FEBS Letters 582 (2008) 1251–1258

Minireview

Predicting biological networks from genomic data

Eoghan D. Harrington[a], Lars J. Jensen[a,b], Peer Bork[a,c,*]

[a] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany
[b] Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, Panum Institutte, Blegdamsvej 3b, DK-2200 Copenhagen N, Denmark
[c] Max Delbrück Centre for Molecular Medicine, Berlin-Buch, Robert-Rössle-Strasse 10, D-13092 Berlin, Germany

Received 8 February 2008; accepted 13 February 2008

Available online 21 February 2008

Edited by Robert B. Russell and Patrick Aloy

Abstract **Continuing improvements in DNA sequencing technologies are providing us with vast amounts of genomic data from an ever-widening range of organisms. The resulting challenge for bioinformatics is to interpret this deluge of data and place it back into its biological context. Biological networks provide a conceptual framework with which we can describe part of this context, namely the different interactions that occur between the molecular components of a cell. Here, we review the computational methods available to predict biological networks from genomic sequence data and discuss how they relate to high-throughput experimental methods.**

Keywords: Prediction; Genomic data; Network; Biological

1. Introduction

As a result of breakthroughs in genome sequencing, we now have access to a huge amount of genomic data from a diverse selection of organisms and environments. However, in order to realise the full potential of this resource we have to be able to convert genomic sequence data into biological knowledge. The first step in this process involves the prediction of genes, which in turn permits some level of functional annotation using domain predictions, homology or orthology. In essence this produces a 'parts list' of genes for that genome, however in order to understand the complexity encoded within we need to uncover which of these parts function together and how. This process is best conceptualised as building a network: the gene predictions give us the parts list (nodes) and the next step is to discover the interactions (edges) connecting them. The most direct way to do this is to experimentally determine which genes interact and how, but this is an expensive process in terms of time and resources and may not be possible for some organisms. Therefore, computational methods have been developed to predict functional interactions between genes, either based on genomic sequence alone or in combination with experimental data. In this review, we will discuss the role of computational methods in the construction of biological networks and discuss the current methods available.

1.1. Experimental methods for determining network structure

The use of the word 'network' has only become common in molecular biology in recent years and reflects a change in the scale of experimental data available rather than the birth of an entirely novel concept. The classical framework for organising and presenting biological models has been the pathway. The traditional approach to building such a pathway would start with a particular phenotype, for example the ability to utilise a particular metabolite, for which a genetic screen could be designed to identify the genes involved. Having identified the components of the system a large variety of techniques can be employed on a gene-by-gene basis, usually in a hypothesis-driven manner, to determine which interact, how and, where possible, in what order. One of the earliest successes of this approach was the discovery of the lac operon and the means by which it is regulated [1] and has since been the dominant paradigm in molecular research [2].

Facilitated by advances in miniaturisation and robotics, some of the classical techniques used to pick apart interactions have been scaled up for application at the genome-wide level. For instance the yeast two-hybrid method used to detect physical interactions between proteins has now been applied on a genome-wide scale in a handful of organisms [3–8]. Similarly, affinity purification methods have been coupled to high-throughput mass spectrometry to uncover the composition of protein complexes in human [9], yeast [10–12] and *Escherichia coli* [13]. The construction of gene deletion libraries in yeast [14] and RNAi libraries in *Caenorhabditis elegans* [15], have allowed the construction of genetic interaction networks. In addition to increasing the throughput of existing methods many novel methods have been developed to map biological networks. One example of this is the development of microarray technology which has allowed the parallel measurement of transcript levels [16], thus allowing the construction of gene co-expression networks [17].

One of the major advantages that these high-throughput approaches offer over the classical methods is the broader scope that they bring. Rather than considering only the components identified to be responsible for a particular phenotype many more are considered, allowing the detection of cross talk between different pathways and the characterisation of multifunctional proteins [18]. However, this increased coverage comes at the expense of resolution. Until now we have used

*Corresponding author. Address: Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany.
E-mail address: bork@embl.de (P. Bork).

0014-5793/$34.00
doi:10.1016/j.febslet.2008.02.033

the word 'edge' to describe any relationship between a pair of nodes, however the exact nature of this relationship can vary widely. For example, in a genetic interaction network two genes are linked by an edge if one gene influences the phenotypic effects of the other [19]. For some pairs of genes in this network the edge might represent a physical interaction, for others it might represent the phosphorylation of one protein by the other and for other pairs it might merely mean that both genes function at opposite ends of the same pathway. The differences between these edges can be thought of as differences in resolution, at a low level of resolution we know that the genes interact and at higher levels of resolution we know how. Recently efforts have been made to formalise the description of these edges [18,19]. An ontology proposed by Lu et al. provides not only a controlled vocabulary to describe the edges, but also a hierarchical relationship between the different edge types [18]. The depth of an edge in this hierarchy can be thought of as its resolution, edges near the root are low and those at the tips are high (see Fig. 1). It is also important to be able to distinguish between functional and non-functional interactions. Non-functional interactions are those that when disrupted have no phenotypic effect and are due to either false positives in the experimental method or biological noise.

