

DOI:10.1145/2160718.2160738

Examining tools that provide valuable insight about molecular components within a cell.

BY NIR ATIAS AND RODED SHARAN

Comparative Analysis of Protein Networks: Hard Problems, Practical Solutions

A HOLY GRAIL of biological research is deciphering the workings of a cell—the elementary unit of life. The main building blocks of the cell are macromolecules called proteins; they are key factors in driving cellular processes and determining the structure and function of cells. Proteins do not work in isolation but rather physically interact to form cellular machineries or transmit molecular signals. A modification of a single protein may have dramatic effects on the cell; indeed,

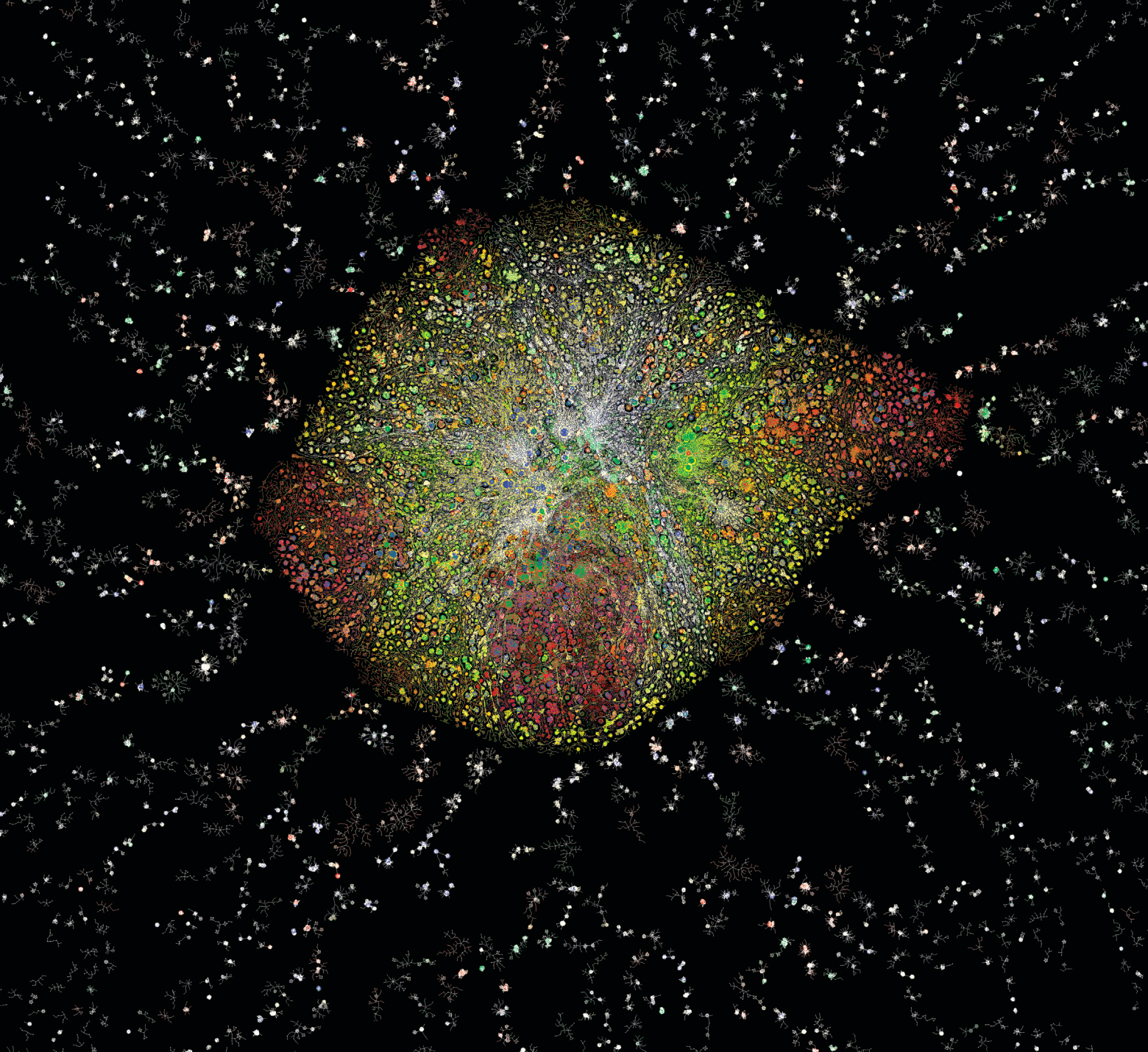
many diseases (for example, Huntington's disease²⁶) are the result of small changes to a single protein and, consequently, to its set of interacting partners and functionality. The mapping of proteins and their interactions and the interpretation of this data are thus a fundamental challenge in modern biology with important applications in disease diagnosis and therapy.¹⁵

