# Function prediction and protein networks

Martijn A Huynen*†, Berend Snel*‡, Christian von Mering§# and Peer Bork§¶

In the genomics era, the interactions between proteins are at the center of attention. Genomic-context methods used to predict these interactions have been put on a quantitative basis, revealing that they are at least on an equal footing with genomics experimental data. A survey of experimentally confirmed predictions proves the applicability of these methods, and new concepts to predict protein interactions in eukaryotes have been described. Finally, the interaction networks that can be obtained by combining the predicted pair-wise interactions have enough internal structure to detect higher levels of organization, such as ‘functional modules’.

**Addresses**
*Nijmegen Center for Molecular Life Sciences, Center for Molecular and Biomolecular Informatics, Toernooiveld 1, 6525 ED Nijmegen,
The Netherlands
†e-mail: huynen@cmbi.kun.nl
‡e-mail: snel@cmbi.kun.nl
§European Molecular Biology Laboratory, Meyerhofstrasse 1,
69117 Heidelberg, Germany
#e-mail: mering@embl-heidelberg.de
¶e-mail: bork@embl-heidelberg.de

**Current Opinion in Cell Biology** 2003, **15**:191–198

This review comes from a themed issue on
Cell regulation
Edited by Pier Paolo di Fiore and Pier Giuseppe Pelicci

0955-0674/03/$ – see front matter

**DOI 10.1016/S0955-0674(03)00009-7**

## Introduction

Genome sequencing provides us with an abundance of genes whose functions are not determined experimentally and have to be predicted by bioinformatics. The classic tool to do so, homology detection, is mainly suited to predict the molecular function of a protein. Because we have complete genome sequences we would also like to know proteins’ functions at a higher level [1], for example the pathway or complex a protein belongs to.

Parallel with experimental developments to determine protein–protein interactions (e.g. [2,3]), bioinformatics supplies us with a growing number of so-called genomic-context methods that exploit the genome sequences themselves to predict such interactions. These methods use the fact that the genes of functionally interacting proteins tend to be associated with each other on genomes. Originally gene fusion [4,5], the conservation of gene order [6,7] and co-occurrence of genes among sequenced genomes [8,9] were proposed (Figure 1), and subsequently also methods that use sequence information of the proteins themselves [10], or that include information from shared regulatory elements [11,12•] have been used. The principles of the above-mentioned methods have been the subject of many reviews already [13–16]. We will therefore focus on their practical applicability. First, we will review how well they perform and survey the predictions that have actually been experimentally verified. Subsequently, we will review how these extensive lists of protein–protein interactions give rise to biological networks and what they mean for biology. Finally, we discuss new principles for interaction prediction from genomic contexts that are specifically applicable to eukaryotes.

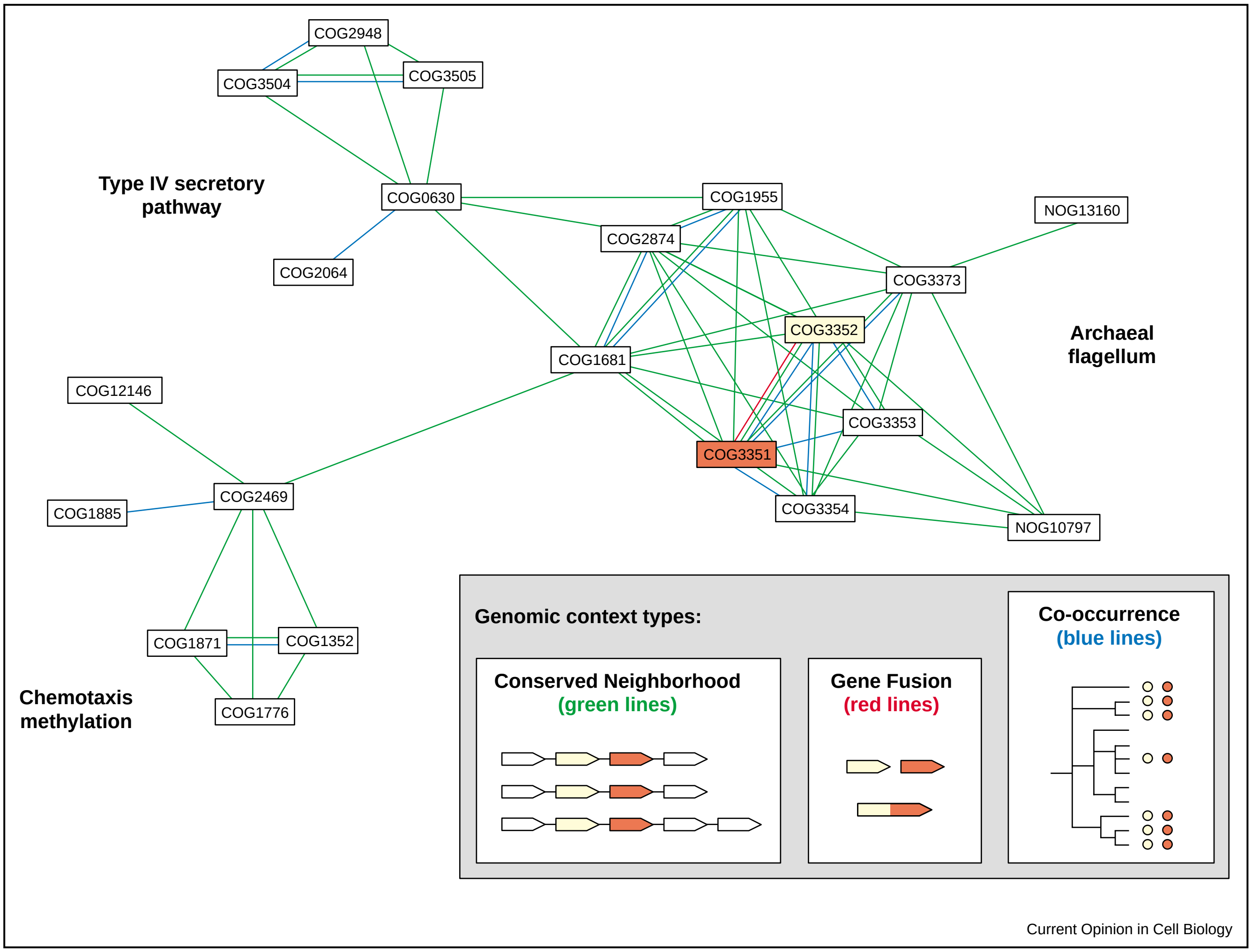
## Performance and applications of context methods