If we now consider the networks derived from high-throughput methods in this light, we can see that they all detect different edge types with different resolutions. For instance the yeast two-hybrid and affinity purification methods both detect physical interactions between proteins. The former mostly detects direct binding between proteins, although may include some indirect interactions [20]. Affinity purification methods on the other hand, detect a mixture of direct binding associations and indirect complex associations. In neither case can these methods distinguish between functional and non-functional interactions. Genetic interaction methods can determine which interactions are functional with respect to a certain phenotype, but remain poorly resolved with respect to edge type. In contrast to the networks generated by high-throughput methods, those determined by classical methods have many different types of nodes (genes, transcripts, proteins, metabolites, etc.) and a huge variety of edges (Fig. 1) and generally only include functional interactions. In addition to the number of edge types, the level of description of these edges is far higher, with information available on the direction of these edges and even aspects such as binding affinity and reaction rate. Moreover, classical pathways often contain spatial and temporal information that allows a hierarchical structure to be imposed on the network. As technology continues to improve, the cost of the trade-off between coverage and resolution will decline, allowing high-throughput methods to create richer descriptions of network structure. In the meantime however, computational methods will continue to play an important role in determining the structure of biological networks.

1.2. The role of computational methods in determining network structure

Computational biology currently contributes to the understanding of biological networks in three main areas: (i) generating predictions of interactions, (ii) determining which interactions are functionally relevant and (iii) integrating different interaction data to provide richer and higher resolution network representations.

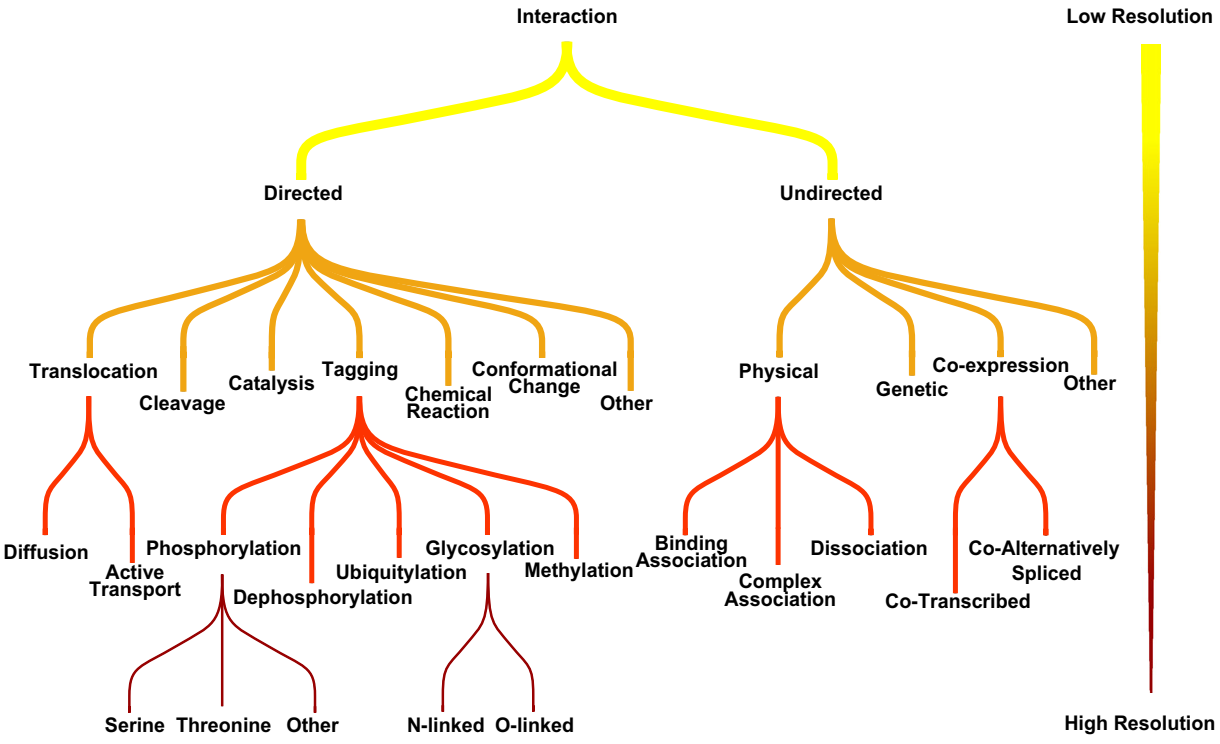

Fig. 1. An adapted version of the edge ontology proposed by Lu et al. [18]. Biological networks may be composed of many different types of interaction (edges). Different methods of detecting and predicting interactions can achieve different levels of resolution. In general, methods that predict interactions based on genomic sequence alone yield low resolution predictions, however these interactions are more likely to be functional.