The last two decades have witnessed a great shift in biological research. While classical research focused on a single gene or subsystem of a specific organism, the emergence of high-throughput technologies for measuring different molecular aspects of the cell has led to a different, systems-level approach. By this approach, genome-wide data is used to build computational models of certain aspects of the cell, thereby generating new biological hypotheses that can be experimentally tested and used to further improve the models in an iterative manner.

A prime example for this technological revolution is the development of techniques for measuring protein-protein interactions (PPIs). Historically, such interactions were measured at small scale—one or few interactions at a time. The development of automated, large-scale measurement technologies such as the yeast two-hybrid system¹⁰ and the co-immunoprecipitation assay¹ has enabled the mapping of

» key insights

- The explosion of biological network data necessitates methods to filter, interpret, and organize this data into modules of cellular machinery.
- The comparative analysis of networks from multiple species has proven to be a powerful tool in detecting significant biological patterns that are conserved across species and in enabling their interpretation.
- Comparative network analysis presents hard computational challenges such as graph and subgraph isomorphism and detecting heavy subgraphs; these can be tackled to near-optimality by a combination of heuristic, parameterized, and integer programming-based approaches.



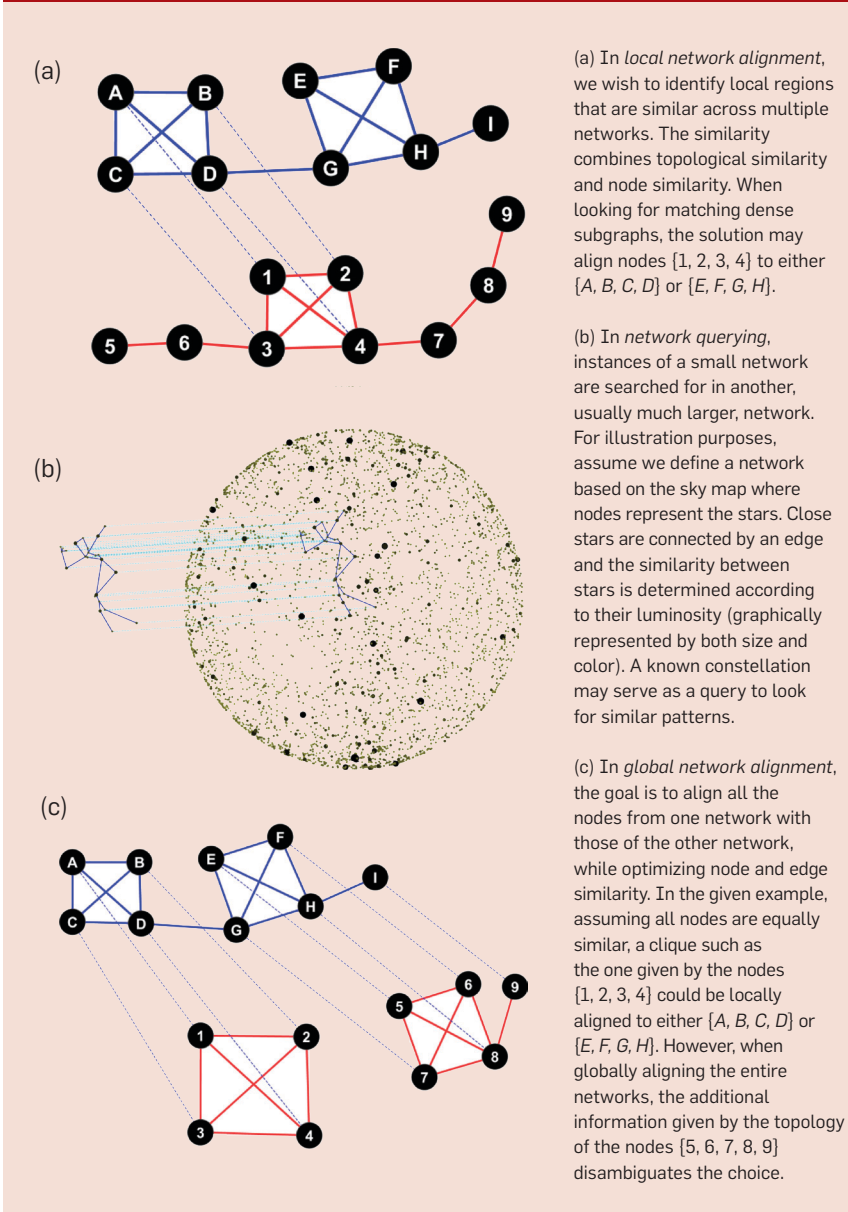
the entire interactome of a species in a single experiment.

