REVIEW

Modeling cellular machinery through biological network comparison

Roded Sharan[1] & Trey Ideker[2]

Molecular networks represent the backbone of molecular activity within the cell. Recent studies have taken a comparative approach toward interpreting these networks, contrasting networks of different species and molecular types, and under varying conditions. In this review, we survey the field of comparative biological network analysis and describe its applications to elucidate cellular machinery and to predict protein function and interaction. We highlight the open problems in the field as well as propose some initial mathematical formulations for addressing them. Many of the methodological and conceptual advances that were important for sequence comparison will likely also be important at the network level, including improved search algorithms, techniques for multiple alignment, evolutionary models for similarity scoring and better integration with public databases.

Data on molecular interactions are increasing exponentially. Just five years ago, no more than several hundred molecular interactions had been measured for any organism. Nowadays, spurred on by advances in technologies such as mass spectrometry[1,2], genome-wide chromatin immunoprecipation[3,4], yeast two-hybrid assays[5–8], combinatorial reverse genetic screens[9] and rapid literature mining techniques[10,11], data on thousands of interactions in humans and most model species have become available. This flood of information parallels that seen for genome sequencing efforts in the recent past, and presents exciting new opportunities for understanding cellular biology and disease in the future.

Given this landscape, the challenge is to develop new strategies and theoretical frameworks to filter, interpret and organize interaction data into models of cellular function. As with biological sequence analysis, a comparative or evolutionary view provides a powerful base from which to address this challenge. However, and although sequence comparison has long been a staple of biological research, the development of a similar toolbox for comparing biological networks is still in its infancy. Nonetheless, a number of recent advances have made it possible to begin to define this field in terms of the computational methodology it requires and the biological questions it may be able to answer.

Conceptually, network comparison is the process of contrasting two or more interaction networks, representing different species, conditions, interaction types or time points. This process aims to answer a number of fundamental biological questions: which proteins, protein interactions and groups of interactions are likely to have equivalent functions across species? Based on these similarities, can we predict new functional information about proteins and interactions that are poorly characterized? What do these relationships tell us about the evolution of proteins, networks and whole species?

[1]School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel. [2]Department of Bioengineering, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. Correspondence should be addressed to R.S. (roded@post.tau.ac.il).

Published online 6 April 2006; doi:10.1038/nbt1196

A final question relates to noise. Given that systematic screens for protein interactions may report large numbers of false-positive measurements[12], which interactions represent true binding events? On the one hand, confidence measures on interactions can and should be taken into account before network comparison[13–17]. On the other hand, because a false-positive interaction is unlikely to be reproduced across the interaction maps of multiple species, network comparison itself increases confidence in the set of molecular interactions found to be conserved.

Such questions have motivated three types, or modes, of comparative methods (**Table 1**). Network alignment is the process of globally comparing two networks, identifying regions of similarity and dissimilarity. Network alignment is commonly applied to detect subnetworks that are conserved across species and, hence, likely to represent true functional modules[18]. Network integration is the process of combining several networks, encompassing interactions of different types over the same set of elements, to study their interrelations. Network integration can assist in predicting protein interactions[19] and uncovering protein modules that are supported by interactions of different types[20,21]. The main conceptual difference from network alignment is that the integrated networks are defined on the same set of elements. The final mode of comparison is network querying, in which a given network is searched for subnetworks that are similar to a subnetwork query of interest[18]. This basic database search operation is aimed at transferring biological knowledge within and across species.

In this review, we survey the key analytical techniques that have served to define each mode of analysis along with the open problems they present. We then describe one possible road ahead, inspired by analogous developments in the history of sequence comparison.

Pairwise network alignment

In basic pairwise network alignment, homologous pairs of interactions, one from each of two molecular interaction networks, are identified. Studies by Matthews *et al.*[22] and Yu *et al.*[23] compared protein-protein interaction networks and regulatory networks across species, identifying pairs of interactions, called interologs and regulogs respectively,

Table 1 Modes of network comparison

Mode	Common application	Main goals	Some current limitations
Alignment	At least two networks of the same type across species	Identification of functional (conserved) protein modules; study of network evolution; interaction prediction	Limited to few (five or fewer) species; nonevolution-based scores
Integration	At least two networks of different types for the same species	Identification of modules (supported by several networks); study of interrelations between data types; interaction prediction	No agreed-upon way to combine scores over different networks
Querying	Subnetwork module versus a network	Identification of duplicated/conserved instances of the module; knowledge transfer	Query is limited to a tree topology; nonevolution-based scores

involving either two genes or two proteins in one species and their best sequence matches in another species. Beyond alignment of single interactions, it is possible to envisage a whole array of network structures that might be conserved between two protein networks. For instance, conserved linear paths may correspond to signaling pathways, and conserved clusters of interactions may be indicative of protein complexes. In certain cases, for example, when the two networks being compared represent linear chains of interactions[24], the network alignment problem admits efficient algorithmic solutions. In general, the problem is computationally hard (generalizing subgraph isomorphism under certain formulations), but heuristic approaches have been devised for it (e.g., Berg & Lassig[25]).