### Accuracy

Recent, large-scale analyses confirm earlier results [17] that the reliability of genomic-context methods to predict functional interactions is high, specifically for gene fusion (72%) [18,19] and gene-order conservation (80%) (Figure 2) [19,20]. One should keep in mind, however, that the benchmarks that are used are often quite general, for example having a similar set of SWISS-PROT keywords [21], or falling on the same metabolic map in KEGG (Kyoto Encyclopedia of Genes and Genomes; http://genome.ad.jp/kegg) [22]. The availability of yeast two-hybrid or identification of protein complexes by mass spectrometry data should allow more-systematic benchmarking of the genomic-context methods for the prediction of physical interaction, were it not that these data themselves are not always of high quality.

By comparing experimental genomics techniques, mRNA-correlated expression, and genomic-context predictions to a classic set of ‘trusted’ physical interactions that were obtained from YPD (Yeast Protein Database) or MIPS (Munich Information Center for Protein Sequences; http://mips.gsf.de/), it was shown that genomic-context predictions actually had both a higher coverage (7.7%) and a higher accuracy (5.3%) not only than mRNA co-expression, but also than direct experimental techniques like yeast two-hybrid or high-throughput mass spectrometric protein-complex identification (HMS-PCI). As the combination of genomic-context data with experimental data increases the fraction of true positives, genomic context can also be used as a filter, to improve the quality of the experimental data [23•,24], albeit at a loss of coverage.

Figure 1



Functional modules in a genomic-context network. Shown are orthologous groups linked via genomic context either directly or indirectly (via one other orthologous group) to COG1681 (Archaeal flagellin). The three types of context evidence — gene order (green), gene fusion (red) and co-occurrence (blue) — are illustrated in the inset and are indicated by separate lines in the network. The three subclusters (type IV secretory pathway, Archaeal flagella and chemotaxis methylation) are only linked to each other through either one orthologous group (COG0630) or one link (between COG2469 and COG1681), yet within each subcluster the orthologous groups are densely linked. The subclusters correspond to separate functional systems. Automatic function prediction for orthologous groups falling within a cluster can be done by transferring the highest common denominator within one cluster to that group; for example, the hypothetical orthologous group COG3373 is predicted to be part of the flagellum, whereas COG2469 is predicted to function in methylation in the regulation of chemotaxis.

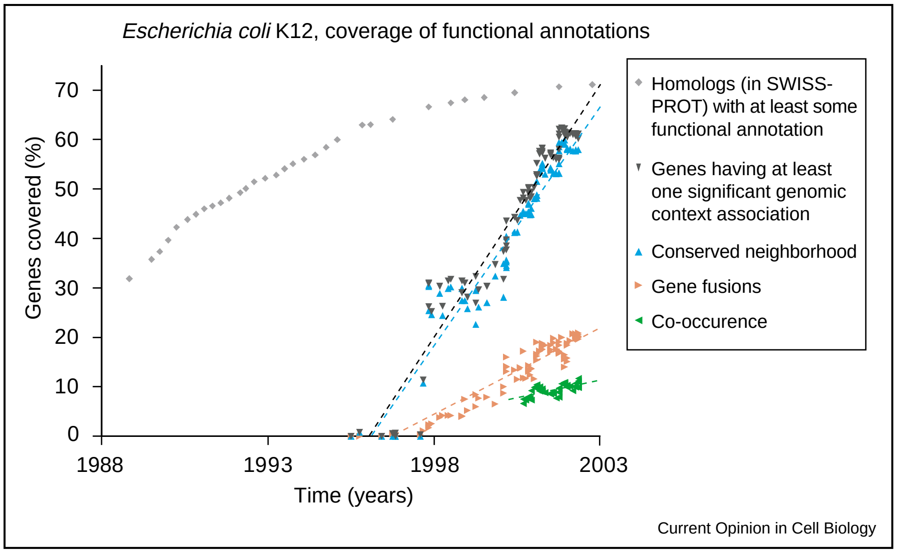
## Comparisons with homology detection and genomic coverage

Classic, homology-based function prediction and genomic-context-based function prediction are complementary, both in the type of functional information they predict (molecular function, versus functional interaction) as well as in the type of information they use (the protein itself, versus its context in the genome).