Since the first publication of PPI data in yeast,³⁷ dozens of large-scale assays have been employed to measure PPIs in a variety of organisms including bacteria,²⁵ yeast, worm,²⁰ fly,¹² and human.^{36,27} Protein interaction data is being accumulated and assessed in numerous databases including DIP,²⁸ BioGRID,³⁵ and more. Nevertheless, PPI data remains noisy and incomplete. The reliability of different experimental sources for protein-protein interactions has been estimated to be in the range of 25%–60%.⁸ A recent experimental assessment of PPIs in

yeast³⁹ estimated that even in this well-mapped organism, the set of reproducible and highly confident interactions covers only 20% of the yeast's interaction repertoire.

The low quality of the data has driven the use of cross-species conservation criteria to focus on the more reliable parts of the network and infer likely functional components. The basic paradigm was borrowed from the genomic sequence world, where sequence conservation (across species) often implies that the conserved region is likely to retain a similar biological function.^{3,24} This evolutionary principle has motivated a series of works

that aim at comparing multiple networks to extract conserved functional components at two different levels: the protein level and the subnetwork level. On the protein level, proteins whose network context is conserved across multiple species are likely to share similar functions.³⁴ On the subnetwork level, conserved subnetworks are likely to correspond to true functional components, such as protein complexes, and to have similar function.³² In both cases, biological knowledge in any one of the species can be transferred to the others, allowing the annotation of networks in an efficient and accurate manner.³⁰

Figure 1. Computational problems in comparative network analysis.

In this review, we survey the field of comparative network analysis with an emphasis on the arising computational problems and the different methods that have been used to tackle them, starting from heuristic approaches, going through parameterized algorithms that perform well on practical instances, and ending with optimal integer linear programming (ILP)-based solutions that rely on powerful, yet available, industrial solvers. We demonstrate the applications of these methods to predict protein function and interaction, infer the organization of protein-protein interaction networks into their underlying functional modules, and link biological processes within and across species.

A Roadmap to Network Comparison Techniques

We view a PPI network of a given species as a graph $G = (V, E)$, where V is the set of proteins of the given species and E is the set of pairwise interactions among them. In a network comparison problem, one is given two or more networks along with sequence information for their member proteins. The goal is to identify similarities between the compared networks, which could be either local or global in nature (Figure 1). The underlying assumption is that the networks have evolved from a common ancestral network, and hence, evolutionarily related proteins should display similar sequence

and interaction patterns. For ease of presentation, we focus in the description below on pairwise comparisons, but the problems and their solutions generalize to multiple networks.

Most algorithms for network comparison score the similarity of two subnetworks by first computing a many-to-many mapping between their vertices (with possibly some unmatched vertices in either network) and then scoring the similarity of proteins and interactions under this mapping. Proteins are commonly compared by their associated amino-acid sequences, using a sequence comparison tool such as BLAST.³ The similarity score of any two sequences is given as a p -value, denoting the chance of observing such sequence similarity at random. Significant p -values imply closer evolutionary distance and, hence, higher chances of sharing similar functions. Interactions are compared in a variety of ways; the simplest and most common of which is to count the number of *conserved interactions*. Formally, given a mapping Φ of proteins between two networks (associating proteins of one network with sets of proteins in the other network), an interaction (u, v) in one species is said to be *conserved* in the other species if there exist $u' \in \Phi(u)$ and $v' \in \Phi(v)$ such that u' and v' interact.

