <300 bases) and conservation of neighboring genes, respectively.

Chromosomal proximity in eukaryotes

Although operons are typically bacterial, some eukaryotes also use proximity in the genome to coordinate regulation [51, 52]. An analysis of gene clustering in five eukaryotic genomes [53] revealed that 30% (in *Drosophila melanogaster*) to 98% (in *S. cerevisiae*) of the pathways in KEGG show a significant clustering of their genes on the chromosomes. Furthermore, gene-order conservation between *S. cerevisiae* and *Candida albicans* appears, at least for divergently transcribed genes, to be correlated with co-expression [54–56]. The signals for function prediction in gene-order and gene-order conservation in eukaryotes are, however, weak and have not been employed for function prediction, although it can be argued that this is in part the result of the still relatively small number of sequenced eukaryotic genomes. Prokaryotic chromosomal proximity can be used for eukaryotic proteins with homologs in prokaryotic species, as in the case of the identification of the human methylmalonyl-coenzyme A (CoA) racemase [57]. Previous to its biochemical characterization, the function of this gene was first inferred based on the conserved chromosomal neighborhood of its prokaryotic homologs with genes involved in propionyl-CoA metabolism.

Similar phylogenetic distribution

Although the fact that two genes are encoded together in one genome provides only a very weak signal that they do interact, when they are encoded in a considerable number of genomes and are both absent from others this signal becomes strong enough for function prediction. This technique, called gene co-occurrence or phylogenetic profiles, was proposed [25, 58] and verified by studies showing that proteins with a similar distribution across species have a high tendency to functionally interact [25, 58–60]. Distribution across species is usually expressed by means of a phylogenetic pattern or profile: a string of letters or numbers that describe the presence or absence of a given gene in a set of genomes, and then detecting genes with a similar profile. Distances between profiles can be measured using a simple count of differences (Hamming distance) or more sophisticated scores such as mutual information [61] or Pearson correlation coefficient [62]. A typical example of a successful function prediction using phylogenetic distribution is that of the frataxin gene. Although the mutation in this gene was known to cause the human neurodegenerative disease Friedreich's ataxia [63], its molecular function remained unclear. In 2001 Huynen et al. [64] indicated that frataxin had the same phylogenetic distribution (fig. 5) as several iron-sulfur cluster assembly protein, suggesting a role in the same process for frataxin. The experimental confirmation that frataxin is actually involved in this process came a year later [65–67].

Because of its large potential for function prediction, the use of phylogenetic patterns to predict protein interactions is continuously undergoing technical improvements. A recent variation includes the use of phylogenetic patterns of neighboring gene pairs [68]. This combination of gene neighborhood and chromosomal proximity was shown to be more accurate than the single-gene phylogenetic profile, at a cost of coverage. Other modifications attempt to filter out the phylogenetic bias in the sequenced genomes (genomes of certain taxa are overrepresented) by using evolutionary information to measure the distance between profiles [69] or collapsing into a single node parts of the profile that represent related species that share the presence or absence of a certain gene [70].

Complementary phylogenetic distribution

A reverse use of phylogenetic profiles to predict function is the identification of proteins with complementary or anticorrelated profiles [10] to detect nonorthologous gene displacements [71]. Cases of experimentally confirmed function prediction are that of a new thymidilate synthase [72], and of seven enzymes involved in thiamine biosynthesis [73]. In general, the detection of nonorthologous gene displacement by complementary phylogenetic profiles is combined with gene-order conservation to increase the signal: i.e. does the 'new' gene occur in conserved operons with the other genes with which it is supposed to interact, replacing the old gene not only in terms of functional context but also in terms of genomic context.

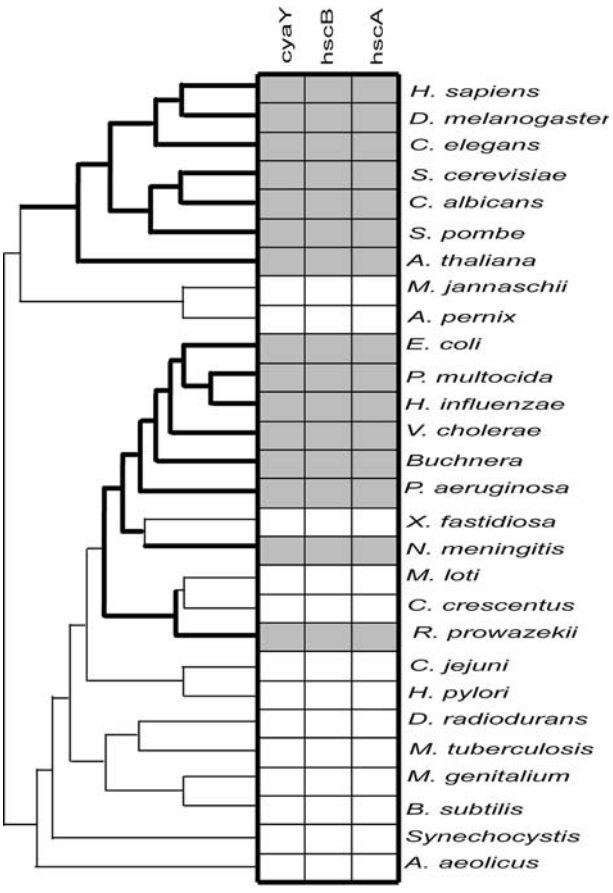


Figure 5. Phylogenetic distribution of frataxin (cyaY) and that of hscB and hscA: Grey and white boxes indicate, respectively, presence and absence of the genes in a certain species (names on the right). The species phylogeny is indicated with thick and thin lines, respectively, indicating presence or absence of the genes.

Correlated gain and loss of genes

One methodological issue in the comparison of phylogenetic profiles is that there is a strong phylogenetic signal in the genes two genomes share [74]: i.e. the fact that two genes occur together in a number of closely related genomes does not necessarily imply a functional interaction. One can solve this by detection of profiles that are not in agreement with the species phylogeny, indicating that a pair of genes have been lost or gained together in a genome. This method was applied to the prediction of

genes responsible for pathogenicity by identifying genes present in a pathogenic species that are absent in closely related nonpathogenic species or strain [75, 76], or genes responsible for host specificity by detecting differences between similar pathogens that affect different hosts [77]. A functional linkage between genes that have been lost in the same lineages has also been shown in eukaryotes by comparing the genomes of the fungi *S. cerevisiae* and *Schizosaccaromyces pombe* [78] and in Archaea by comparing the three sequenced genomes from the *Pyrococcus* genus [79].