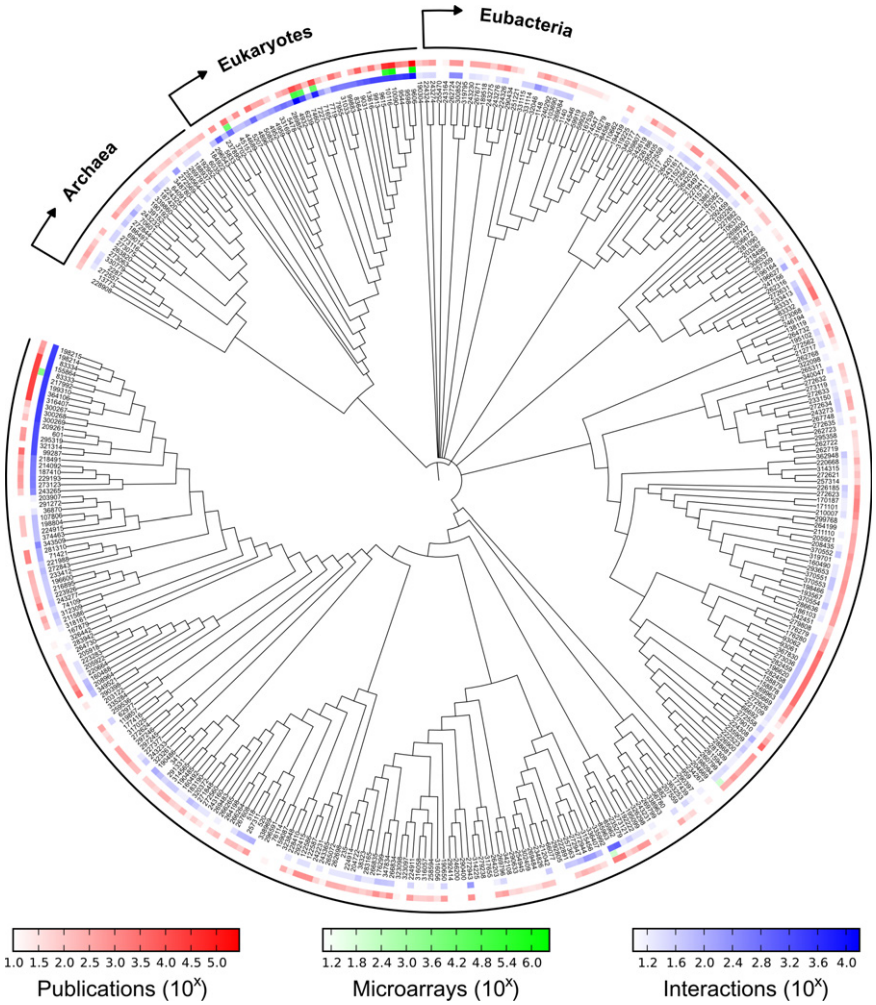

Fig. 2. Experimentally derived interactions are not equally distributed across the tree of life. The inner ring (blue) shows the number of physical protein–protein interactions for each species contained in the STRING7 database [69]. The middle ring (green) shows the number of microarray experiments, the basis for co-expression networks. The outer ring (red) shows the number of Pubmed abstracts that mention this species, providing the basis for collections of manually-curated interactions. This tree was created using iTOL [73], note that in order to display the full range of data the colour intensity is on a log-scale.

Since the publication of the first whole bacterial genome sequence [21] both the cost of sequencing and the time taken to do so have dropped by orders of magnitude, to the point where we now have the genome sequences of hundreds of organisms [22]. Over the same period, however, the phylogenetic distribution of experimentally confirmed interactions has failed to spread much beyond a handful of model organisms (Fig. 2). This is partly due to the fact that a whole genome sequence is often a prerequisite for the high-throughput methods of interaction detection. However, in addition to this natural lag, there are many more technical barriers to the application of high-throughput methods of interaction detection. Currently the only requirement for an organism to be sequenced is that enough source material is available for sequencing, meaning that only unculturable prokaryotes are unamenable to whole genome sequencing [23]. In contrast, most of the methods of detecting interactions require significant investments of time and resources before they can be transferred to another organism. Many of the high-throughput methods mentioned above require the construction of large libraries of gene deletion mutants or fusion proteins, which prevents their widespread application. Similarly the investment required to develop a microarray platform has meant that such experiments have only been carried out in a handful of organisms. In other cases species-specific biology prevents the general application of a method. For instance the methods to detect synthetic lethal genetic interactions in yeast rely on libraries of gene deletion mutants, while in eukaryotes RNAi is used [24]. All of these factors have lead to a very peaked distribution of experimentally determined interactions (Fig. 2).

Over the next few years it is likely that the need to culture organisms for whole genome sequencing will be eliminated [25,26], meaning that there will be many more organisms for which complete genome sequences but no experimental interaction data are available. The computational methods described below are based on evolutionary principles and therefore do not require experimentally determined interactions for each organism in order to make predictions. Therefore, these methods can be used to reconstruct some of the networks in these organisms [27,28] and indeed will become more powerful as additional genomes are sequenced.