One heuristic approach creates a merged representation of the two networks being compared, called a network alignment graph, and then applies a greedy algorithm for identifying the conserved subnetworks embedded in the merged representation. In a network alignment graph, the nodes represent sets of molecules, one from each network, and the links represent conserved molecular interactions across the different networks (**Fig. 1**). The alignment is particularly simple when there exists a one-to-one correspondence between molecules across the two networks, but in general there may be a complex many-to-many correspondence.

A network alignment graph facilitates the search for conserved network regions, as these will appear as subnetworks with specific structure. For instance, conserved protein complexes might appear as clusters of densely interacting nodes. This technique was first used by Ogata *et al.*[26], who searched for correspondences between the reactions of specific metabolic pathways and the genomic locations of the genes encoding the enzymes catalyzing those reactions. Their network alignment graph combined the genome ordering information, represented as a network of genes arranged in a linear (or circular) path, with a network of successive enzymes in metabolic pathways. Single-linkage clustering was applied to this graph to identify pathways for which the enzymes clustered along the genome (**Fig. 2a**).

Kelley *et al.*[18] applied the concept of network alignment to the study of protein interaction networks. They translated the problem of finding conserved pathways to that of finding high-scoring paths in the alignment graph. Their algorithm, PathBLAST, identified five regions that were conserved across the protein networks of *Saccharomyces cerevisiae* and *Helicobacter pylori*. This comparison was later extended to detect conserved protein clusters rather than paths[27], employing a likelihood-based scoring scheme that weighs the denseness of a given subnetwork versus the chance of observing such topology at random (**Box 1**). The latter approach was recently used by Suthram *et al.*[28] to show that the protein-protein interaction network of *Plasmodium falciparum* differs substantially from those of other eukaryotes. Finally, Koyuturk *et al.*[29] developed an evolution-based scoring scheme to detect conserved protein clusters, which takes into account interaction insertion/deletion and protein duplication events (**Box 1**). Their MaWish algorithm was applied to detect human-mouse conserved subnetworks.

The methodology of network alignment can also be applied to predict various properties of genes and proteins on a global scale. First and foremost, a conserved subnetwork that contains many proteins of the same known function suggests that the remaining proteins also have that function. We have recently used this concept to predict thousands of new protein functions for yeast, worm (*Caenorhabditis elegans*) and fly (*Drosophila melanogaster*), with an estimated success rate of 58–63% (ref. 13). More complex relationships, such as protein interactions, functional orthology and links between cellular processes, can also be inferred from the network alignment[13,30,31].

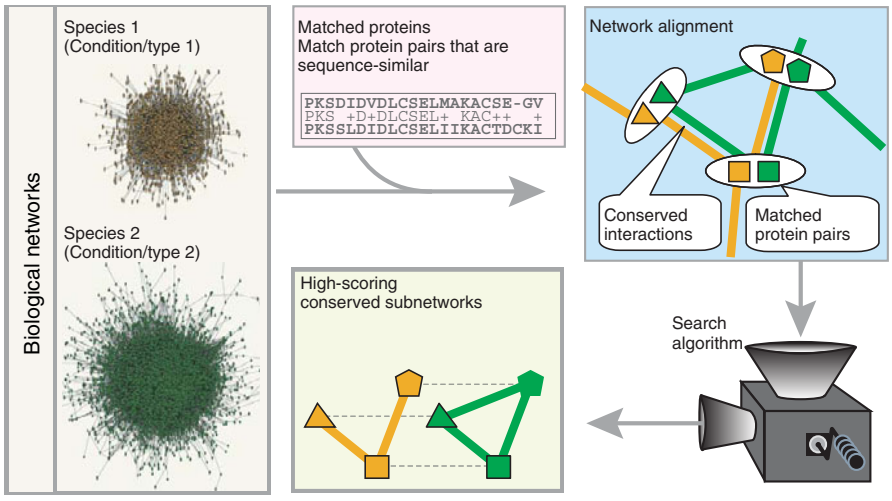

Figure 1 Network alignment. Network alignment combines protein interaction data that are available for each of at least two species with orthology information based on the corresponding protein sequences. A detailed probabilistic model is used to identify protein subnetworks within the aligned network that are conserved across the species. Each node in this aligned network represents a set of sequence-similar proteins (one from each species) and each link represents a conserved interaction. Other than species, the networks being compared can also be sampled across different biological conditions or interaction types.