We analyzed how their predictive potential has improved with the increase of reported genomes and of experimentally determined functions. For one reference genome, *Escherichia coli*, we determined two aspects: firstly, for how many of its proteins we can detect homologs with a known molecular function in various releases of the SWISS-PROT protein database [21]; and secondly, for how many of its proteins we can find significant context information (Figure 2). The plots show a clearly increasing but saturating trend for homology detection, whereas the increase in context detection is almost linear, albeit with a slight saturation in the last year. On the basis of the extrapolation of these curves, genomic context is expected to pass homology detection in terms of coverage in 2003.

Presently we can predict with 80% confidence functional links for the majority of the proteome of prokaryotes

**Figure 2**



Coverage of homology-based methods and of context-based methods for function prediction. Coverage of homology methods was determined by comparing the proteins encoded in the reference genome, *E. coli*, with archived releases of SWISS-PROT dating back to 1988 from which the proteins without functional information were removed (Smith-Waterman searches, e-value $< 0.01$). The coverage of genomic context methods is given at an estimated average accuracy of 80%; the three types of evidence are indicated separately.

(64% in *Mycoplasma genitalium* and 60% in *E. coli*) and for a substantial fraction of the proteome of the eukaryote *Saccharomyces cerevisiae* (26%). It should be noted that some hypothetical proteins with a significant genomic context are only linked to other hypothetical proteins. Links between hypothetical proteins cannot be used for function prediction, but they are relevant because they provide information about the topology of the network of interactions in a cell (see below). Using genomic context we can thus already obtain a view on the network of interactions within a cell, even if we do not know or cannot predict the functions of its individual elements.

## Experimentally verified context predictions

Real applicability of genomic-context methods can, in the long run, only be established by experiments based on their predictions. We identified 13 cases where functional interactions and function were predicted to a varying level of specificity and either published before the experimental verification or published with it (Table 1). In these

**Table 1**

**Experimental verification of context predictions.**

| Protein/gene | Context | Type of interaction | Function | References |
|---|---|---|---|---|
| Mt-Ku | Gene order | Physical interaction | Double-stranded-DNA repair | [46] |
| GlnK | Gene order | Physical interaction | Signal transduction for ammonium transport | [57,58] |
| PH0272 | Gene order | Metabolic pathway | Methylmalonyl-CoA racemase | [45] |
| PrpD | Gene order | Metabolic pathway | 2-Methylcitrate dehydratase | [17,59] |
| arok | Gene order | Metabolic pathway | Shikimate kinase | [60] |
| ComB | Gene order | Metabolic pathway | 2-Phosphosulfolactate phosphatase | [61] |
| Yfh1 | Co-occurrence | Process | Iron–sulfur protein maturation | [27,28] |
| YchB | Co-occurrence | Metabolic pathway | Terpenoid synthesis | [62] |
| SmpB | Co-occurrence | Process | Trans-translation | [8,63] |
| ThyX | Complement | Enzymatic activity | Thymidilate synthase* | [14,64] |
| Prx | Fusion | Pathway | Peroxiredoxin | [65] |
| YgbB | Fusion/gene order | Metabolic pathway | Terpenoid synthesis | [66] |
| SelR | Fusion/gene order/co-occurrence | Enzymatic activity | Methionine sulfoxide reductase | [14,67,68] |

In all cases genomic context was used to predict a functional interaction between proteins, and this interaction was subsequently experimentally verified. In the cases where more than one reference is given, the functional link was published separately and before the experimental verification.
* In a variation of using the phylogenetic distribution of genes to predict functional interaction, a complementary distribution of two orthologous groups was used to predict that they have the same enzymatic function.

cases gene fusion, gene-order conservation, and gene co-occurrence have been used successfully to predict new functional interactions, with gene-order conservation contributing the largest share.