In addition to their limited phylogenetic coverage, experimental methods are also limited with respect to the proportion of a given species network that they cover. This is partly due to the dynamic nature of the networks being examined, with many interactions present only in certain cellular, developmental or environmental contexts [29]. Therefore to achieve full coverage of the network, one would face the daunting task of have to experimentally sample the network over all of these contexts. On the other hand, by detecting relationships between genes that are evolutionarily conserved, computational methods implicitly include all contexts in which this interaction functions.

The coverage of these networks is also limited by the technical shortcomings of the methods used [30]. For example, it is estimated that only 50% of the yeast and 10% of the human physical protein–protein interactions have been mapped [31]. This is partly due to the undersampling described in the previous paragraph, however is also due to technical aspects of the methods used. For example the false negative rate (the proportion of true interactions missed by the method) in yeast two-hybrid datasets ranges from 75% in *C. elegans* to 90% in *Drosophila melanogaster*, up to 85% of which are missed due to systematic errors in the technique [32]. On top of this, many of the interactions determined by these methods, although real, have little or no functional impact on the organism. The interactions determined by computational methods are conserved and therefore should only contain functional interactions.

2. Computational methods for predicting interactions

Most computational methods for predicting interactions from genomic sequence are based on the rationale that if genes functionally interact, then they are likely to have a shared evolutionary history. This can be detected as correlations between different aspects of their evolution in multiple lineages.

2.1. Phylogenetic profiles

Perhaps the simplest correlation that can be used is the correlation between the phylogenetic distributions of two genes. The justification for this method is that if two genes are functionally related, they should tend to be co-inherited, as the loss of either one of these genes would be detrimental to that particular function [33,34]. This pattern of inheritance can be detected by creating a vector of the presence or absence of a particular orthologous group across a set of species (Fig. 3a). These vectors can then be clustered together to identify genes that have similar inheritance profiles and are therefore likely to be functionally related. In addition to the direct detection of functional interactions between genes, phylogenetic profiles can also be used to indirectly infer interactions between genes. By looking for anti-correlated phylogenetic profiles, Morett et al. could detect analogous proteins, functionally equivalent but non-homologous, in different species [35]. As well as detecting pairwise relationships between genes, phylogenetic profiles can be used to detect higher-order relationships. By comparing the profiles of three genes at a time, Bowers et al. were able identify logic relationships behind the presence or absence of genes across genomes [36]. For instance, the function of a certain gene might depend on the function of two other genes, in which case that gene would only be seen in genomes where both genes are present.

As the number of fully sequenced genomes increases this method will become more powerful, however methods will need to account for the phylogenetic biases in these genomes. The whole genome sequences that are present in the public dat-

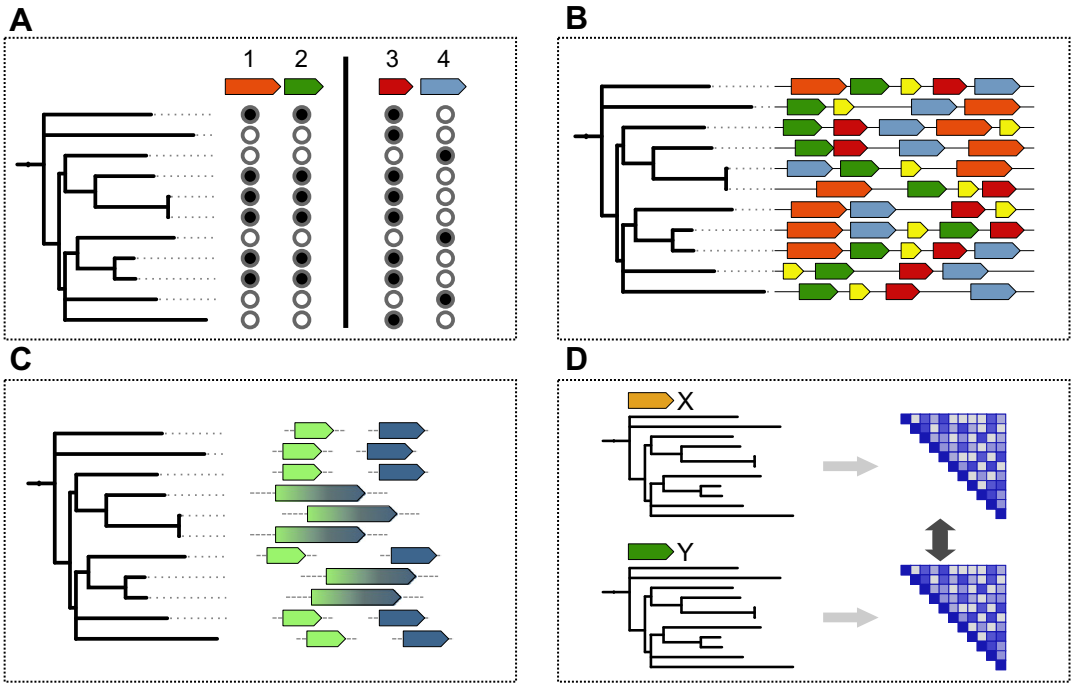

Fig. 3. Computational methods for predicting protein interactions. (a) Phylogenetic profiles; (b) genomic neighbourhood; (c) gene fusion and (d) sequence co-evolution.