Multiple network alignment

The generalization of the network alignment process to more than two networks entails devising an appropriate scoring scheme and

extending the notion of a network alignment graph. Stuart *et al.*[32] tackled the latter problem in the context of cross-species coexpression networks by forcing a consistent one-to-one mapping across all the networks, obtaining an alignment graph in which each gene is a member of at most one node. Another relatively simple scenario occurs when the compared networks are linear paths. The network alignment problem then becomes completely analogous to the sequence case, and one could adapt multiple sequence alignment techniques, such as progressive alignment, for its solution[24].

Recently, we have described a framework for multiple network alignment, which handles general correspondence relationships across networks[13]. The scoring scheme extends the likelihood approach described in **Box 1**. The search problem is handled by extending the notion of a network alignment graph to multiple networks, albeit with an increased computational complexity, which scales as nh^{k-1} for k networks of size n with an average number of h possible orthologs per protein per species. This method was applied to systematically identify conserved protein subnetworks across yeast, worm and fly, uncovering 71 conserved network regions that fell into well-defined functional categories. Two representative alignments are shown in **Figure 2b**. Such graphical layouts can be automatically generated using a variant of the spring-embedder algorithm for graph drawing[13].

Network integration

Far from being homogenous, molecular interactions come in an assortment of different types, including protein-protein, transcriptional, coexpression and genetic interactions. Together, they have populated molecular interaction databases for a large number of species (e.g., Biomolecular Interaction Network Database (BIND)[33], Database of Interacting Proteins (DIP)[34], Molecular Interactions Database (MINT)[35] and General Repository for Interaction Datasets (GRID)[36]).

Because each type of network lends insight into a different slice of biological information, integrating different network types may paint a more comprehensive picture of the overall biological system under study. Commonly, networks to be integrated are defined over the same set of elements (e.g., the set of proteins of a certain species), and the integration is achieved by merging them into a single network with multiple types of interactions, each drawn from one of the original networks. A fundamental problem is to identify in the merged network functional modules that are supported by interactions of multiple types.