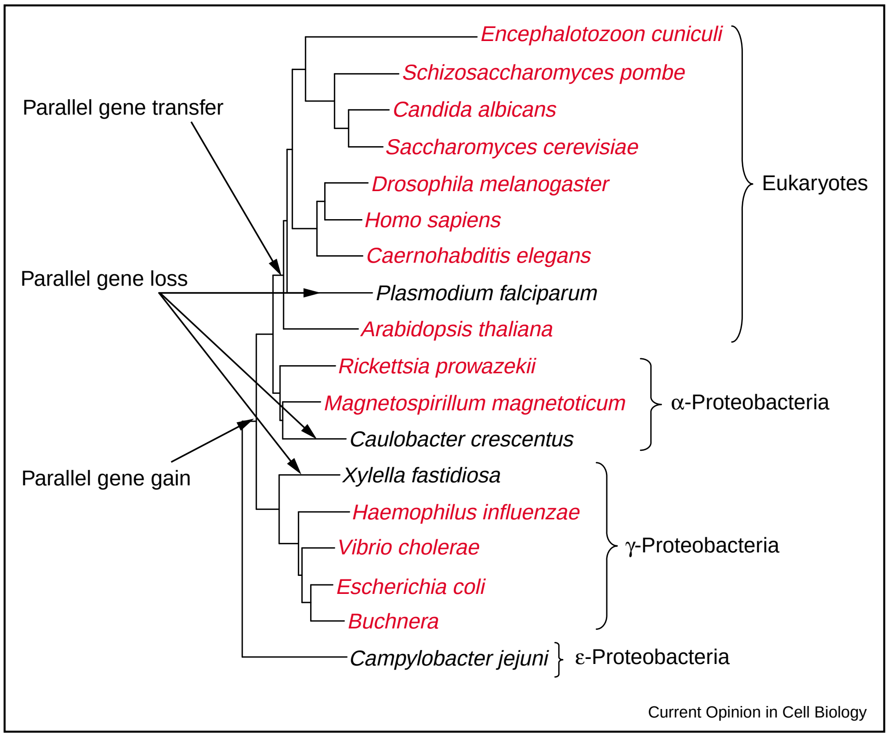
Note that using genomic-context methods to design an experiment is not that trivial because the leads are not very specific. The methods do not predict what the type of interaction between the proteins is: it could for example be regulatory, physical or being part of the same pathway or process (Table 1) [17]; nor do they tell you, for example in the case of a metabolic pathway, where in that pathway to place the hypothetical protein.

One way to increase the prediction specificity is to include the degree to which the genomic context is conserved. The stronger the evolutionary conservation of a genomic-context pattern (e.g. the more often that the genes are neighbors), the more likely not only that the proteins functionally interact, but also that they interact in the most direct way; that is, by being involved in the same reaction and forming a protein complex [25]. Generally, however, it is left to the researcher to combine the genomic-context information with data on homology relations and with data on, for example, the phenotypic effects of deletion of the protein or on missing steps in a pathway, to make a specific, testable prediction about the protein's function.

## A case story of a successful context-based prediction: frataxin

An example of such a successful prediction about a protein's involvement in a biological process is that of the well-known disease gene frataxin. The protein's function has remained elusive despite the fact that its gene was identified in 1996 as being responsible for the neurodegenerative disorder Friedreich's ataxia [26]. The main hypothesis, based on the observation of the accumulation of iron in mitochondria in the protein's absence, was that it is directly involved in maintaining iron homeostasis. Genomic-context analysis indicated that frataxin has the same evolutionary history, involving at least three cases of parallel gene loss, as two chaperones involved in the assembly of iron–sulfur clusters in proteobacteria and mitochondria — HscA/Ssq1 and HscB/Jac1 (Figure 3) [27]. This has lead to the hypothesis that frataxin is involved in iron–sulfur cluster assembly on proteins as well [27], for which there is now a rapidly increasing body

**Figure 3**



Parallel evolution of the genes *frataxin* and *HscB/Jac1*. The species that have genes for frataxin and HscB/Jac1 in their genome are shown in red. Both genes are only present in the eukaryotes and in the α-, β- and γ-proteobacteria. They probably originated at the onset of the α-, β- and γ-proteobacteria, as the genes are absent from other prokaryotes, here represented by *Campylobacter jejuni*. Subsequently, they have both been transferred to the eukaryotes with the origin of the mitochondria from an a-proteobacterium. The genes have been lost together at least three times, from the prokaryote *Xylella fastidiosa*, from a set of α-proteobacteria represented by *Caulobacter crescentus*, and from the eukaryote *Plasmodium falciparum*.

of experimental evidence [28–30]. The more specific hypothesis, that it is involved in the same subprocess as HscB/Jac1, awaits confirmation.

## Interaction networks

### From pair-wise interactions to networks

The many pair-wise interactions that are proposed on the basis of genomic-context analyses, or that are present in metabolic maps or experimental approaches to large-scale identification of protein–protein interactions, present us with networks of interactions in which the large majority of proteins are linked to each other, either directly or indirectly.

To study the intrinsic properties of these networks and to be able to compare them, some general statistics are measured, for example the average minimal path length (the number of intermediate links) between any two nodes, the clustering co-efficient (see below) and the distribution of the number of connections per node. Comparing the number of connections per node with data on the lethality of mutations indicates that the larger the number of physical interactions a protein has, the higher the probability that it is essential for survival [31•]. Comparing the number of connections per protein with its evolutionary rate has also revealed that the more physical interactions a protein has, the lower its rate of evolution, not because it is relatively essential for the species, but rather because a larger part of the protein is involved in interactions with other proteins [32••]. Furthermore, and consistent with this, proteins that physically interact with each other tend to evolve at similar rates [32••].

Determination of metabolic-network statistics on the basis of genome annotations also allows cross-species comparisons of network topologies [33]. The topology of networks puts constraints on the process by which they could have evolved [33]. It has, however, not been shown conclusively that the topology of the network is subject to selection and therewith of value to the understanding of a cell [34•]. Furthermore, apparently interesting patterns in the networks [35], for example the tendency of highly connected nodes not to be linked to each other, can reflect systematic biases in the experimental technique used to detected the links [36•], rather than patterns in the underlying biology.