Historically, the first considered problem variant was *local network alignment* (Figure 1a), where the goal is to identify local regions that are similar across the networks being compared. To this end, one defines a scoring function that measures the similarity of a pair of subnetworks, one from each species, in terms of their topology and member proteins. To guide the search for high scoring, or significant matches, the scoring function is often designed to favor a certain class of subnetworks, such as dense subnetworks that serve as a model for protein complexes,^{13,16,32} or paths that serve as a model for protein pathways.^{17,18} In the related *network querying* problem (illustrated in Figure 1b in an astronomical context), a match is sought between a query subnetwork, representing a known functional component of a well-studied species, and a relatively unexplored network of some other organism. The match could be exact (that is, an isomorphic

subgraph under some mapping of the proteins between the two species) or inexact, allowing unmatched nodes on either subnetwork. This problem was first studied by Kelley et al.¹⁷ in the context of local network alignment; its later development accompanied the growth in the number of mapped organisms.^{5,7,9,33} The third problem that has been considered is *global network alignment* (Figure 1c), where one wishes to align whole networks, one against the other.^{4,34} In its simplest form, the problem calls for identifying a 1-1 mapping between the proteins of two species so as to optimize some conservation criterion, such as the number of conserved interactions between the two networks.

All these problems are NP-hard as they generalize graph and subgraph isomorphism problems. However, heuristic, parameterized, and ILP approaches for solving them have worked remarkably well in practice. Here, we review these approaches and demonstrate their good performance in practice both in terms of solution quality and running time.

Heuristic Approaches

As in other applied fields, many problems in network biology are amenable to heuristic approaches that perform well in practice. Here, we highlight two such methods: a local search heuristic for local network alignment and an eigenvector-based heuristic for global network alignment.

NetworkBLAST³² is an algorithm for local network alignment that aims to identify significant subnetwork matches across two or more networks. It searches for conserved paths and conserved dense clusters of interactions; we focus on the latter in our description. To facilitate the detection of conserved subnetworks, NetworkBLAST first forms a network alignment graph,^{17,23} in which nodes correspond to pairs of sequence-similar proteins, one from each species, and edges correspond to conserved interactions (see Figure 2). The definition of the latter is flexible and allows, for instance, a direct interaction between the proteins of one species versus an indirect interaction (via a common network neighbor) in the other species. Any subnetwork of the alignment graph naturally corre-

Figure 2. The NetworkBLAST local network alignment algorithm. Given two input networks, a network alignment graph is constructed. Nodes in this graph correspond to pairs of sequence-similar proteins, one from each species, and edges correspond to conserved interactions. A search algorithm identifies highly similar subnetworks that follow a prespecified interaction pattern. Adapted from Sharan and Ideker.³⁰