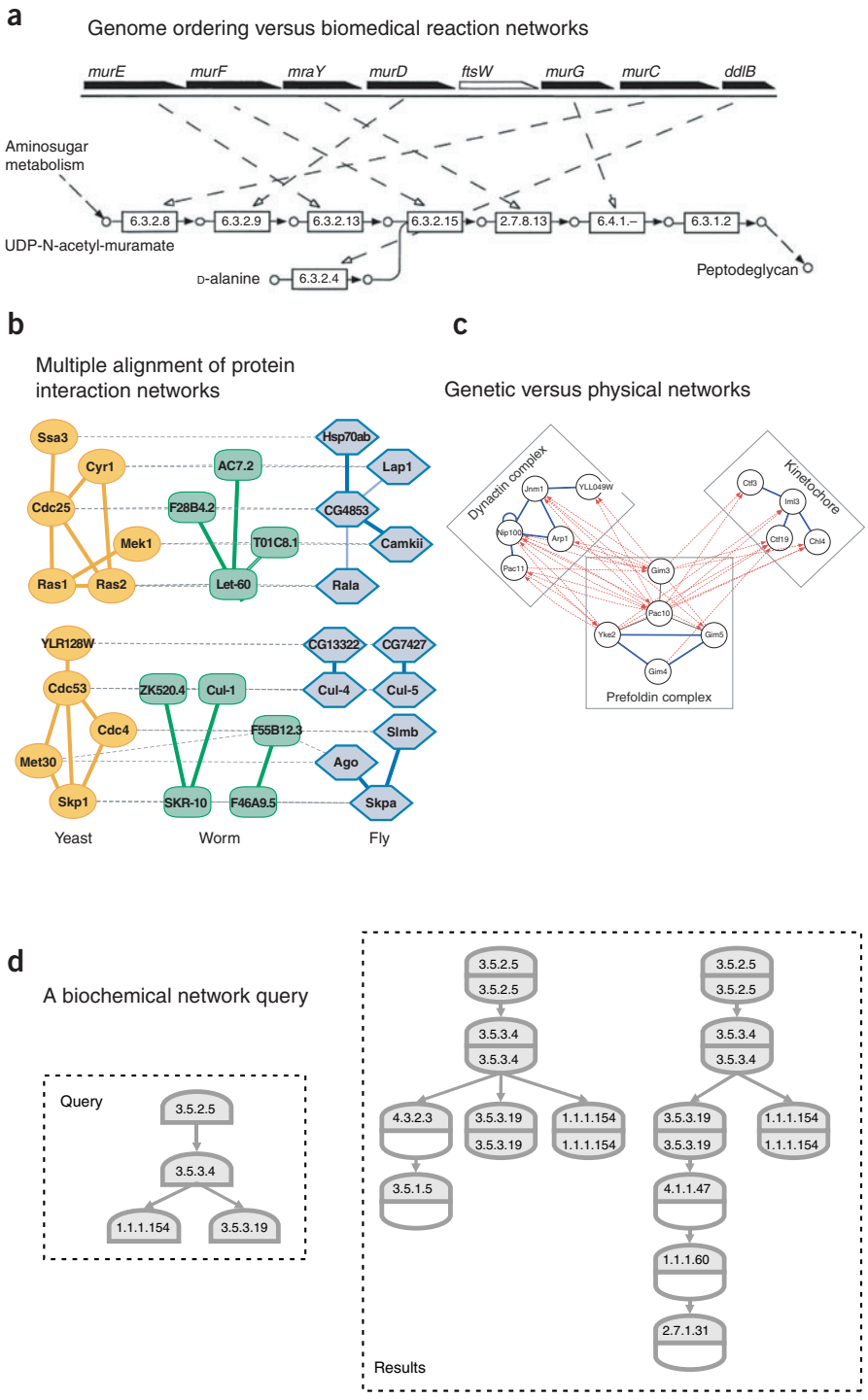

Figure 2 A gamut of network comparative approaches. Four recent examples of biological network comparisons are shown. (**a**) Ogata *et al.*[26], (**b**) Sharan *et al.*[13], (**c**) Kelley *et al.*[20] and (**d**) Pinter *et al.*[44] (figures reproduced with permission). These examples span the three modes of comparison described in the text: alignment (**a**+**b**), integration (**c**) and querying (**d**). Comparisons have involved networks of metabolic reactions (**a**+**d**), protein interactions (**b**), genetic interactions (that is, synthetic lethals) (**c**) and gene linkages on the chromosome (**a**). The networks being compared can be of the same type (**b**+**d**) or of different types (**a**+**c**). In **a**, a contiguous region of the *E. coli* genome is aligned against a contiguous pathway of peptidoglycan biosynthesis in its metabolic network. In **b**, matching proteins are linked by dotted lines, and yellow, green or blue links represent measured protein-protein interactions between yeast, worm or fly proteins, respectively. In **c**, blue versus dotted-red links correspond to protein-protein versus genetic interactions. In **d**, each alignment position groups matching enzyme commission numbers vertically.

Box 1 Algorithmics of network alignment

Identifying conserved subnetworks within a given pair of networks relies on two important algorithmic components: a scoring function and a search procedure. The scoring function measures the similarity of each subnetwork to a predefined structure of interest and the level of conservation of this structure across the subnetworks being scored. Koyuturk *et al.*[29] suggested an evolution-based scoring scheme for the alignment of protein interaction networks of two species. Define M to be the set of interologs (matches) among the two subnetworks being compared (that is, two pairs of interacting proteins, one in each subnetwork, with orthology relations between them). Define N to be the set of mismatched interactions (that is, two pairs of proteins with orthology relations between them, such that only one pair interacts). Define D to be the union of the sets of duplicated protein pairs within each subnetwork. Given scoring functions for a match (m), mismatch (n) and duplication (d), Koyuturk *et al.* define the overall score as:

$$\sum_{\alpha \in M} m(\alpha) - \sum_{\beta \in N} n(\beta) - \sum_{\chi \in D} d(\chi)$$

Thus, conservation is awarded on matches and penalized on mismatches and duplications. Statistical significance is estimated by assuming that the score is approximately normally distributed.

Another proposed solution to the scoring problem is based on maximum likelihood[27]. For each of the two aligned subnetworks, one computes a log likelihood ratio that measures the fit of the subnetwork to the desired structure (subnetwork model) versus the chance that the subnetwork is observed at random (null model). One can assume, for instance, that as a simple model of a protein complex, each protein pair within a complex interacts with high probability β, independently of other protein pairs. The null model assumes that every two proteins u,v interact with probability $p(u,v)$ that depends on their node degrees (numbers of network connections). (More precisely, $p(u,v)$ is the fraction of networks with the same node degrees that link u and v.) The likelihood that a set of proteins C with a set of interactions $E(C)$ forms a complex is thus:

$$L(C) = \sum_{(u,v) \in E(C)} \log \frac{\beta}{p(u,v)} + \sum_{(u,v) \notin E(C)} \log \frac{1-\beta}{1-p(u,v)}$$

These log likelihood ratios are summed over the aligned subnetworks to yield the overall score, which can be further extended to include information on the reliabilities of the reported interactions. To assess the significance level of the score, it is compared to those obtained under randomized versions of the input networks (that is, shuffling their edges while maintaining node degrees).