### Functional modules

An interesting aspect of the network topologies that does have biological relevance and that can be used for function prediction is the detection of higher levels of functional organization, or 'functional modules' [37,38••,39•] — sets of proteins that together function in a single process (Figure 1). The presence of such modules can be deduced when networks have a high clustering coefficient. This clustering coefficient is the fraction of cases where, if a protein (A) is linked to two other proteins (B and C), the latter two proteins also have a direct link to each other.

In a genomic-context network, the clustering coefficient was observed to be much higher (0.6) than that of a random network with the same number of nodes and connections (0.005) [38••]. Identifying the modules in a network with a high clustering coefficient basically involves 'cutting-up' the network in the less densely clustered areas. The modules in a genomic-context network tend to be functionally homogeneous; that is, they contain proteins that are part of a single pathway [38••]. Delineating the cluster-structure thus also facilitates protein-function prediction, as the highest common denominator of the proteins with known function can automatically be transferred to a hypothetical protein in that cluster (Figure 1). Similarly, the network of metabolites as extracted from the metabolic pathway database WIT (What Is There; http://wit.mcs.anl.gov/WIT2/) [40] can be shown to have a modular organization [41]. Furthermore, functional modules have also been extracted from gene-expression data, although without explicitly deriving the network topology [42,43••].

The search for functional modules immediately raises issues, not only with the objectivity of pathway databases like KEGG and WIT, but also with our definitions of biological processes and to what extent boundaries can be drawn between them. On the one hand it holds the promise to generate functional module definitions that are independent of specific experimental conditions, including the species being studied, but purely based on comparative genome analysis. On the other hand it is questionable to what extent a species-independent pathway definition makes sense at all. It denies the variation and evolution of pathways, one of the most interesting results to come out of comparative genome analysis. A middle ground here would be to compare sets of genomes from a single taxon, to identify taxon-specific pathways [44].

## Genomic context in eukaryotes

Gene-order conservation is the most powerful genomic context technique in prokaryotes [17] (Table 1, Figure 2). It can also be used for the functional characterization of those genes in eukaryotes that have orthologs in bacteria, as was shown for the human methylmalonyl-CoA racemase [45]. In *S. cerevisiae*, 1302 proteins (21% of its proteome) have orthologs with conserved gene-order in prokaryotes and for an increasing number of proteins that were originally regarded as purely eukaryotic, homologs with similar functions in prokaryotes are being detected (e.g. [46]). Thus there is still a large potential in using prokaryotic gene-order conservation for protein-function prediction in eukaryotes.

Nevertheless, there are some observations that point to the potential of using gene-order in eukaryotes themselves. One observation is the presence of polycistronic RNA transcripts in nematodes which were recently estimated to contain 15% of the genes [47•]. The evidence for functional interactions between the proteins encoded by such polycistronic transcripts is, however, anecdotal. A second observation is the chromosomal clustering of co-expressed genes in *Caenorhabditis elegans* at a higher level than that of the polycistronic RNAs [48]. Also in *Homo sapiens* highly expressed genes are clustered in the genome [49]. This pattern, however, appears to be caused by the clustering of housekeeping genes [50••], and therewith to give only a very weak signal for function prediction.

Weak signals from genomics data can generally be enhanced by exploiting evolutionary conservation. A recent example concerns gene co-expression: whereas for the co-expressed genes within yeast and worm the fraction of physically interacting proteins was 22% and 32%, respectively, for conservedly co-expressed genes the fraction rose to 89% [51•]. With expression data on more species becoming available, conservation of co-expression is a promising technique for function-prediction in eukaryotes.

Finally, combining gene-order conservation with gene co-expression points to the potential of divergently transcribed, co-regulated genes. In *S. cerevisiae*, co-regulated, divergently transcribed genes have a relatively high chance of having conserved gene order in *Candida albicans* compared with those that are not co-regulated [25,52•]. These conserved gene pairs include not only well-known cases of functionally interacting genes such as the histone gene pairs H2A–H2B and H3–H4, but also experimentally uncharacterized ones like a hexose permease (YJL219W) and an a glucosidase (YJL221C).

## Conclusions and future challenges

Parallel to large-scale experimental efforts, genomic-context methods are giving us a new view on function, one that focuses on the functional interactions between proteins and on the functional modules that they form. Paradoxically, the challenges in increasing the coverage and accuracy of these genomic-context prediction tools are partly on the experimental side. We not only need more experimental verification of the specific predictions that have been made in various context papers [17,53], or that can be retrieved from the web-servers (Box 1). We also need more high-quality interaction data from genomics to provide protein–protein interaction benchmarks as well as more eukaryotic genome sequences and other types of genomics data to fully apply the tools of comparative genomics. Further integration of genomic context in experimental genomics is also invaluable to increase the accuracy of the results [23•].

On the bioinformatics side we face major technical hurdles in developing a higher-resolution orthology prediction than is currently available in the state-of-the-art COG (Clusters of Orthologous Groups; http://www.ncbi.nlm.nih.gov/COG/) database [54], specifically with the rampant gene duplication in eukaryotes. On the network side, taking a step further than functional modules, recent studies have started to delineate 'network motifs' in experimentally determined regulatory networks [55,56]. If we can observe specific motifs in genomic-context networks and are able to link them to specific types of functions, functional modules are poised to be of the same importance for our understanding of cellular systems as protein domains have proven to be for our understanding of proteins.