abases are biased towards industrially and medically relevant organisms. As a consequence many more genomes have been sequenced from some parts of the tree of life (for example the γ-proteobacteria) than from others. This raises the problem that the correlation detected in the presence/absence profile of a pair of genes may be due to their shared ancestry rather than any selective pressure. In other words not enough time has elapsed since these species diverged for one or other or both of these genes to be lost in a lineage. Therefore correlations between the profiles in closely related sequences are less informative than those from distantly related species. Perhaps the strongest evidence that a pair of genes are functionally linked is if the same pair of genes are lost (or gained by horizontal gene transfer) together in several lineages independently [37]. By using a phylogenetic tree in combination with phylogenetic profiles, the detection of functional links can be improved both in terms of sensitivity and specificity [37]. As methods to construct phylogenetic trees for large numbers of species improve [38], the combination of phylogenetic trees and profiles are likely to play a large role in the prediction of functional interactions.

The type of interaction detected by this method are similar to those detected in genetic interaction experiments, which look at the phenotypic effects of double gene deletion mutants relative to the effects of the single gene mutants [14]. The difference between the interactions detected is that genetic interaction experiments only assay the interactions for a specific phenotype, for example growth rate, whereas those detected by the phylogenetic profile method detect those responsible for overall fitness. Phylogenetic profiles of genes can be also coupled to phenotypic information of the species in which they are found. Korbel et al. employed literature mining to connect genes to certain phenotypic properties, many of which would not have been revealed by profiling or genomic neighbourhood methods [39].

2.2. Genomic neighbourhood

The genomic neighbourhood methods can be seen as an extension to the phylogenetic profiles, in addition to looking for a tendency for genes to be co-inherited they also look for a tendency to cluster together on the genome (Fig. 3b). Similarly the reasoning behind these methods is that there is a selective pressure to keep these genes close on the chromosome, indicative of a functional relationship between the genes in a cluster. Just what this selective pressure is and what it tells us about the possible interactions will be discussed below, but first we will describe the approaches used to detect such clusters.

The goal of genomic neighbourhood methods is to detect clusters of genes that recur across multiple genomes. This is more difficult than it may seem at first glance due to the fact that the order of genes within a cluster is not necessarily important for its function and therefore there can potentially be disrupted by shuffling within the cluster [40,41]. Therefore, assuming that you have a starting set of clusters which you think might be conserved, to look for the conservation of these clusters in the genome of another organism one would have to look for each permutation of the genes within these clusters. With the number of whole genome sequences currently available, this task would be very computationally intensive. However, if we do not have a starting set of clusters the task becomes virtually impossible. Therefore, methods have been employed to reduce the computational complexity of the problem. The first methods defined gene clusters for each genome individually and then looked for pairs of orthologs that were seen in the same cluster in different genomes [42]. This method uses some of the properties of operons, the most common type of conserved gene clusters, to define the starting set of clusters. The intergenic distance between a pair of genes within an operon is on average much shorter than at the boundary of an operon [43], therefore the starting sets of genomic neighbourhoods were defined as set of genes on the same strand, uninterrupted by genes on the opposite strand, where the maximum intergenic distance between a pair of genes was 300 nucleotides [42]. Another way of avoiding the computational cost of considering all permutations is to only consider pairs of genes at a time, either direct neighbours [44,45] or allowing for a small number of intervening genes [46,47]. This has the added advantage of removing the requirement that the genes are on the same strand, allowing the detection of conserved genomic neighbourhoods other than operons. Korbel et al. found that pairs of genes on opposite strands can be conserved together over long evolutionary timescales and are likely to form regulatory circuits, where one gene is a transcription factor that auto-regulates its own synthesis and also regulates its conserved partner [46].

It has been shown that recently formed operons are much less likely to contain functionally related genes than more ancient ones [48] and therefore, similar to the methods based on phylogenetic profiles, it is important to be able to distinguish between gene clusters that recur due to shared ancestry and those that are maintained by selection [33]. One way to distinguish between these scenarios is to remove closely related species from the comparison, thus ensuring that a sufficient number of recombination events have occurred to disrupt a non-functional cluster [45,47]. For metagenomic data, where the exact taxonomic origin of the sequence is unknown, this method is impossible. Therefore a method was developed that downweights evidence from closely related clusters relative to those that occur over longer distances, thus permitting the prediction of thousands of novel interaction [44].