Once the scoring function is set, one can use an array of methods for searching the network alignment graph for conserved subnetworks of interest. The most commonly used method is a greedy search[13,31], which starts from promising seeds and refines them using a local search. The local search iteratively performs a modification (addition or deletion of a protein) that contributes most to the score, until no such modification is possible. There are also efficient detection methods for certain graph classes, such as paths or trees[70], which rely on color coding[71] or similar techniques.

As one example of network integration, Kelley *et al.*[20] studied the interrelations between protein-protein and genetic (synthetic lethal) interactions in yeast. They searched for two structures in the integrated network: pairs of subnetworks of protein-protein interactions interconnected to each other by a dense pattern of genetic interactions (**Fig. 2c**); and clusters enriched for both physical and genetic interactions. The first structure was found to be more prevalent, suggesting that genetic interactions tend to bridge genes operating in two pathways with redundant or complementary functions, rather than occurring between protein subunits within a single pathway. Gunsalus *et al.*[37] combined protein interaction, coexpression and phenotypic similarity networks to predict protein complexes in worm, modeled as dense clusters of interactions. This work was based on a 'reinforcement' principle: a cluster of interactions supported by several networks is more likely to represent a true protein complex than a cluster arising in one network only.

Several groups have integrated multiple networks for the purpose of predicting protein function[38] and interaction[19,39–42]. In all cases, supporting data from multiple sources, ranging from coexpression relationships to similarity of phylogenetic profiles, were used to corroborate each of the predictions. The occurrence of composite network motifs, built of several interaction types, has also been studied. Yeger-Lotem *et al.*[43] investigated the overrepresentation of network motifs in a combined network of protein-protein and transcriptional interactions. Most of the identified motifs combined both types of interactions, exhibiting a tendency toward coregulation and complex formation. Zhang *et al.*[21] integrated coexpression, transcriptional, protein-protein and genetic interactions in yeast and studied motifs in the combined network. Their findings support the 'redundant pathway' interpretation of genetic interactions (see above) and highlight a tendency of protein complexes to exhibit coregulation of their members.

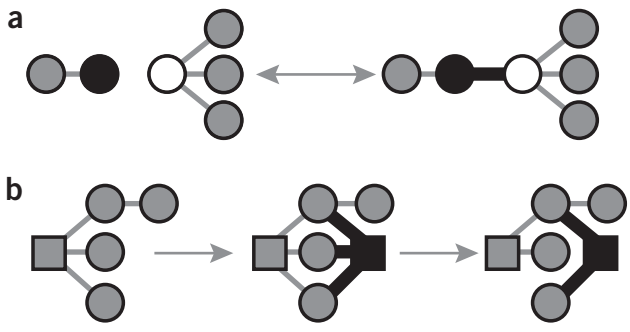

Figure 3 Evolutionary processes shaping protein interaction networks. The progression of time is symbolized by arrows. (**a**) Link attachment and detachment occur through mutations in a gene encoding an existing protein. These processes affect the connectivity of the protein whose coding sequence undergoes mutation (shown in black) and of one of its binding partners (shown in white). Empirical data shows that attachment occurs preferentially towards partners of high connectivity. (**b**) Gene duplication produces a new protein (black square) with initially identical binding partners (gray square). Empirical data suggest that duplications occur at a much lower rate than link attachment/detachment and that redundant links are lost subsequently (often in an asymmetric fashion), which affects the connectivities of the duplicate pair and of all its binding partners. Modified with permission from Berg *et al.*[47].

Network querying

Network alignment and integration are focused on *de novo* discovery of biologically significant regions embedded in a network, based on the assumption that regions supported by multiple networks are functional. In contrast, a supervised approach to the module detection problem relies on a query subnetwork that is previously known to be functional. The goal is to identify subnetworks in a given network that are similar to the query. Kelley *et al.*[18] approached the query problem in the context of the PathBLAST network alignment algorithm, by designating one of the networks as the query. When PathBLAST is applied in this setting, it identifies all matches to the query in the network under study. As in the comparison case, the treatment here is only in queries that take the form of a linear path of interacting proteins.

Recently, Pinter *et al.*[44] devised an algorithm for querying metabolic networks. Their algorithm allows querying metabolic pathways that take the form of a tree within a collection of such pathways. **Figure 2d** shows an example of their approach: here, a query of a core pathway revealed an allantoin degradation pathway in *E. coli* and a ureide degradation pathway in yeast.

Network querying tools are still at an early stage and are currently limited to sparse topologies, such as paths and trees. Approaches to handle more general queries could benefit from the rich literature on graph mining techniques in the data mining community[31,45].