Coevolution of sequences

Another variant of the use of coevolution to predict protein function uses the evolutionary information that is contained at a lower level than the distribution across species: that of the sequences themselves. For specific cases of protein families known to interact, such as insulin and its receptors [80] or the chemokine-receptor system [81, 82], their phylogenetic trees are more similar to each other than expected based on the general divergence between the corresponding species. This was interpreted as an indication of correlated evolution reflecting similar evolutionary constrains. Valencia et al. [83, 84] made use of this property to search for interaction partners within the *E. coli* proteome by measuring the correlation between the distance matrices used to build the phylogenetic trees (fig. 6). Provided good species coverage and quality of the multiple sequence alignment, the technique can indeed distinguish statistically true interactions among many possible alternatives. Ramani and Marcotte [85] used a similar approach (fig. 6) to predict the binding specificities among members of 18 ligand and receptor families that posses many paralogs in the human genome. The coevolution of interacting partners can be followed more closely by searching for mutations that are correlated in both protein families (they occur in the same species). These positions may correspond to residues on the interface that undergo compensatory mutations in one protein to compensate the effects of mutations in the other. This information was initially used to detect proximal residues to predict protein folding [86] or discriminate between different structural models [87], and was later extended to the prediction of interacting partners based on the finding of pairs of proteins with correlated mutations [88, 89]. The method has the advan-

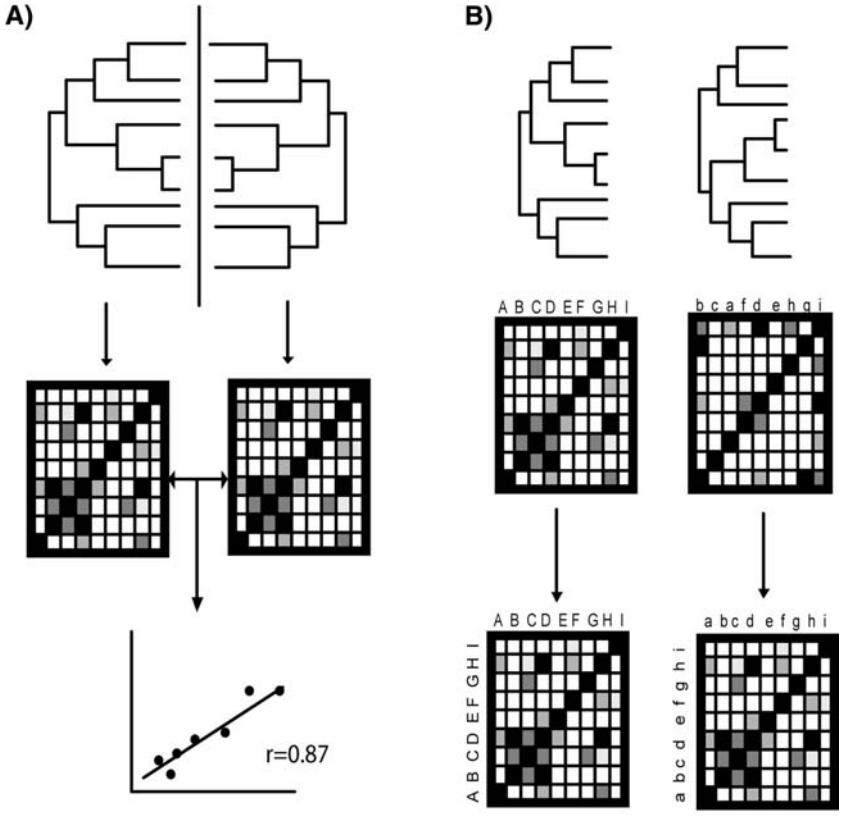


Figure 6. Valencia (*A*) and Ramani (*B*) methods for predicting protein interactions: (*A*) Similarity between the phylogenies of two protein families is estimated by measuring the correlation between their respective distance matrices. (*B*) The interaction specificity between the members of two interacting protein families is predicted by means of shuffling the columns and rows of the distance matrix of one of the families until the agreement with the other one is maximized. The interactions are then predicted for proteins heading the same columns.

tage that it provides not only the prediction of interacting partners but also of potentially interacting residues. Currently, 3D information about proteins is rarely used to predict which proteins interact with which ones. An exception is the use of 3D information combined with sequence covariation on the potential interaction surfaces to predict whether homologs of interacting proteins will interact in the same way [90]. 3D information is also used by docking procedures for interaction prediction, but these methods predict how and where two proteins interact with each other rather than which ones interact with which. See [91] for a recent update.

Conservation of coexpression

Besides genome sequences, the only experimental context data to date that are truly genome wide are expression data such as microarrays [4, 5] or SAGE [92]. These are similar to genome-context data in the sense that they reflect functional interactions between proteins [93–96], and can therewith also be used to predict them [97–101]. The relative high level of noise that is usually encountered in the expression profiles can be reduced by detection of a correlation between two genes among different experiments [98, 101]. Still, this technique detects functional interactions of a general kind [97, 102]. Recently the observation from other types of genomic-context methods that evolutionary conservation dramatically increases the reliability of the prediction has been shown to apply to coexpression as well [36, 103, 104]. An interesting extension to the idea that conservation of coexpression increases the likelihood of functional interaction is to apply it to conservation after gene duplication: the likelihood that two genes (A and B) that are co-expressed interact increases when their paralogs (A′ and B′) are also coexpressed [36]. The concept of evolutionary conservation of interaction has also been applied to yeast-two hybrid data to detect pathways conserved between *Helicobacter pylori* and *S. cerevisiae* or duplicated within *S. cerevisiae* [105].

Accuracy and coverage of context-based methods

Large-scale analyses of context-based methods indicate a high accuracy, especially gene fusion, with estimates ranging from 72 to 95% [41, 70], gene-order conservation (80 to 95%) [70, 106] and conserved coexpression (> 95%) [36]. In general, the likelihood that any prediction is true can be increased, at a cost of coverage, by either combining different context-based methods (vertical comparative genomics) [70] (see also [107] for the combination of other types of genomics data) or increasing the required evolutionary conservation (horizontal comparative genomics) [55, 70, 106]. In terms of coverage, context-based methods will likely surpass that of the homology-based ones in the near future, at least for prokaryotes: in 2002 conserved neighborhood could provide reliable prediction for roughly 60% of *E. coli* genes, a fraction approaching that of genes with an annotated homolog in SWISS-PROT (70%) [108]. When compared with experimental genome-wide methods, genome-context predictions have a higher coverage and accuracy than several genomics experimental techniques such as yeast two-hybrid or simple (not-conserved) coexpression [109]. More important than such 'beauty contests' is that combining experimental and computational techniques increases accuracy while maintaining a reasonable coverage [109]. In other words, it makes sense, after a high-throughput experiment, to combine the results with genomic-context analyses of one's proteins to identify the most likely interactions, e.g. with a public domain database such as STRING [70].