**Box 1** Web accessibility.

Web servers for the prediction of protein–protein interactions based on genomic context that are reasonably up to date in terms of the included genomes are: the COG database itself (http://www.ncbi.nlm.nih.gov/COG/) [54]; predictome (http://predictome.bu.edu, containing the same genomes as the COG database) [17]; and STRING (http://www.bork.embl-heidelberg.de/STRING/). STRING has the largest coverage in terms of published genomes, and integrates the three types of genomic context into a single score function, increasing the number of orthologous groups for which function predictions can be made [19].

## References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

• of special interest
•• of outstanding interest

1. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y: **Predicting function: from genes to genomes and back**. *J Mol Biol* 1998, **283**:707-725.

2. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al.*: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae***. *Nature* 2000, **403**:623-627.

3. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM *et al.*: **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature* 2002, **415**:141-147.

4. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein–protein interactions from genome sequences**. *Science* 1999, **285**:751-753.

5. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events**. *Nature* 1999, **402**:86-90.

6. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **Use of contiguity on the chromosome to predict functional coupling**. *In Silico Biol* 1998, **1**:93-108.

7. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact**. *Trends Biochem Sci* 1998, **23**:324-328.

8. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles**. *Proc Natl Acad Sci USA* 1999, **96**:4285-4288.

9. Huynen MA, Bork P: **Measuring genome evolution**. *Proc Natl Acad Sci USA* 1998, **95**:5849-5856.

10. Pazos F, Valencia A: **Similarity of phylogenetic trees as indicator of protein-protein interaction**. *Protein Eng* 2001, **14**:609-614.

11. McGuire AM, Hughes JD, Church GM: **Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes**. *Genome Res* 2000, **10**:744-757.

12. van Nimwegen E, Zavolan M, Rajewsky N, Siggia ED: **Probabilistic**
• **clustering of sequences: inferring new bacterial regulons by comparative genomics**. *Proc Natl Acad Sci USA* 2002, **99**:7323-7328.

A novel Bayesian clustering methodology for promotor prediction by comparative genome analysis, leading to the prediction of around 100 new regulons in *E. coli*. The paper argues that there are inherent mathematical limits to our ability to discover promoters in sequence data and that comparative genomic information may be crucial in this respect.

13. Marcotte EM: **Computational genetics: finding protein function by nonhomology methods**. *Curr Opin Struct Biol* 2000, **10**:359-365.

14. Galperin MY, Koonin EV: **Who's your neighbor? New computational approaches for functional genomics**. *Nat Biotechnol* 2000, **18**:609-613.

15. Valencia A, Pazos F: **Computational methods for the prediction of protein interactions**. *Curr Opin Struct Biol* 2002, **12**:368-373.

16. Huynen M, Snel B, Lathe W, Bork P: **Exploitation of gene context**. *Curr Opin Struct Biol* 2000, **10**:366-370.

17. Huynen M, Snel B, Lathe W III, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences**. *Genome Res* 2000, **10**:1204-1210.

18. Yanai I, Derti A, DeLisi C: **Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes**. *Proc Natl Acad Sci USA* 2001, **98**:7940-7945.

19. von Mering C, Huynen MA, Jaeggi D, Schmidt S, Bork P, Snel B: **STRING: a database of predicted functional associations between proteins**. *Nucleic Acids Res* 2003, **31**:258-261.

20. Yanai I, Mellor JC, DeLisi C: **Identifying functional links between genes using conserved chromosomal proximity**. *Trends Genet* 2002, **18**:176-179.

21. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000**. *Nucleic Acids Res* 2000, **28**:45-48.

22. Kanehisa M, Goto S, Kawashima S, Nakaya A: **The KEGG databases at GenomeNet**. *Nucleic Acids Res* 2002, **30**:42-46.

23. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork
• P: **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* 2002, **417**:399-403.

The paper summarizes protein interaction data in the yeast *S. cerevisiae*, and performs a rough quality assessment through several independent benchmarks. Interaction data are found to be largely complementary — having a surprisingly small overlap. The best accuracy is found for combined datasets.

24. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: **Protein function in the post-genomic era**. *Nature* 2000, **405**:823-826.

25. Huynen MA, Snel B: Exploiting the variations in the genomic associations of genes to predict pathways and reconstruct their evolution. In *Frontiers in Computational Genomics*. Edited by Galperin MY, Koonin EV. Norfolk: Caisters Academic Press; 2003:145-166.

26. Campuzano V, Montermini L, Molto MD, Pianese L, Cossee M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A *et al.*: **Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion**. *Science* 1996, **271**:1423-1427.

27. Huynen MA, Snel B, Bork P, Gibson TJ: **The phylogenetic distribution of frataxin indicates a role in iron-sulfur cluster protein assembly**. *Hum Mol Genet* 2001, **10**:2463-2468.

28. Muhlenhoff U, Richhardt N, Ristow M, Kispal G, Lill R: **The yeast frataxin homolog Yfh1p plays a specific role in the maturation of cellular Fe/S proteins**. *Hum Mol Genet* 2002, **11**:2025-2036.