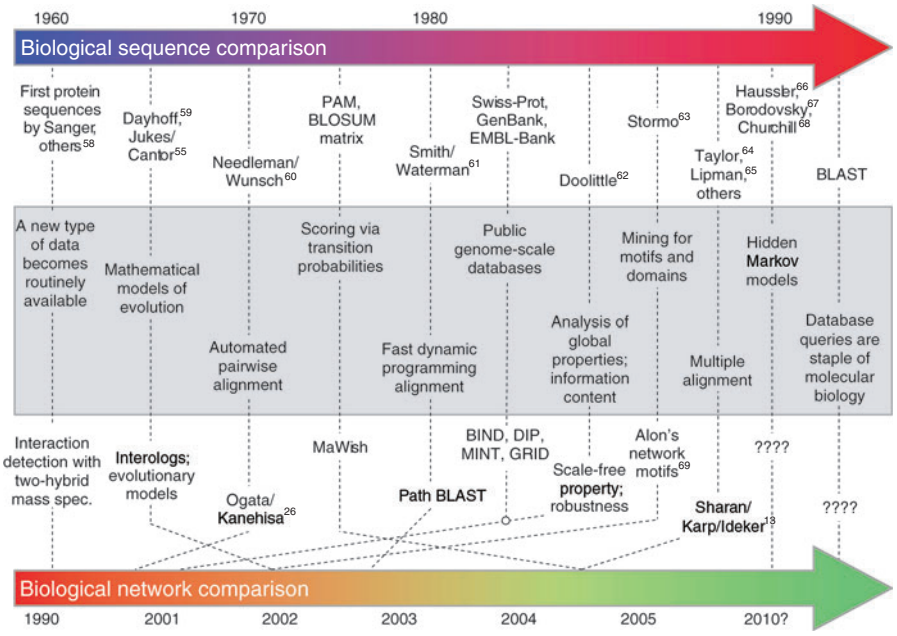

Figure 4 Parallels between sequence and network comparison on a timeline. The recent and possibly future developments in methods for network comparison are shown in the context of the analogous developments as they occurred in the field of sequence comparison. General milestones for both fields are shown in the middle (gray box), with the specific instances for sequence versus network comparison appearing directly above or below, respectively.

Network evolution

Understanding how networks evolve is a fundamental issue, which affects each of the above analysis modes as well as the study of networks in general. Two kinds of processes have been invoked to explain network evolution. The first consists of sequence mutations in a gene, which result in modifications of the interface between interacting proteins[46] (**Fig. 3a**). Consequently, the corresponding protein may gain new connections (attachment) or lose (detachment) some of the existing connections to other proteins. The second type of evolutionary process consists of gene duplication, followed by either silencing of one of the duplicated genes or by functional divergence of the duplicates (**Fig. 3b**). In terms of the network, a gene duplication corresponds to the addition of a node with links identical to the original node, followed by the divergence of some of the redundant links between the two duplicate nodes.

Berg *et al.*[47] referred to link attachment and detachment processes collectively as link dynamics. They estimated the empirical rates of link dynamics and gene duplication in the yeast protein network, finding the former to be at least one order of magnitude higher than the latter. Based on this observation, they proposed a model for the evolution of protein networks in which link dynamics are the major evolutionary forces shaping the topology of the network, whereas slower gene duplication processes mainly affect its size. Rzhetsky & Gomez[48] formulated a model that uses these two evolutionary processes, but whose underlying basic elements are domains rather than whole proteins. Barabasi & Albert[49] suggested gene duplication as the major mechanism for generating the scale-free topology of protein interaction networks. Their network growth model predicts that molecules that appeared early in the network are the most connected ones. Several lines of empirical evidence support this hypothesis: metabolites of some of the most ancient pathways, such as glycolysis and the tricarboxylic acid cycle, are among the most connected substrates in metabolic networks[50]; for protein interaction networks, one observes a positive correlation between the evolutionary age of a protein and its degree of connectivity[51].

Network comparison: the next ten years?

Notwithstanding the recent advances, the field of network comparison is still very young. However, by exploiting the close analogy to sequence comparison, one can envision some of the key milestones on the road ahead (**Fig. 4**). Methods for sequence comparison have been the main focus of bioinformatics for most of its history, starting in 1970 with the publication of the first comparison algorithm by Needleman & Wunsch[52]. Since that initial work, major advances have included better alignment score functions to more accurately reflect evolutionary distance, methods for multiple sequence alignment and numerous optimizations to the search algorithm (**Fig. 4**). In recent years, the development of sequence analysis tools has been largely driven by the immense amounts of data emerging from the human genome[53,54] and other sequencing projects.