Benchmarking and experimental verification

One of the bottlenecks in the development of methods for the prediction of functional interaction is the availability of benchmarks with experimentally determined interactions. Manually curated databases of protein interactions [110–112] have a relatively small coverage, and especially the fraction of false negatives (what fraction of true interactions are not predicted by a method) is therefore generally hard to estimate. It should be noted that 'functional interaction' is, even more than 'function', not a strictly defined term that can range from a direct stable physical interaction to less direct ones such as 'being part of the same biological process'. Any reference set used for benchmarking should therefore be categorized into different types of interactions, allowing not only quantitative (what is the fraction of false positives?) but also qualitative evaluation (what types of interaction do we detect?) of function prediction. Manual analyses can of course be more thorough in evaluation of the experimental evidence and can make such a distinction between different types of interactions, but they are necessarily limited to relatively small sets of proteins, such as the 480 proteins of *Mycoplasma genitalium* [12]. Large-scale benchmarking of context-based functional annotation use classifications such as presence of the same key words describing function in SWISS-PROT [113], having the same gene ontology annotation [114], belonging to the same functional class according to a database such as COG [115] or being part of the same pathway according to KEGG [116]. It should be noted that benchmarking is often done in terms of enrichment of proteins of a certain functional class, e.g. when proteins in a cluster have a statistically significant higher than average probability of

being involved in the same pathway. Such a significant pattern does not necessarily imply that highly reliable predictions can be made [113]. Reliable prediction can only be made when a large fraction (e.g. 90%) of the proteins with known function in a cluster belong to the same pathway.

Experimental verification of predictions made by genomic context methods still lags far behind the many predictions that have been made. In a recent survey we identified 13 cases of predictions that were experimentally verified [108], which is a small fraction compared with the literally hundreds that have been made, e.g. [12, 46, 117]. We expect that improving accessibility of genomic-context data [70] will facilitate the usage of experimental groups to exploit them and to couple the predictions directly to experimental verification.

From functional interactions to biochemical pathways and networks

By combining pairwise interactions, one can derive networks of protein interactions. The study of such networks, in which nodes are connected when they are (predicted to be) involved in the same biological process, has revealed that they are so-called scale-free networks. This means that there is not a typical number of connections per node; rather, the distribution of the number of connections (k) per node *(N)* follows a power law [$N(k) \sim k^{-\gamma}$]. In other words, there are many nodes with few connections and a small but still significant number of nodes with many interactions. These highly connected nodes tend to be relatively essential to an organism [118] and to evolve relatively slowly [119]. Protein interaction networks can be analyzed for structures that reflect function and selection at a level higher than pairwise interactions, e.g. that of pathways and protein complexes, and that can also be used for higher-order function prediction. Such 'functional modules' are represented in the network by sets of proteins that are locally highly interconnected to each other and less connected to other proteins [120, 121]. Detecting that an uncharacterized protein is part of such a highly connected cluster of proteins allows the prediction that the protein is part of that module [120, 122]. Furthermore, finding interactions between an uncharacterized protein and multiple proteins from the same module increases the likelihood that the predicted involvement is indeed correct [27, 121, 123].

Global constraints on network structure?

Besides the high local clustering coefficient, another aspect of protein networks that has been argued to reflect selection at a level higher than pairwise interactions is the diameter, i.e. what is, on average, the minimal number of steps one needs to get from any node to any other node? It should be noted that the direction of this argument has been rather arbitrary: both the relatively small diameter of metabolic networks [124] as well as a relatively large diameter of protein interaction networks [125] have been argued to be the result of selection. Subsequent analyses have, however, shown that in either case the networks were more random than proposed and that the observed biases in the diameter size were either due to the choice of the network nodes [126] or experimental bias in the underlying dataset [127]. Whether the global architecture of biological interaction networks is relevant for function or merely puts boundaries on the evolutionary process that created it is still open to debate. Neutral models for the evolution of the networks are able to capture aspects such as their scale-free architecture [128] and high cliquishness [129], and in our view there is no convincing evidence for selection in the evolution of the global network architecture. This does not of course imply that evolution has not capitalized on the neutrally evolved network structure as such, or that the individual links between proteins are meaningless. It means that the scale-free architecture as such is no evidence for selection.

Metabolic pathway prediction

A special case of functional networks are metabolic pathways. Their reconstruction from a species genome sequence has been possible through the combination of homology-based methods, to determine the molecular function of the proteins, with the identification of their functional partners by context-based techniques [130–133]. The comparison of reconstructed central metabolic pathways such as glycolysis [134] and citric acid cycle [135] from different organisms revealed a surprising plasticity with the existence of many species-specific variations. Such deviations from the canonical pathway can be used to identify drug targets when certain alternative enzymes or bypasses are specific to the pathogen [136].

Combining homology and context for function prediction: a case history

A typical example of what is and is not possible in function prediction by the methods discussed here is the protein RNase L inhibitor (RLI). Experimental results have indicated that this protein reversibly associates with RNase L, which it inhibits [137], although the exact mechanism of inhibition has not been elucidated. Furthermore, RLI interacts with the human immunodeficiency virus (HIV)-1 protein VIF [138]. Functionally, these two activities of RLI are not likely to be the whole