As an extension of the phylogenetic profiles method, genomic neighbourhood methods also predict functional relationships between proteins. The exact nature of this functional relationship can be inferred from the selective pressure acting to maintain the cluster [41,48]. Under Fisher's model of gene clustering, the increased proximity of functionally related genes reduces the chance that recombination will break apart a pair of co-adapted alleles [33,49]. In this case, the conservation of a cluster merely tells us that the genes function together, but gives no hint as to the nature of this interaction. A more specific prediction is derived from the theory that gene clusters are maintained by the need to coordinately regulate the expression of the genes within. There are several mechanisms by which this can be achieved. In prokaryotes and some eukaryotes it takes the form of operons, where a promoter drives expression of a single transcript containing several genes. In eukaryotes coordinate regulation of gene clusters may be achieved through the remodelling of local chromatin structure [50], or bi-directional promoters [51]. Co-regulation of expression allows sets of genes to be activated simultaneously, a requirement for many functional interactions. For instance, the lac operon is activated by the presence of

allolactose, activating proteins involved in the uptake and metabolism of lactose [1]. In addition to being on the same biochemical pathway, some operons, especially the more conserved ones, contain members of the same protein complex [45]. It is thought that translation from a single transcript aids complex formation by reducing stochastic differences in protein levels [52] and increasing local protein concentrations [45]. However, it should be noted that the functional interaction implied by co-expression is not always be so clear cut. For instance, some conserved operons contain a mixture of ribosomal proteins and enzymes of central metabolism, the link being that expression levels of these genes correlate with growth rate [47].

2.3. Gene fusion

An extreme case of gene clustering occurs when a pair of genes becomes fused into a single open reading frame (Fig. 3c) [53,54]. Early examples of such genes came from the comparison of bacterial and fungal genomes. For instance the alpha and beta polypeptides of tryptophan synthetase in *E. coli* are observed as a single gene in *S. cerevisiae* [55]. In addition to permitting the tight co-regulation of expression, gene fusion is thought to increase the efficiency of biochemical and signalling pathways by coupling together successive steps [54]. As a result the interactions predicted by gene fusion methods are somewhat more specific than those predicted by most methods. For instance, rather than merely predicting that two genes are likely to be on the same pathway, gene fusion methods predict that the genes are likely to carry out successive steps along a pathway [56]. Moreover, some gene fusion events have occurred independently in multiple lineages, an indicator of positive selection [56].

2.4. Sequence co-evolution

The methods described so far have examined the existence and genomic location of genes to detect the co-evolution of functionally related genes. Another collection of methods go into more detail to look at the level of sequence to detect correlations between sequence changes in different genes [57,58]. These methods are based on the observation that the phylogenetic trees of proteins that physically interact are more similar to each other than expected [59]. This pattern is thought to be caused by the process whereby mutations in a gene that are detrimental to an interaction can be compensated for by mutations in its interacting partner, implying a relatively tight functional linkage. Alternatively this pattern may be caused by a more general trend, whereby functionally related proteins evolve at more similar rates than unrelated ones, in which case the functional linkage may be more loose [60].

The methods used to detect this pattern all take the same general approach to quantifying the similarity of phylogenetic trees (Fig. 3d). Firstly a multiple sequence alignment is constructed for each gene family under consideration, from which an evolutionary distance matrix is derived. These matrices are then compared to each other, the similarity between matrices quantified by Pearson's correlation coefficient, and those with high coefficients are likely to contain interacting proteins [57,58]. As it is described this method only works reliably on single-copy orthologs, however given that the divergence of binding specificities among duplicate genes is an important source of functional complexity, it is also important to be able to deal with families that contain paralogs [61]. In order to do so, an additional step has to be carried out where the distance matrix family of one family is aligned to the other [61].

As with all the methods examined so far these methods have to account for correlations that arise due to shared ancestry rather than selection [62–64]. One approach to this has been to construct a distance matrix containing the speciation signal, for example from a tree built using 16S rRNA sequences, which can then be subtracted from each of the distance matrices, improving the quality of the predictions [63].

3. Integrating computational predictions and experimental data

While these methods are important in that they permit us to predict functional interactions using genomic sequence alone, in reality they provide a relatively low-resolution picture of biological networks. In order to increase the resolution and provide more detailed representations of biological networks they must be combined with each other and with other datasets of experimentally derived interactions [19,65–69].

The first advantage of combining datasets is that it can improve the accuracy of the network. For instance, an interaction with a low confidence score in any one dataset may be upgraded if seen in another dataset, thus reducing the false negative rate. For this reason it has been suggested that the raw data from interaction experiments are reported rather than just the ones above a confidence threshold [31]. Similarly the false positive rate of a method may be reduced by using the networks derived by other methods to filter out spurious interactions. This approach was recently used to derive a network of protein kinases and their targets [70]. Thousands of protein phosphorylation sites have been identified by high-throughput in vitro experiments, allowing the construction of consensus motifs for many of the known kinases. However, not all proteins containing such a motif are true targets of a kinase, as the contextual factors such as protein localisation, coexpression and structure can all determine whether a motif is a true target. By using the STRING network [69] to filter out the interactions that were unlikely to be functional, thus implicitly incorporating this contextual information, Linding et al. achieved a 2.5-fold increase in the accuracy of the phosphorylation network.

As well as creating more accurate representations of networks, the integration of data will allow us to create richer representation of biological networks, approaching the level of detail created by classical approaches. For instance by integrating the predictions derived from the methods described in this chapter we can infer the likely nature of a functional interaction between genes (and vice versa infer which experimentally derived interactions are indeed functional). Moreover, for some of the experimentally derived networks, such as genetic interaction, phosphorylation and transcription factor networks, the edges are directed which allow them to be ordered into pathways [19]. A complete representation of biological networks also requires that they are expanded beyond genes and their products to include small molecules [71].