Unlike the more mature field of sequence alignment, network alignment has a conceptual framework and several proof-of-principle studies, but relatively little in terms of advanced computational methodology. Nevertheless, it is exciting that virtually all of the major advances that occurred for sequence alignment can be envisioned for network alignment. For instance, a clear parallel goal is to progress from pairwise to multiple alignment of networks. At present, a method for three-species network alignment has been described (see above discussion of Sharan *et al.*[13]), but this algorithm scales poorly with the number of networks/species and may reach a practical limit at four or five. As yet another example, save perhaps a single study[29], the score functions for assessing network similarity are not

yet strongly rooted in an evolutionary model of how networks evolve. Innovations similar to the Jukes/Cantor model of nucleotide substitution[55] may therefore also prove fruitful for the study of networks. Just as the genome projects spurred on bioinformatics over the past decade, it is clear that high-quality interactome mapping projects will be essential to the development of these new comparison techniques and to opening new research frontiers, such as the association of network features with disease[56,57].

The analogy between sequence and network comparison is not perfect, and in fact reveals some interesting differences. For sequences, alignment methods were proposed long before large sequence databases were widely available. In contrast, large network and interaction databases have been available from the late 1990s onwards, three to four years before the first network comparisons were performed. On the other hand, computational searches for motifs and systematic characterization of global properties (e.g., amino acid content for sequences, degree distribution for networks) arose relatively late in the history of sequence analysis, but occurred early in the field of network comparison. The two problems also have crucial computational differences that stem from the linearity of sequences as opposed to the nonlinearity of networks. In particular, the problem of local sequence alignment can be solved efficiently, whereas the analogous problem of identifying conserved protein modules is computationally hard. Moreover, certain data types, such as metabolic reactions and protein complexes, are best modeled as hypergraphs, in which hyperedges link multiple (more than two) molecules together, presenting even harder computational challenges.

Finally, the possibility to integrate networks across a wide variety of biological origins appears distinct to the field of network comparison. Biological sequences are based on either nucleotides or amino acids; owing to the well-understood interdependency between these data types, integrating them has not posed a major problem. In contrast, in network space we are just beginning to understand how the different data types interrelate. For instance, it is well known that protein-protein interactions can transmit signaling events, genetic interactions can connect parallel pathways and protein-DNA interactions are the scaffold of gene regulatory control; however, the rules and configurations by which the cell integrates all of these data types to form a coordinated response are almost completely unexplored. In this regard, the growing abundance of interaction data will enable models combining different types to be systematically formulated and tested.

In summary, network comparison techniques promise to take a leading role in bioinformatics research by providing the means to contrast and query complex biological systems. Recent advances in the field, inspired by developments in sequence comparison, demonstrate the power of network comparison in elucidating network organization, function and evolution.

ACKNOWLEDGMENTS

R.S. is supported by an Alon Fellowship; T.I., by the David and Lucille Packard Foundation. This work was also supported by the National Center for Research Resources (RR018627) and the National Science Foundation (NSF 0425926).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
2. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
3. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
4. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
5. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
6. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246 (1989).
7. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
8. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
9. Tong, A.H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368 (2001).
10. Peri, S. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13**, 2363–2371 (2003).
11. Nikitin, A., Egorov, S., Daraselia, N. & Mazo, I. Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics* **19**, 2155–2157 (2003).
12. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
13. Sharan, R. *et al.* Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA* **102**, 1974–1979 (2005).
14. Bader, J.S., Chaudhuri, A., Rothberg, J.M. & Chant, J. Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* **22**, 78–85 (2004).
15. Qi, Y., Klein-Seetharaman, J. & Bar-Joseph, Z. Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac. Symp. Biocomput.* **10**, 531–542 (2005).
16. Deng, M., Sun, F. & Chen, T. Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac. Symp. Biocomput.* **8**, 140–151 (2003).
17. Suthram, S., Shlomi, T., Ruppin, E., Sharan, R. & Ideker, T. in *Proceedings of the First Annual RECOMB Systems Biology Workshop*, vol. 1 (2005).
18. Kelley, B.P. *et al.* Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci. USA* **100**, 11394–11399 (2003).
19. Rhodes, D.R. *et al.* Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.* **23**, 951–959 (2005).
20. Kelley, R. & Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* **23**, 561–566 (2005).
21. Zhang, L.V. *et al.* Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network. *J. Biol.* **4**, 6 (2005).
22. Matthews, L.R. *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs." *Genome Res.* **11**, 2120–2126 (2001).
23. Yu, H. *et al.* Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.* **14**, 1107–1118 (2004).
24. Tohsato, Y., Matsuda, H. & Hashimoto, A. in Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB) 376–383 (2000).
25. Berg, J. & Lassig, M. Local graph alignment and motif search in biological networks. *Proc. Natl. Acad. Sci. USA* **101**, 14689–14694 (2004).
26. Ogata, H., Fujibuchi, W., Goto, S. & Kanehisa, M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res.* **28**, 4021–4028 (2000).
27. Sharan, R., Ideker, T., Kelley, B., Shamir, R. & Karp, R.M. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J. Comput. Biol.* **12**, 835–846 (2005).
28. Suthram, S., Sittler, T. & Ideker, T. The Plasmodium protein network diverges from those of other eukaryotes. *Nature* **438**, 108–112 (2005).
29. Koyuturk, M., Grama, A. & Szpankowski, W. in Proceedings of the Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB) 48–65 (2005).
30. Bandyopadhyay, S., Sharan, R. & Ideker, T. Systematic identification of functional orthologs based on protein network comparison. *Genome Res.* **16**, 428–435 (2006).
31. Koyuturk, M., Grama, A. & Szpankowski, W. An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics* **20** suppl. 1, I200–I207 (2004).
32. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
33. Bader, G.D. *et al.* BIND-The biomolecular interaction network database. *Nucleic Acids Res.* **29**, 242–245 (2001).
34. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305 (2002).
35. Zanzoni, A. *et al.* MINT: a Molecular INTeraction database. *FEBS Lett.* **513**, 135–140 (2002).
36. Breitkreutz, B.J., Stark, C. & Tyers, M. The GRID: the General Repository for Interaction Datasets. *Genome Biol.* **4**, R23 (2003).
37. Gunsalus, K.C. *et al.* Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. *Nature* **436**, 861–865 (2005).
38. Kemmeren, P. *et al.* Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell* **9**, 1133–1143 (2002).
39. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453 (2003).
40. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
41. Lu, L.J., Xia, Y., Paccanaro, A., Yu, H. & Gerstein, M. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.* **15**, 945–953 (2005).
42. Wong, S.L. *et al.* Combining biological networks to predict genetic interactions. *Proc. Natl. Acad. Sci. USA* **101**, 15682–15687 (2004).