story, however. RLI is present in all eukaryotes and all Archaea that have been sequenced so far (fig. 7), but an examination of the SMART database [21] indicates that only mammals have proteins with the domain organization of its interaction partner in human, RNase L. Furthermore the interaction with HIV-1 proteins can hardly be the original reason for the proteins existence. We examined information from genomic context data and from homology to derive a new hypothesis for the function of RLI. First, we examined whether there are other orthologous groups that specifically tend to cooccur with RLI. An examination of the STRING database indicates that only 55 orthologous groups have a phylogenetic distribution that is identical to RLI. For the orthologous groups in this set of which we know the function (44), nearly all are either involved in translation or ribosome biogenesis (33, 60%), in transcription (7, 16%) and in DNA replication, recombination and repair (3, 7%), matching the general pattern that the proteins that the eukaryotes obtained from the Archaea are mainly involved in the replication and processing of DNA and RNA. These correlations point to a role of RLI in DNA replication and transcription or RNA processing. From a second type of genomic context, the conservation of coregulation [36, 104], comes a prediction that is consistent with this observation, but that is more specific. Between *S.cerevisiae* and *Caenorhabditis elegans* RLI is conservedly coexpressed with a number of proteins involved the processing of ribosomal RNA (rRNA) such as the nucleolar protein SIK1 (NOP56) from yeast that is involved in rRNA methylation [36]. A conservation of coexpression between four species study [104] hints furthermore at interactions with (i) another nucleolar protein involved in rRNA processing, NOP4, (ii) a DEAD box RNA helicase, DBP2, and (iii) the B and C subunits of RNA polymerase I (fig. 7). A possible link with an RNA helicase can also be inferred from genomic context information: in one archaeal genus, *Methanosarcina*, RLI is located in a potential operon with a DEAD box helicase. A third type of information, that of the domain structure of the protein, is quite specific. The protein contains two domains each with four conserved cysteines, one unique to RLI, and one a 4Fe-4S binding domain, and furthermore two ATPase domains (fig. 7). One possibility to link this domain organization to RNA interaction lies in the 4Fe-4S binding domain. Aside from playing a role in redox reactions, these domains have also been observed in DNA-binding protein endonuclease III [139] and the DNA glycosylase MutY [140]. The 4Fe-4S cluster is hypothesized to stabilize the fold, presenting a loop that extends from the backbone of the DNA [139]. Consistent with a role of both the 4-cysteine domains of RLI is that they both contain conserved lysines (positively charged) that could interact with the phosphate backbone of rRNA (negatively charged).

Examples such as this one show the potential and limits of using comparative genomics in protein function prediction. We can pinpoint a role in a process, but cannot always predict exactly what that role is. In the case of metabolic pathways a molecular function such as enzymatic activity and context function such as the pathway can often be matched to obtain to a specific prediction. In the case of RLI case the situation is less obvious.

Discussion

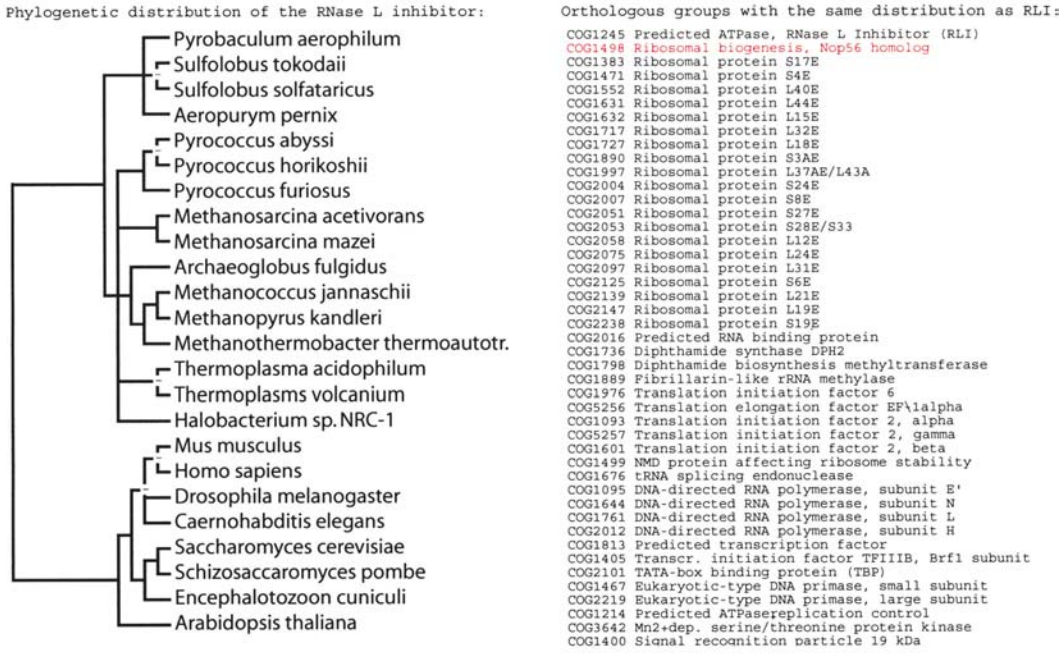
The use of genome sequences to predict protein function is in its adolescence. As we have reviewed here, a number of concepts have been introduced, compared and combined for the prediction of function and of functional interaction. Furthermore, benchmarking against a variety of databases indicates the generally high reliability of the predictions. These advances have not only been made possible by the availability of genomics data, they also contribute to the exploitation of these data. The use of context-based techniques has proven useful in improving the annotation of complete genomes [12, 141], and annotation at the level of orthologous groups is included in genome annotation [115].

There are, however, a number of challenges that we will need to tackle if genomic context wants to reach the same level of success as homology-based function prediction. On the practical side, we need better integration of the various signals, both from homology as well as from genomic context. Not to make more predictions, but rather to make more specific predictions that are directly amenable for experimental testing: i.e. we would like to predict not only in which biological process protein plays a role, but what the protein does there. As the example from the RLI shows, one can use many different sources of information that are relevant to function: combining those in a (semi-)automatic way will not be trivial. On the theoretical side we need a better understanding of why our methods actually work. While vertical comparative genomics (comparing different types of data from one species) appears to be a straightforward way of filtering out noise, in horizontal comparative genomics (the comparison of the same type of data from different species), a second effect is likely to play a role: filtering out species-specific interactions with little general relevance. Classic examples of such species-specific interactions are the coregulation of ribosomal genes with genes in glycolysis in *Halobacterium* and *S. cerevisiae* [142, 143]. Such interactions do not fit into our platonic view of what constitutes a functional interaction, yet we can very well imagine why genes for glycolysis and the ribosome are coregulated. This does confront us with the question how to define functional interactions between proteins, when protein function itself is already a concept that can be de-

A. The co-expression of RNase L inhibitor (RLI) with rRNA processing proteins, including NOP56, is conserved between multiple species:

gene	protein function	co-expression conservation with RLI, p-value
SIK1	Nucleolar, rRNA processing (NOP56)	0.00128
NOP4	Nucleolar, rRNA processing	0.00133
RPA135	RNA polymerase I, beta subunit	0.00197
DBP2	ATP dependent RNA helicase (DEAD-box)	0.00404
RPC40	RNA polymerase I/III, c subunit	0.00607

B. RLI has the same phylogenetic distribution as a number of proteins (COGs) involved in replication, transcription and translation, including NOP56:



C. The ferredoxin-like, N-terminal domains of RLI contain conserved, positively charged lysines (K), consistent with a role in binding the negatively charged rRNA backbone:

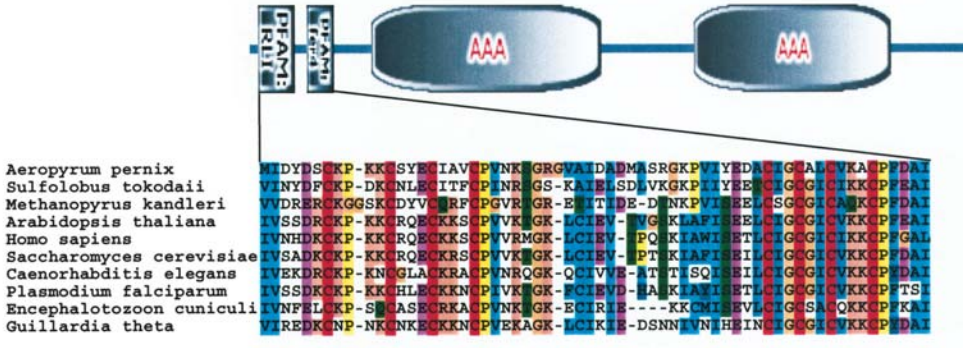


Figure 7. Genomic context (phylogenetic distribution and conserved coexpression) and sequence homology (domain organization and conserved residues in the sequence alignment) hint at a role of the RLI in the processing of RNA. (*A*) Conserved coexpression from http://cmgm.stanford.edu/~kimlab/multiplespecies [104]. The genes (names from *S. cerevisiae*) that are conservedly coexpressed with RLI (P value < 0.001) in *S. cerevisiae, C. elegans, D. melanogaster* and *H. sapiens* and for which functional information is available can all be linked to transcription and processing of rRNA. (*B*) The orthologous groups (right panel) that have the same phylogenetic distribution as RLI (present in all Archaea and all eukaryotes, left panel), based on the COGs [26] as implemented STRING http://string.embl-heidelberg.de [70]. Most genes are involved in replication, transcription and translation. The overlap with the conservedly coexpressed set of genes (*A*) is NOP56 (in red). (*C*) The domain organization of RLI from SMART http://smart.embl-heidelberg.de [21] with an alignment of the N-terminal, four-cysteine domains of a representative set of sequences, constructed with clustalx [147]. In each sequence both sets of four cysteines contain at least one intercysteine loop (between the first two cysteines and between the last two cysteines) with a positively charged residue (lysine). See text for further details.

fined at many levels [144]. The genome era has made this question much less academic than it used to be, leading to important, if simplifying standarization [145], and that itself is an important and unanticipated achievement in the field of biology to which the concept of function is as central as ever.

1 Fleischmann R. D., Adams M. D., White O., Clayton R. A., Kirkness E. F., Kerlavage A. R. et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269:** 496–512
2 van Nimwegen E. (2003) Scaling laws in the functional content of genomes. Trends Genet. **19:** 479–484
3 Benson D. A., Karsch-Mizrachi I., Lipman D. J., Ostell J. and Wheeler D. L. (2003) GenBank. Nucleic Acids Res **31:** 23–27
4 Schena M., Shalon D., Davis R. W. and Brown P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 467–470
5 Lockhart D. J., Dong H., Byrne M. C., Follettie M. T., Gallo M. V, Chee M. S. et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. **14:** 1675–1680
6 Chien C. T., Bartel P. L., Sternglanz R. and Fields S. (1991) The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. Proc. Natl. Acad. Sci. USA **88:** 9578–9582
7 Fire A., Xu S., Montgomery M. K., Kostas S. A., Driver S. E. and Mello C. C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature **391:** 806–811
8 Kamath R. S. and Ahringer J. (2003) Genome-wide RNAi screening in *Caenorhabditis elegans*. Methods **30:** 313–321
9 S. Vidan and M. Snyder (2001) Large-scale mutagenesis: yeast genetics in the genome era. Curr. Opin. Biotechnol. **12:** 28–34
10 Galperin M. Y. and Koonin E. V. (2000) Who's your neighbor? New computational approaches for functional genomics. Nat. Biotechnol. **18:** 609–613
11 Iliopoulos I., Tsoka S., Andrade M. A., Janssen P., Audit B., Tramontano A. et al. (2001) Genome sequences and great expectations. Genome Biol. **2:** interactions 0001.1– 0001.3
12 Huynen M., Snel B., Lathe W., 3rd and Bork P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res. **10:** 1204–1210
13 Marcotte E. M., Pellegrini M., Ng H. L., Rice D. W., Yeates T. O. and Eisenberg D. (1999) Detecting protein function and protein-protein interactions from genome sequences. Science **285:** 751–753
14 Fitch W. M. (1970) Distinguishing homologous from analogous proteins. Syst. Zool. **19:** 99–113
15 Teichmann S. A. (2002) The constraints protein-protein interactions place on sequence divergence. J. Mol. Biol. **324:** 399–407
16 Boeckmann B., Bairoch A., Apweiler R., Blatter M. C., Estreicher A., Gasteiger E. et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31:** 36–370
17 Smith T. F. and Waterman M. S. (1981) Identification of common molecular subsequences. J. Mol. Biol. **147:** 195–197
18 Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402
19 Park J., Karplus K., Barrett C., Hughey R., Haussler D.,. Hubbard T et al. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J. Mol. Biol. **284:** 1201–1210
20 Baldi P., Chauvin Y., Hunkapiller T. and McClureM. A. (1994) Hidden Markov models of biological primary sequence information. Proc. Natl. Acad. Sci. USA **91:** 1059–1063
21 Letunic I., Goodstadt L., Dickens N. J., T. Doerks, Schultz J., Mott R. et al. (2002) Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res. **30:** 242–244
22 Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S. R. et al. (2002) The Pfam protein families database. Nucleic Acids Res. **30:** 276–280
23 Mulder N. J., Apweiler R., Attwood T. K., Bairoch A., Bateman A., Binns D. et al. (2002) InterPro: an integrated documentation resource for protein families, domains and functional sites. Brief Bioinform. **3:** 225–235
24 Copley R. R., Doerks T., Letunic I. and Bork P. (2002) Protein domain analysis in the era of complete genomes. FEBS Lett. **513:** 129–134
25 Huynen M. A. and Bork P. (1998) Measuring genome evolution. Proc. Natl. Acad. Sci. USA **95:** 5849–5856
26 Tatusov R. L., Natale D. A,. Garkavtsev I. V, Tatusova T. A., Shankavaram U. T., Rao B. S. et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. **29:** 22–28
27 Sonnhammer E. L. and Koonin E. V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet. **18:** 619–620
28 Lespinet O., Wolf Y. I., Koonin E. V. and Aravind L. (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res. **12:** 1048–1059
29 Tatusov R. L., Fedorova N. D., Jackson J. J., Jacobs A. R., Kiryutin B., Koonin E. V. et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:** 41
30 Eisen J. A. and Fraser C. M. (2003) Phylogenomics: intersection of evolution and genomics. Science **300:** 1706–1707
31 Saier M. H., Jr., Eng B. H., Fard S., Garg J., Haggerty D. A., Hutchinson W. J. et al. (1999) Phylogenetic characterization of novel transport protein families revealed by genome analyses. Biochim. Biophys. Acta **1422:** 1–56
32 Eisen J. A., SwederK. S. and Hanawalt P. C. (1995) Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. Nucleic Acids Res. **23:** 2715–2723
33 Storm C. E. and Sonnhammer E. L. (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics **18:** 92–99
34 Yuan Y. P., Eulenstein O., Vingron M. and Bork P. (1998) Towards detection of orthologues in sequence databases. Bioinformatics **14:** 285–289
35 Arvestad L., Berglund A. C., Lagergren J. and Sennblad B. (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics **19 Suppl. 1:** I7–I15
36 van Noort V., Snel B. and Huynen M. A. (2003) Predicting gene function by conserved co-expression. Trends Genet. **19:** 238–242
37 Snel B., Bork P. and Huynen M. (2000) Genome evolution. Gene fusion versus gene fission. Trends Genet. **16:** 9–11
38 Welch G. R. and Easterby J. S. (1994) Metabolic channeling versus free diffusion: transition-time analysis. Trends Biochem. Sci. **19:** 193–197
39 Burns D. M., Horn V., Paluh J. and Yanofsky C. (1990) Evolution of the tryptophan synthetase of fungi. Analysis of experimentally fused *Escherichia coli* tryptophan synthetase alpha and beta chains. J. Biol. Chem. **265:** 2060–2069
40 Enright A. J., Iliopoulos I., Kyrpides N. C. and Ouzounis C. A. (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature **402:** 86–90