As these richer network representations are integrated with temporal and spatial information we will achieve a better quantitative understanding of biological systems, approaching the level of the classically defined pathways [29,72]. However,

in order to fully exploit these data we have to develop comparative methods that go beyond the genome sequence to compare networks between species.

Acknowledgements: This work was supported through the GeneFun Specific targeted Research Project, Contract No. LSHG-CT-2004-503567, and through the BioSapiens Network of Excellence, Contract No. LSHG-CT-2003-503265, both funded by the European Community FP6 programme.

References

[1] Jacob, F. and Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3, 318–356.
[2] Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999) From molecular to modular cell biology. Nature 402, C47–C52.
[3] Giot, L. et al. (2003) A protein interaction map of *Drosophila melanogaster*. Science 302, 1727–1736.
[4] Ito, T. et al. (2000) Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc. Natl. Acad. Sci. USA 97, 1143–1147.
[5] Li, S. et al. (2004) A map of the interactome network of the metazoan *C. elegans*. Science 303, 540–543.
[6] Rual, J.F. et al. (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437, 1173–1178.
[7] Stelzl, U. et al. (2005) A human protein–protein interaction network: a resource for annotating the proteome. Cell 122, 957–968.
[8] Uetz, P. et al. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. Nature 403, 623–627.
[9] Ewing, R.M. et al. (2007) Large-scale mapping of human protein–protein interactions by mass spectrometry. Mol. Syst. Biol. 3, 89.
[10] Gavin, A.-C. et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440, 631–636.
[11] Gavin, A.-C. et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147.
[12] Krogan, N.J. et al. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 440, 637–643.
[13] Butland, G. et al. (2005) Interaction network containing conserved and essential protein complexes in *Escherichia coli*. Nature 433, 531–537.
[14] Tong, A.H.Y. et al. (2004) Global mapping of the yeast genetic interaction network. Science 303, 808–813.
[15] Lehner, B., Crombie, C., Tischler, J., Fortunato, A. and Fraser, A.G. (2006) Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. Nat. Genet. 38, 896–903.
[16] 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.
[17] 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.
[18] Lu, L.J. et al. (2007) Comparing classical pathways and modern networks: towards the development of an edge ontology. Trends Biochem. Sci. 32, 320–331.
[19] Beyer, A., Bandyopadhyay, S. and Ideker, T. (2007) Integrating physical and genetic maps: from genomes to interaction networks. Nat. Rev. Genet. 8, 699–710.
[20] Aloy, P. and Russell, R.B. (2002) The third dimension for protein interactions and complexes. Trends Biochem. Sci. 27, 633–638.
[21] Fleischmann, R.D. et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269, 496–512.
[22] Liolios, K., Tavernarakis, N., Hugenholtz, P. and Kyrpides, N.C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res. 34, D332–D334.
[23] Torsvik, V. and Øvreås, L. (2002) Microbial diversity and function in soil: from genes to ecosystems. Curr. Opin. Microbiol. 5, 240–245.
[24] Costanzo, M., Giaever, G., Nislow, C. and Andrews, B. (2006) Experimental approaches to identify genetic networks. Curr. Opin. Biotechnol. 17, 472–480.
[25] Marcy, Y. et al. (2007) Dissecting biological "dark matter with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc. Natl. Acad. Sci. USA 104, 11889–11894.
[26] Zhang, K., Martiny, A.C., Reppas, N.B., Barry, K.W., Malek, J., Chisholm, S.W. and Church, G.M. (2006) Sequencing genomes from single cells by polymerase cloning. Nat. Biotechnol. 24, 680–686.
[27] Osterman, A. and Overbeek, R. (2003) Missing genes in metabolic pathways: a comparative genomics approach. Curr. Opin. Chem. Biol. 7, 238–251.
[28] von Mering, C., Zdobnov, E.M., Tsoka, S., Ciccarelli, F.D., Pereira-Leal, J.B., Ouzounis, C.A. and Bork, P. (2003) Genome evolution reveals biochemical networks and functional modules. Proc. Natl. Acad. Sci. USA 100, 15428–15433.
[29] Bork, P. and Serrano, L. (2005) Towards cellular systems in 4D. Cell 121, 507–509.
[30] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417, 399–403.
[31] Hart, G.T., Ramani, A.K. and Marcotte, E.M. (2006) How complete are current yeast and human protein-interaction networks? Genome Biol. 7, 120.
[32] Huang, H., Jedynak, B.M. and Bader, J.S. (2007) Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. PLoS Comput. Biol. 3, e214.
[33] Huynen, M.A. and Bork, P. (1998) Measuring genome evolution. Proc. Natl. Acad. Sci. USA 95, 5849–5856.
[34] Pellegrini, M., Marcotte, E.M., 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.
[35] Morett, E. et al. (2003) Systematic discovery of analogous enzymes in thiamin biosynthesis. Nat. Biotechnol. 21, 790–795.
[36] Bowers, P.M., Cokus, S.J., Eisenberg, D. and Yeates, T.O. (2004) Use of logic relationships to decipher protein network organization. Science 306, 2246–2249.
[37] Barker, D. and Pagel, M. (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput. Biol. 1, e3.
[38] Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B. and Bork, P. (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311, 1283–1287.
[39] Korbel, J.O., Doerks, T., Jensen, L.J., Perez-Iratxeta, C., Kaczanowski, S., Hooper, S.D., Andrade, M.A. and Bork, P. (2005) Systematic association of genes to phenotypes by genome and literature mining. PLoS Biol. 3, e134.
[40] Lathe 3rd, W.C., Snel, B. and Bork, P. (2000) Gene context conservation of a higher order than operons. Trends Biochem. Sci. 25, 474–479.
[41] Rogozin, I.B., Makarova, K.S., Wolf, Y.I. and Koonin, E.V. (2004) Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes. Brief Bioinform. 5, 131–149.
[42] 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.
[43] Salgado, H., Moreno-Hagelsieb, G., Smith, T.F. and Collado-Vides, J. (2000) Operons in *Escherichia coli*: genomic analyses and predictions. Proc. Natl. Acad. Sci. USA 97, 6652–6657.
[44] Harrington, E.D., Singh, A.H., Doerks, T., Letunic, I., von Mering, C., Jensen, L.J., Raes, J. and Bork, P. (2007) Quantitative assessment of protein function prediction from metagenomics shotgun sequences. Proc. Natl. Acad. Sci. USA 104, 13913–13918.
[45] 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.