43. Yeger-Lotem, E. *et al.* Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc. Natl. Acad. Sci. USA* **101**, 5934–5939 (2004).
44. Pinter, R.Y., Rokhlenko, O., Yeger-Lotem, E. & Ziv-Ukelson, M. Alignment of metabolic pathways. *Bioinformatics* **21**, 3401–3408 (2005).
45. Giugno, R. & Shasha, D. in Proceeding of the 16th International Conference on Pattern Recognition (ICPR) 112–115 (2002).
46. Jones, S. & Thornton, J.M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **93**, 13–20 (1996).
47. Berg, J., Lassig, M. & Wagner, A. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol. Biol.* **4**, 51 (2004).
48. Rzhetsky, A. & Gomez, S.M. Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics* **17**, 988–996 (2001).
49. Barabasi, A.L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
50. Wagner, A. & Fell, D.A. The small world inside large metabolic networks. *Proc. Biol. Sci.* **268**, 1803–1810 (2001).
51. Eisenberg, E. & Levanon, E.Y. Preferential attachment in the protein network evolution. *Phys. Rev. Lett.* **91**, 138701 (2003).
52. Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
53. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
54. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
55. Jukes, T.H. & Cantor, C.R. in Mammalian Protein Metabolism (ed. Munro, H.N.) 21–123 (Academic Press, New York, 1969).
56. Goehler, H. *et al.* A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol. Cell* **15**, 853–865 (2004).
57. Calvano, S.E. *et al.* A network-based analysis of systemic inflammation in humans. *Nature* **437**, 1032–1037 (2005).
58. Sanger, F. & Tuppy, H. The amino acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem. J.* **49**, 463–481 (1951).
59. Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. A model of evolutionary change in proteins. in Atlas of Protein Sequence and Structure, vol. 5, suppl. 3, (Dayhoff, M.O., ed.) 345–352 (National Biomedical Research Foundation, Silver Spring, MD, 1978).
60. Needleman, S.B. & Wunsch, C.D. A general method applicable to the search of similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
61. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
62. Kyte, J. & Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
63. Stormo, G.D. & Hartzell, G.W. III. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA* **86**, 1183–1187 (1989).
64. Taylor, W.R. Multiple sequence alignment by a pairwise algorithm. *Comput. Appl. Biosci.* **3**, 81–87 (1987).
65. Lipman, D.J., Altschul, S.F. & Kececioglu, J.D. A tool for multiple sequence alignment. *Proc. Natl Acad. Sci. USA* **86**, 4412–4415 (1989).
66. Krogh, A., Brown, M., Mian, S., Sjolander, K. & Haussler, D. Hidden Markov models in computational biology: applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531 (1994).
67. Borodovsky, M. & McIninch, J. GENMARK: parallel gene recognition for both DNA strands. *Comput. Chem.* **17**, 123–133 (1993).
68. Churchill, G.A. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51**, 79–94 (1989).
69. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
70. Scott, J., Ideker, T., Karp, R.M. & Sharan, R. in Proceedings of the Ninth Annual International Conference on Research in Computational Molecular Biology (RECOMB) 1–13 (2005).
71. Alon, N., Yuster, R. & Zwick, U. Color-coding. *J. ACM* **42**, 844–856 (1995).