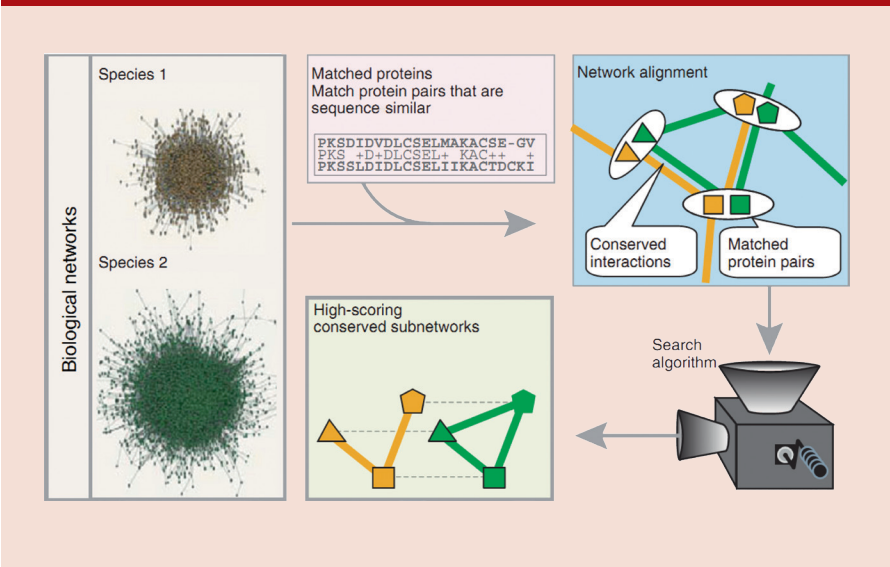
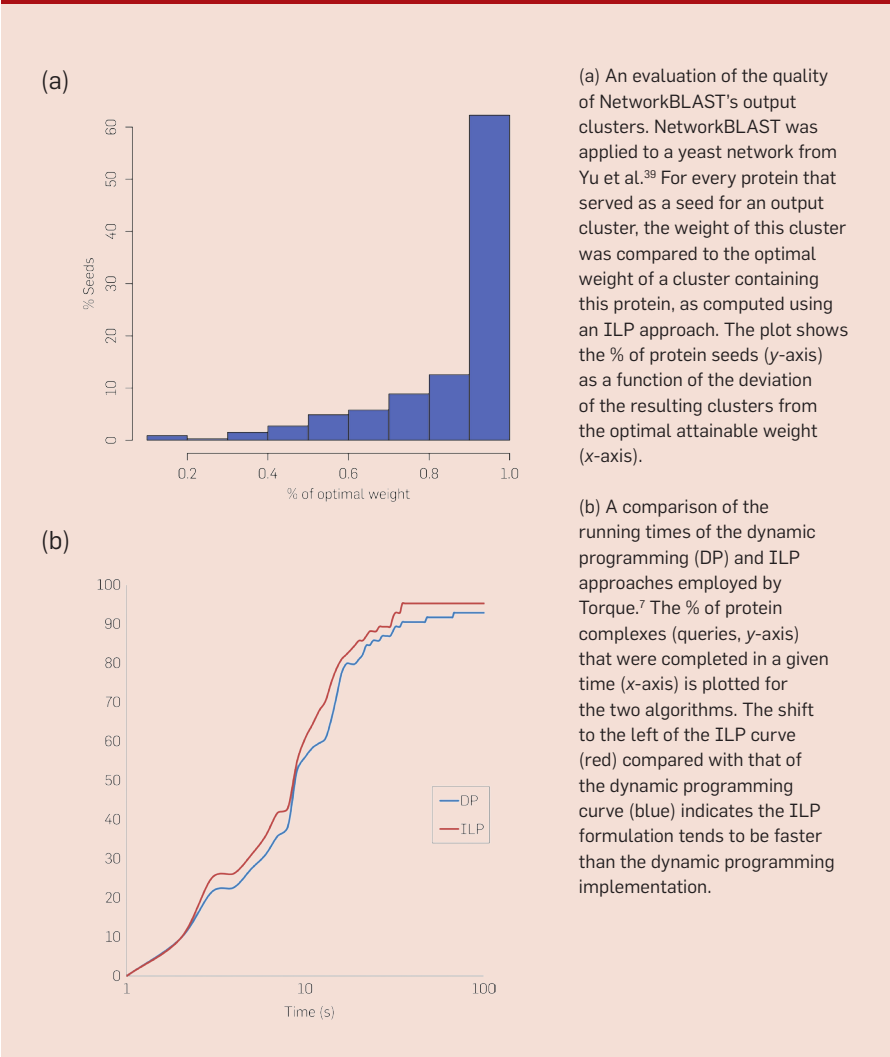


Figure 3. Performance comparison of computational approaches.



(a) An evaluation of the quality of NetworkBLAST's output clusters. NetworkBLAST was applied to a yeast network from Yu et al.³⁹ For every protein that served as a seed for an output cluster, the weight of this cluster was compared to the optimal weight of a cluster containing this protein, as computed using an ILP approach. The plot shows the % of protein seeds (y-axis) as a function of the deviation of the resulting clusters from the optimal attainable weight (x-axis).

(b) A comparison of the running times of the dynamic programming (DP) and ILP approaches employed by Torque.⁷ The % of protein complexes (queries, y-axis) that were completed in a given time (x-axis) is plotted for the two algorithms. The shift to the left of the ILP curve (red) compared with that of the dynamic programming curve (blue) indicates the ILP formulation tends to be faster than the dynamic programming implementation.

Finding a Hairpin in a Haystack

Cells react to stimulation by propagating signals from sensory proteins to a set of target proteins. Given a signaling pathway of a well-studied species, it is interesting to query it within networks of less well-studied species. QPath is an exact path query method that is based on color coding. It extends the basic color coding formulation by allowing one to query a weighted network for inexact matches. That is, each edge of the network has a confidence associated with it, and the goal is to find high-scoring subnetworks that are similar to the query while allowing some flexibility in the matches (see Figure 5b). Specifically, there could be up to N_{ins} insertions of vertices to the match that are not aligned against the query's proteins, and up to N_{del} deletions of vertices from the query that are not aligned against vertices in the matching subnetwork. The algorithm aims at finding an inexact match that optimizes a scoring function that combines: (i) sequence similarity—every pair of matched proteins (q, v) contributes a sequence similarity term $\sigma(q, v)$; (ii) interaction confidence—every edge (u, v) on the matched path contributes a weight term $w(u, v)$; (iii) insertion penalty— c_{ins} per insertion; and (iv) deletion penalty— c_{del} per deletion. The relative importance of each of these terms is learned automatically by QPath; for clarity, we assume that all terms are equally important in the description below.