[46] Korbel, J.O., Jensen, L.J., von Mering, C. and Bork, P. (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat. Biotechnol. 22, 911–917.
[47] Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., Tatusov, R.L., Szekely, L.A. and Koonin, E.V. (2002) Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res. 30, 2212–2223.
[48] Price, M.N., Arkin, A.P. and Alm, E.J. (2006) The life-cycle of operons. PLoS Genet. 2, e96.
[49] Fisher, R.A. (1930) The Genetical Theory of Natural Selection, Oxford University Press.
[50] Batada, N.N., Urrutia, A.O. and Hurst, L.D. (2007) Chromatin remodelling is a major source of coexpression of linked genes in yeast. Trends Genet. 23, 480–484.
[51] Kruglyak, S. and Tang, H. (2000) Regulation of adjacent yeast genes. Trends Genet. 16, 109–111.
[52] Swain, P.S. (2004) Efficient attenuation of stochasticity in gene expression through post-transcriptional control. J. Mol. Biol. 344, 965–976.
[53] 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.
[54] 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.
[55] 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.
[56] Yanai, I., Wolf, Y.I. and Koonin, E.V. (2002) Evolution of gene fusions: horizontal transfer versus independent events. Genome Biol. 3, research0024.
[57] Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D. and Cohen, F.E. (2000) Co-evolution of proteins with their interaction partners. J. Mol. Biol. 299, 283–293.
[58] Pazos, F. and Valencia, A. (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng. 14, 609–614.
[59] Fryxell, K.J. (1996) The coevolution of gene family trees. Trends Genet. 12, 364–369.
[60] Chen, Y. and Dokholyan, N.V. (2006) The coordinated evolution of yeast proteins is constrained by functional modularity. Trends Genet. 22, 416–419.
[61] 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.
[62] Juan, D., Pazos, F. and Valencia, A. (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc. Natl. Acad. Sci. USA 105, 934–939.
[63] Pazos, F., Ranea, J.A.G., Juan, D. and Sternberg, M.J.E. (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J. Mol. Biol. 352, 1002–1015.
[64] Sato, T., Yamanishi, Y., Kanehisa, M. and Toh, H. (2005) The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21, 3482–3489.
[65] Bowers, P.M., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O. and Eisenberg, D. (2004) Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol. 5, R35.
[66] Hu, Z. et al. (2007) VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. Nucleic Acids Res. 35, W625–W632.
[67] 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.
[68] Srinivasan, B.S., Shah, N.H., Flannick, J.A., Abeliuk, E., Novak, A.F. and Batzoglou, S. (2007) Current progress in network research: toward reference networks for key model organisms. Brief Bioinform. 8, 318–332.
[69] von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Krüger, B., Snel, B. and Bork, P. (2007) STRING 7 – recent developments in the integration and prediction of protein interactions. Nucleic Acids Res. 35, D358–D362.
[70] Linding, R. et al. (2007) Systematic discovery of in vivo phosphorylation networks. Cell 129, 1415–1426.
[71] Kuhn, M., von Mering, C., Campillos, M., Jensen, L.J. and Bork, P. (2008) STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res. 36, D684–D688.
[72] de Lichtenberg, U., Jensen, L.J., Brunak, S. and Bork, P. (2005) Dynamic complex formation during the yeast cell cycle. Science 307, 724–727.
[73] Letunic, I. and Bork, P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23, 127–128.