29. Chen OS, Hemenway S, Kaplan J: **Inhibition of Fe-S cluster biosynthesis decreases mitochondrial iron export: Evidence that Yfh1p affects Fe-S cluster synthesis**. *Proc Natl Acad Sci USA* 2002, **99**:12321-12326.

30. Duby G, Foury F, Ramazzotti A, Herrmann J, Lutz T: **A non-essential function for yeast frataxin in iron-sulfur cluster assembly**. *Hum Mol Genet* 2002, **11**:2635-2643.

31. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and**
• **centrality in protein networks**. *Nature* 2001, **411**:41-42.

This fresh and original study integrates data from yeast two-hybrids and lethal mutations in yeast to convincingly show that mutations in proteins central in the network of proteins interactions are indeed more lethal, as predicted by theory of scale-free networks.

32. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW:
•• **Evolutionary rate in the protein interaction network**. *Science* 2002, **296**:750-752.

A rigorous statistical analysis of variations in evolutionary rates and their causes. Notably the authors can reject the hypothesis that the slow evolutionary rate of highly connected proteins is because these proteins are relatively essential for the organism. Instead it appears to be caused by co-evolutionary constraints that are caused by protein–protein interactions.

33. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks**. *Nature* 2000, **407**:651-654.

34. Wolf YI, Karev G, Koonin EV: **Scale-free networks in biology: new**
• **insights into the fundamentals of evolution?** *Bioessays* 2002, **24**:105-109.

The authors give a clear exposition on findings on biological networks. They deem it trivial that the majority of the biological networks are scale-free networks because this is easily explained through slow evolutionary growth by preferential attachment. They conclude that most network studies have not yet 'crossed the line between abstract discourse and actual research tools and techniques'.

35. Maslov S, Sneppen K: **Specificity and stability in topology of protein networks**. *Science* 2002, **296**:910-913.

36. Aloy P, Russell RB: **Potential artefacts in protein-interaction**
• **networks**. *FEBS Lett* 2002, **530**:253-254.

A detailed analysis of the asymmetry in the yeast two-hybrid experiments between situations where proteins act as 'prey' versus where they act as 'bait'. In the network of interactions, the most highly connected nodes (hubs) turn out to have a strong preference for being 'bait' rather than 'prey'. This systematic experimental bias explains why the hubs do not tend to interact with each other, and no unproven selection effect, as proposed by Maslov and Sneppen (2002) [35], is required to explain this bias in the network topology.

37. Koonin EV, Wolf YI, Aravind L: **Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach**. *Genome Res* 2001, **11**:240-252.

38. Snel B, Bork P, Huynen MA: **The identification of functional**
•• **modules from the genomic association of genes**. *Proc Natl Acad Sci USA* 2002, **99**:5890-5895.

The network properties of a genomic-context network (based on conservation of gene order). The network is shown to be a scale-free network with a high clustering coefficient. Subclusters are separated from each other by 'cutting' at orthologous groups that are locally the only link between them. The subclusters are functionally homogeneous, and can be regarded as functional modules. Orthologous groups that link separate subclusters to each other tend to be multifunctional.

39. Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov
• RL, Szekely LA, Koonin EV: **Connected gene neighborhoods in prokaryotic genomes**. *Nucleic Acids Res* 2002, **30**:2212-2223.

This work outlines an approach to find clusters of gene arrays through connected trails of conserved gene pairs. It is successful in obtaining functionally coherent sets of genes. The authors coin the term genomic hitchhiking for those genes that associate themselves to other genes for what, to the human eye, seem unclear reasons.

40. Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E Jr, Kyrpides N, Fonstein M, Maltsev N, Selkov E: **WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction**. *Nucleic Acids Res* 2000, **28**:123-125.

41. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks**. *Science* 2002, **297**:1551-1555.

42. Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ: **Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters**. *Nat Genet* 2002, **31**:255-265.

43. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N:
•• **Revealing modular organization in the yeast transcriptional network**. *Nat Genet* 2002, **31**:370-377.
The authors rightly point out that conventional hierarchical clustering of co-expression data often fails, because genes can play a role in multiple cellular processes and their common regulatory element can only be detected in a subset of experiments. They have therefore developed a method to detect genes that are co-expressed under a subset of conditions. This method finds a comprehensive set of overlapping 'transcriptional modules' that promise to be very useful for function prediction.

44. Snel B, Bork P, Huynen M: **Conservation of gene co-regulation in prokaryotes and eukaryotes**. *Trends Biotechnol* 2002, **20**:410.

45. Bobik TA, Rasche ME: **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* 2001, **276**:37194-37198.

46. Weller GR, Kysela B, Roy R, Tonkin LM, Scanlan E, Della M, Devine SK, Day JP, Wilkinson A, di Fagagna F *et al.*: **Identification of a DNA nonhomologous end-joining complex in bacteria**. *Science* 2002, **297**:1686-1689.

47. Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-
• Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M *et al.*: **A global analysis of *Caenorhabditis elegans* operons**. *Nature* 2002, **417**:851-854.
Direct experimental identification of the majority of polycistronic transcription units in the genome of *C. elegans*, yielding an estimate of 13%–15% of genes being operon members. Case studies hint at functional links within such operons — but the overall signal, if any, appears much weaker than in prokaryotes.