41 Yanai I., Derti A. and DeLisi C. (2001) Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc. Natl. Acad. Sci. USA **98:** 7940–7945
42 Tsoka S. and Ouzounis C. A. (2000) Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. Nat. Genet. **26:** 141–142
43 Yanofsky C. (1984) Comparison of regulatory and structural regions of genes of tryptophan metabolism. Mol. Biol. Evol. **1:** 143–161
44 Mushegian A. R. and Koonin E. V. (1996) Gene order is not conserved in bacterial evolution. Trends Genet **12:** 289–290
45 Watanabe H., Mori H., Itoh T. and Gojobori T. (1997) Genome plasticity as a paradigm of eubacteria evolution. J. Mol. Evol. **44 Suppl. 1:** S57–64
46 Wolf Y. I., Rogozin I. B., Kondrashov A. S. and Koonin E. V. (2001) Genome alignment, evolution of prokaryotic genome organization and prediction of gene function using genomic context. Genome Res. **11:** 356–372
47 Dandekar T., Snel B., Huynen M. and Bork P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. **23:** 324–328
48 Overbeek R. F., D'Souza M., Pusch G. D. and Maltsev N. (1998) Use of contiguity on the chromosome to infer functional coupling. In Silico Biol. **2:** 93–108
49 Moreno-Hagelsieb G., Trevino V., Perez-Rueda E., Smith T. F. and Collado-Vides J. (2001) Transcription unit conservation in the three domains of life: a perspective from *Escherichia coli*. Trends Genet. **17:** 175–177
50 Overbeek R., Fonstein M., D'Souza M., Pusch G. D. and Maltsev N. (1999) The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA **96:** 2896–2901
51 Blumenthal T. (1998) Gene clusters and polycistronic transcription in eukaryotes. Bioessays **20:** 480–487
52 Spieth J., Brooke G., Kuersten S., Lea K. and Blumenthal T. (1993) Operons in *C. elegans*: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. Cell **73:** 521–532
53 Lee J. M. and Sonnhammer E. L. (2003) Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. **13:** 875–882
54 Hurst L. D., Williams E. J. and Pal C. (2002) Natural selection promotes the conservation of linkage of co-expressed genes. Trends Genet. **18:** 604–606
55 Huynen M. A. and Snel B. (2003) Exploiting the variation in the genomic associations of genes to predict pathways and reconstruct their evolution. In: Frontiers in Computational Genomics, pp. 145–166, Galperin M. Y. and Koonin E. V. (eds), Caister Academic Press, Norfolk, UK
56 Huynen M. A., Snel B. and Bork P. (2001) Inversions and the dynamics of eukaryotic gene order. Trends Genet. **17:** 304–306
57 Bobik T. A. and Rasche M. E. (2001) Identification of the human methylmalonyl-CoA racemase gene based on the analysis of prokaryotic gene arrangements. Implications for decoding the human genome. J. Biol. Chem. **276:** 37194–37198
58 Pellegrini M., M. Marcotte E., Thompson M. J., Eisenberg D. and Yeates T. O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA **96:** 4285–4288
59 Huynen M. A. and Snel B. (2000) Gene and context: integrative approaches to genome analysis. Adv. Protein Chem. **54:** 345–379
60 Gaasterland T. and Ragan M. A. (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. Microb. Comp. Genomics **3:** 199–217
61 Huynen M., Snel B.,. Lathe W and. Bork P (2000) Exploitation of gene context. Curr. Opin. Struct. Biol. **10:** 366–370
62 Wu J., Kasif S. and DeLisi C. (2003) Identification of functional links between genes using phylogenetic profiles. Bioinformatics **19:** 1524X–1530
63 Campuzano V., Montermini L., Molto M. D., Pianese L., Cossee M., Cavalcanti F. et al. (1996) Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science **271:** 1423–1427
64 Huynen M. A., Snel B., Bork P. and Gibson T. J. (2001) The phylogenetic distribution of frataxin indicates a role in iron-sulfur cluster protein assembly. Hum. Mol. Genet. **10:** 2463–2468
65 Muhlenhoff U., Richhardt N., Ristow M., Kispal G. and Lill R. (2002) The yeast frataxin homolog Yfh1p plays a specific role in the maturation of cellular Fe/S proteins. Hum. Mol. Genet. **11:** 2025–2036
66 Chen O. S., Hemenway S. and Kaplan J. (2002) Inhibition of Fe-S cluster biosynthesis decreases mitochondrial iron export: evidence that Yfh1p affects Fe-S cluster synthesis. Proc. Natl. Acad. Sci. USA **99:** 12321–12326
67 Duby G., Foury F., Ramazzotti A., Herrmann J. and Lutz T. (2002) A non-essential function for yeast frataxin in iron-sulfur cluster assembly. Hum. Mol. Genet. **11:** 2635–2643
68 Zheng Y., Roberts R. J. and Kasif S. (2002) Genomic functional annotation using co-evolution profiles of gene clusters. Genome Biol. **3:** research 0060.1–0060.9
69 Liberles D. A., Thoren A., von Heijne G. and Elofsson A. (2002) The use of phylogenetic profiles for gene prediction. Current Genomics **3:** 131–137
70 von Mering C., Huynen M., Jaeggi D., Schmidt S., Bork P. and Snel B. (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. **31:** 258–261
71 Koonin E. V., Mushegian A. R. and Bork P. (1996) Non-orthologous gene displacement. Trends Genet. **12:** 334–336
72 Myllykallio H., Lipowski G., Leduc D., Filee J., Forterre P. and Liebl U. (2002) An alternative flavin-dependent mechanism for thymidylate synthesis. Science **297:** 105–107
73 Morett E., Korbel J. O., Rajan E., Saab-Rincon G., Olvera L., Olvera M. et al. (2003) Systematic discovery of analogous enzymes in thiamin biosynthesis. Nat. Biotechnol. **21:** 790–795
74 Snel B., Bork P. and Huynen M. A. (1999) Genome phylogeny based on gene content. Nat. Genet. **21:** 108–110
75 Perna N. T., Plunkett G., 3rd, Burland V., Mau B., Glasner J. D., Rose D. J. et al. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature **409:** 529–533
76 Blattner F. R., Plunkett G., 3rd, Bloch C. A., Perna N. T., Burland V., Riley M. et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science **277:** 1453–1474
77 Buchrieser C., Rusniok C., Kunst F., Cossart P. and Glaser P. (2003) Comparison of the genome sequences of *Listeria monocytogenes* and *Listeria innocua*: clues for evolution and pathogenicity. FEMS Immunol. Med. Microbiol. **35:** 207–213
78 Aravind L., Watanabe H., Lipman D. J. and Koonin E. V. (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes. Proc. Natl. Acad. Sci. USA **97:** 11319–11324
79 Ettema T., van der Oost J. and Huynen M. (2001) Modularity in the gain and loss of genes: applications for function prediction. Trends Genet. **17:** 485–487
80 Fryxell K. J. (1996) The coevolution of gene family trees. Trends Genet. **12:** 364–369
81 Goh C. S., Bogan A. A., Joachimiak M., Walther D. and Cohen F. E. (2000) Coevolution of proteins with their interaction partners. J. Mol. Biol. **299:** 283–293
82 Hughes A. L. and Yeager M. (1999) Coevolution of the mammalian chemokines and their receptors. Immunogenetics **49:** 115–124