48. Roy PJ, Stuart JM, Lund J, Kim SK: **Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans***. *Nature* 2002, **418**:975-979.

49. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA *et al.*: **The human transcriptome map: clustering of highly expressed genes in chromosomal domains**. *Science* 2001, **291**:1289-1292.

50. Lercher MJ, Urrutia AO, Hurst LD: **Clustering of housekeeping
•• genes provides a unified model of gene order in the human genome**. *Nat Genet* 2002, **31**:180-183.
The authors demonstrate through rigorous statistical analysis that the at-first-glance surprising finding of clustered highly expressed genes on human chromosomes can be completely explained through co-localization of housekeeping genes.

51. Teichmann S, Babu M: **Conservation of gene co-regulation
• in prokaryotes and eukaryotes**. *Trends Biotechnol* 2002, **20**:407.
To our knowledge, this is the first comparative (i.e. between species) co-expression analysis. The results appear to suggest that very little co-expression is conserved, but those gene pairs that are conservedly co-expressed invariably are part of the same protein complex.

52. Hurst LD, Wiliams EJB, Pal C: **Natural selection promotes the
• conservation of linkage of co-expressed genes**. *Trends Genet* 2002, **18**:604-606.
First demonstration (together with Huynen and Snel [2003] [25]) that divergently transcribed, co-expressed genes in *S. cerevisiae tend to have conserved gene order in C. albicans. Such conserved gene order of co-expressed genes hints at the usage of gene-order conservation for function prediction in fungi.*

53. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV: **Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context**. *Genome Res* 2001, **11**:356-372.

54. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes**. *Nucleic Acids Res* 2001, **29**:22-28.

55. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al.*: **Transcriptional regulatory networks in *Saccharomyces cerevisiae***. *Science* 2002, **298**:799-804.

56. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs: simple building blocks of complex networks**. *Science* 2002, **298**:824-827.

57. Thomas G, Coutts G, Merrick M: **The glnKamtB operon. A conserved gene pair in prkaryotes**. *Trends Genet* 2000, **16**:11-14.

58. Coutts G, Thomas G, Blakey D, Merrick M: **Membrane sequestration of the signal transduction protein GlnK by the ammonium transporter AmtB**. *EMBO J* 2002, **21**:536-545.

59. Horswill AR, Escalante-Semerena JC: ***In vitro* conversion of propionate to pyruvate by *Salmonella enterica* enzymes: 2-methylcitrate dehydratase (PrpD) and aconitase enzymes catalyze the conversion of 2-methylcitrate to 2-methylisocitrate**. *Biochemistry* 2001, **40**:4703-4713.

60. Daugherty M, Vonstein V, Overbeek R, Osterman A: **Archaeal shikimate kinase, a new member of the GHMP-kinase family**. *J Bacteriol* 2001, **183**:292-300.

61. Graham DE, Graupner M, Xu H, White RH: **Identification of coenzyme M biosynthetic 2-phosphosulfolactate phosphatase. A member of a new class of Mg(2+)-dependent acid phosphatases**. *Eur J Biochem* 2001, **268**:5176-5188.

62. Luttgen H, Rohdich F, Herz S, Wungsintaweekul J, Hecht S, Schuhr CA, Fellermeier M, Sagner S, Zenk MH, Bacher A *et al.*: **Biosynthesis of terpenoids: YchB protein of *Escherichia coli* phosphorylates the 2-hydroxy group of 4-diphosphocytidyl-2C-methyl-D-erythritol**. *Proc Natl Acad Sci USA* 2000, **97**:1062-1067.

63. Karzai AW, Susskind MM, Sauer RT: **SmpB, a unique RNA-binding protein essential for the peptide-tagging activity of SsrA (tmRNA)**. *EMBO J* 1999, **18**:3793-3799.

64. Myllykallio H, Lipowski G, Leduc D, Filee J, Forterre P, Liebl U: **An alternative flavin-dependent mechanism for thymidylate synthesis**. *Science* 2002, **297**:105-107.

65. Rouhier N, Gelhaye E, Sautiere PE, Brun A, Laurent P, Tagu D, Gerard J, de Fay E, Meyer Y, Jacquot JP: **Isolation and characterization of a new peroxiredoxin from poplar sieve tubes that uses either glutaredoxin or thioredoxin as a proton donor**. *Plant Physiol* 2001, **127**:1299-1309.

66. Herz S, Wungsintaweekul J, Schuhr CA, Hecht S, Luttgen H, Sagner S, Fellermeier M, Eisenreich W, Zenk MH, Bacher A *et al.*: **Biosynthesis of terpenoids: YgbB protein converts 4-diphosphocytidyl-2C-methyl-D-erythritol 2-phosphate to 2C-methyl-D-erythritol 2,4-cyclodiphosphate**. *Proc Natl Acad Sci USA* 2000, **97**:2486-2490.

67. Huynen M, Snel B, Lathe W III, Bork P: **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences**. *Genome Res* 2000, **10**:1204-1210.

68. Kryukov GV, Kumar RA, Koc A, Sun Z, Gladyshev VN: **Selenoprotein R is a zinc-containing stereo-specific methionine sulfoxide reductase**. *Proc Natl Acad Sci USA* 2002, **99**:4245-4250.