83 Pazos F. and Valencia A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng. **14:** 609–614

84 Valencia A. and Pazos F. (2003) Prediction of protein-protein interactions from evolutionary information. Methods Biochem. Anal. **44:** 411–426

85 Ramani A. K. and Marcotte E. M. (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. J. Mol. Biol. **327:** 273–284

86 Ortiz A. R., Kolinski A., Rotkiewicz P., Ilkowski B. and Skolnick J. (1999) Ab initio folding of proteins using restraints derived from evolutionary information. Proteins **Suppl. 3:** 177–185

87 Gobel U., Sander C., Schneider R. and Valencia A. (1994) Correlated mutations and residue contacts in proteins. Proteins **18:** 309–317

88 Pazos F., Helmer-Citterich M., Ausiello G. and Valencia A. (1997) Correlated mutations contain information about protein-protein interaction. J. Mol. Biol. **271:** 511–523

89 Pazos F. and Valencia A. (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins **47:** 219–227

90 Aloy P. and Russell R. B. (2002) Interrogating protein interaction networks through structural biology. Proc. Natl. Acad. Sci. USA **99:** 5896–5901

91 Mendez R., Leplae R., De Maria L. and Wodak S. J. (2003) Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins **52:** 51–67

92 Velculescu V. E., Zhang L., Vogelstein B. and Kinzler K. W. (1995) Serial analysis of gene expression. Science **270:** 484–487

93 Grigoriev A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. Nucleic Acids Res. **29:** 3513–3519

94 Jansen R., Greenbaum D. and Gerstein M. (2002) Relating whole-genome expression data with protein-protein interactions. Genome Res. **12:** 37–46

95 Ge H.,. Liu Z, Church G. M. and Vidal M. (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. Nat. Genet. **29:** 482–486

96 DeRisi J. L., Iyer V. R. and Brown P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278:** 680–686

97 Wu L. F., Hughes T. R., Davierwala A. P., Robinson M. D., Stoughton R. and Altschuler S. J. (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. Nat. Genet **31:** 255–265

98 Eisen M. B., Spellman P. T., Brown P. O. and Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95:** 14863–14868

99 Niehrs C. and Pollet N. (1999) Synexpression groups in eukaryotes. Nature **402:** 483–487

100 Wen X., Fuhrman S., Michaels G. S., Carr D. B., Smith S., Barker J. L. et al. (1998) Large-scale temporal gene expression mapping of central nervous system development. Proc. Natl. Acad. Sci. USA **95:** 334–339

101 Hughes T. R., Marton M. J., Jones A. R., Roberts C. J., Stoughton R., Armour C. D. et al. (2000) Functional discovery via a compendium of expression profiles. Cell **102:** 109–126

102 Noordewier M. O. and Warren P. V. (2001) Gene expression microarrays and the integration of biological knowledge. Trends Biotechnol. **19:** 412–415

103 Teichmann S. A. and Babu M. M. (2002) Conservation of gene co-regulation in prokaryotes and eukaryotes. Trends Biotechnol **20:** 407-10; discussion 410

104 Stuart J. M, Segal E., Koller D. and Kim S. K. (2003) A gene coexpression network for global discovery of conserved genetic modules. Science **302:** 249–255

105 Kelley B. P., Sharan R., Karp R. M., Sittler T., Root D. E., Stockwell B. R. et al. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc. Natl. Acad. Sci. USA **100:** 11394–11399

106 Yanai I., Mellor J. C. and DeLisi C. (2002) Identifying functional links between genes using conserved chromosomal proximity. Trends Genet. **18:** 176–179

107 Jansen R., Lan N., Qian J. and Gerstein M. (2002) Integration of genomic datasets to predict protein complexes in yeast. J Struct Funct Genomics **2:** 71–81

108 Huynen M. A., Snel B., Mering C. and Bork P. (2003) Function prediction and protein networks. Curr. Opin. Cell Biol. **15:** 191–198

109 von Mering C., Krause R., Snel B., Cornell M., Oliver S. G., Fields S. et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature **417:** 399–403

110 Xenarios I., Salwinski L., Duan X. J., Higney P., Kim S. M. and Eisenberg D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. **30:** 303–305

111 Zanzoni A., Montecchi-Palazzi L., Quondam M., Ausiello G., Helmer-Citterich M. and Cesareni G. (2002) MINT: a Molecular INTeraction database. FEBS Lett. **513:** 135–140

112 Bader G. D., Betel D. and Hogue C. W. (2003) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. **31:** 248–250

113 Marcotte E. M., Pellegrini M., Thompson M. J., Yeates T. O. and Eisenberg D. (1999) A combined algorithm for genome-wide prediction of protein function. Nature **402:** 83–86

114 Hill D. P., Blake J. A., Richardson J. E. and Ringwald M. (2002) Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. Genome Res. **12:** 1982–1991

115 Tatusov R. L., Galperin M. Y., Natale D. A. and Koonin E. V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. **28:** 33–36

116 Ogata H., Goto S., Sato K., Fujibuchi W., Bono H. and Kanehisa M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. **27:** 29–34

117 Matte-Tailliez O., Zivanovic Y. and Forterre P. (2000) Mining archaeal proteomes for eukaryotic proteins with novel functions: the PACE case. Trends Genet. **16:** 533–536

118 Jeong H., Mason S. P., Barabasi A. L. and Oltvai Z. N. (2001) Lethality and centrality in protein networks. Nature **411:** 41–42

119 Fraser H. B., Hirsh A. E., Steinmetz L. M., Scharfe C. and Feldman M. W. (2002) Evolutionary rate in the protein interaction network. Science **296:** 750–752

120 Snel B., Bork P. and Huynen M. A. (2002) The identification of functional modules from the genomic association of genes. Proc. Natl. Acad. Sci. USA **99:** 5890–5895

121 Spirin V. and Mirny L. A. (2003) Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. USA **100:** 12123–12128

122 Vazquez A., Flammini A., Maritan A. and Vespignani A. (2003) Global protein function prediction from protein-protein interaction networks. Nat. Biotechnol. **21:** 697–700

123 Goldberg D. S. and. Roth F. P (2003) Assessing experimentally derived interactions in a small world. Proc. Natl. Acad. Sci. USA **100:** 4372–4376

124 Jeong H., Tombor B., Albert R., Oltvai Z. N. and Barabasi A. L. (2000) The large-scale organization of metabolic networks. Nature **407:** 651–654

125 Maslov S. and Sneppen K. (2002) Specificity and stability in topology of protein networks. Science **296:** 910–913

126 Ma H. and Zeng A. P. (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics **19:** 270–277

127 Aloy P. and Russell R. B. (2002) Potential artefacts in protein-interaction networks. FEBS Lett. **530:** 253–254
128 Barabasi A. L. and Albert R. (1999) Emergence of scaling in random networks. Science **286:** 509–512
129 Pastor-Satorras R., Smith E. and Sole R. V. (2003) Evolving protein interaction networks through gene duplication. J. Theor. Biol. **222:** 199–210
130 Paley S. M. and Karp P. D. (2002) Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*. Bioinformatics **18:** 715–724
131 Overbeek R., Larsen N., Pusch G. D., D'Souza M., Selkov E., Jr., Kyrpides N. et al. (2000) WIT**:** integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res. **28:** 123–125
132 Schilling C. H.,. Covert M. W, Famili, I., Church G. M., Edwards J. S. and Palsson B. O. (2002) Genome-scale metabolic model of *Helicobacter pylori* 26695. J. Bacteriol. **184:** 4582–4593
133 Schilling C. H. and Palsson B. O. (2000) Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. J. Theor. Biol. **203:** 249–283
134 Dandekar T., Schuster S., Snel B., Huynen M. and Bork P. (1999) Pathway alignment: application to the comparative analysis of glycolytic enzymes. Biochem. J **343 Pt 1:** 115–124
135 Huynen M. A., Dandekar T. and Bork P. (1999) Variation and evolution of the citric-acid cycle: a genomic perspective. Trends Microbiol. **7:** 281–291
136 Becker K., Schirmer M., Kanzok S. and Schirmer R. H. (1999) Flavins and flavoenzymes in diagnosis and therapy. Methods Mol. Biol. **131:** 229–245
137 Bisbal C., Martinand C., Silhol M., Lebleu B. and Salehzada T. (1995) Cloning and characterization of a RNAse L inhibitor. A new component of the interferon-regulated 2-5A pathway. J. Biol. Chem. **270:** 13308–13317
138 Zimmerman C., Klein K. C., Kiser P. K., Singh A. R., Firestein B. L., Riba S. C. et al. (2002) Identification of a host protein essential for assembly of immature HIV-1 capsids. Nature **415:** 88–92
139 Fromme J. C. and Verdine G. L. (2003) Structure of a trapped endonuclease III-DNA covalent intermediate. EMBO J. **22:** 3461–3471
140 Porello S. L., Cannon M. J. and David S. S. (1998) A substrate recognition role for the [4Fe-4S]2+ cluster of the DNA repair glycosylase MutY. Biochemistry **37:** 6465–6475
141 Strong M., Mallick P., Pellegrini M., Thompson M. J. and Eisenberg D. (2003) Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach. Genome Biol. **4:** R59
142 Kromer W. J. and Arndt E. (1991) Halobacterial S9 operon. Three ribosomal protein genes are cotranscribed with genes encoding a tRNA(Leu), the enolase and a putative membrane protein in the archaebacterium *Haloarcula* (Halobacterium) *marismortui.* J. Biol. Chem. **266:** 24573–24579
143 Santangelo G. M. and Tornow J. (1990) Efficient transcription of the glycolytic gene ADH1 and three translational component genes requires the GCR1 product, which can act through TUF/GRF/RAP binding sites. Mol Cell Biol **10:** 859–862
144 Bork P., Dandekar T., Diaz-Lazcoz Y., Eisenhaber F., Huynen M. and Yuan Y. (1998) Predicting function: from genes to genomes and back. J. Mol. Biol. **283:** 707–725
145 (2001) Creating the gene ontology resource: design and implementation. Genome Res. **11:** 1425–1433
146 Snel B., Lehmann G., Bork P. and Huynen M. A. (2000) STRING**:** a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res. **28:** 3442–3444
147 Chenna R., Sugawara H., Koike T., Lopez R., Gibson T. J., Higgins D. G. et al. (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. **31:** 3497–3500