For a given coloring of the network by $k + N_{\text{ins}}$ colors, QPath employs a dynamic programming formulation to find the highest scoring match with up to N_{ins} insertions and N_{del} deletions. Denote the color of a vertex v by $c(v)$. Denote by $W(i, v, S, \theta_{\text{del}})$ the score of an optimal alignment of the first i nodes of the query that ends at a vertex v in the network, induces θ_{del} deletions, and visits nodes of each color in S . Then,

$$W(i, v, S, \theta_{\text{del}}) = \max_{u \in N(v)} \begin{cases} W(i-1, u, S \setminus \{c(v)\}, \theta_{\text{del}}) + w(u, v) + \sigma(q_i, v) & (u, v) \in E \\ W(i, u, S \setminus \{c(v)\}, \theta_{\text{del}}) + w(u, v) - c_{\text{ins}} & (u, v) \in E \\ W(i-1, v, S, \theta_{\text{del}} - 1) - c_{\text{del}} & 0 < \theta_{\text{del}} \leq N_{\text{del}} \end{cases}$$

The best scoring path is obtained using a standard dynamic programming backtracking starting at

$$\operatorname{argmax}_{v \in V, S \subseteq \{1, \dots, k + N_{\text{ins}}\}, \theta_{\text{del}} \leq N_{\text{del}}} W(k, v, S, \theta_{\text{del}}).$$

sponds to a pair of potentially matching subnetworks. NetworkBLAST scores such a subnetwork by the density of its corresponding intra-species subnetworks versus the chance that they arise at random, assuming a random model that preserves the node degrees.

After constructing the alignment graph, the algorithm proceeds to identify high-scoring subnetworks. This is done by starting with a seed of at most four nodes, and applying a local search to expand it. Each node serves as the center of a seed, along with at most three of its neighbors. The search iteratively adds or removes a node that contributes most to the score, as long as the score increases (and up to an upper bound of 15 nodes). The effectiveness of this search strategy can be quantified by comparing to an exhaustive search when such is possible. Figure 3(a) presents such a comparison when analyzing a single (yeast) network from Yu et al.,³⁹ searching the best cluster containing each of the network's proteins. It can be

seen that the greedy heuristic produces near-optimal clusters (up to 20% deviation in score) in about 75% of the cases, with an average of merely 13% deviation from the optimal score. Notably, NetworkBLAST requires only a few minutes to run, while the exhaustive (ILP-based) approach took several hours while limiting the solver to five minutes per seed. For seven out of a total of 326 seeds, the solver could not find an optimal solution within the allotted time.

While NetworkBLAST can be used to align multiple networks, the size of the alignment graph grows exponentially with the number k of networks and becomes prohibitive for $k = 4$. Interestingly, the NetworkBLAST alignment strategy can be mimicked without having to explicitly construct the alignment graph. Instead, Kalaev et al.¹⁶ show that one can build a linear-size *layered alignment graph* where each layer contains the PPI network of a single species and inter-layer edges connect similar proteins. The main

observation is that a set of proteins that are sequence similar, one from each species, translates to a size- k subgraph that contains a protein (vertex) from each species and is connected through the sequence similarity edges. Such a subgraph must have a spanning tree, which can be looked for using dynamic programming.

To exemplify the algorithm, consider the implementation of NetworkBLAST's local search strategy and let us focus on the addition of new k -protein "nodes" (that is, these would have been nodes of the alignment graph) to the growing seed. The latter requires identification of k inter-species proteins that induce a connected graph on the interlayer edges and contribute most to the seed's weight. As the contribution of each protein to the score can be easily computed and the total contribution is the sum of individual protein contributions, the optimal "node" to be added can be identified in a recursive fashion. That is, the corresponding spanning tree is obtained by merging two neighboring subtrees that span distinct species subsets whose union is the entire species set. This computation takes $O(3^{kl})$ time in total, where l is the number of inter-layer edges.

For global network alignment, both heuristic and exact (ILP) approaches exist. Here, we highlight one such approach by Singh et al.³⁴ that is based on Google's PageRank algorithm. The idea is to score pairs of proteins, one from each species, based on their sequence similarity as well as the similarity of their neighbors. A maximum matching algorithm is then used to find a high-scoring 1-1 alignment between the compared networks.

Singh et al. formulate the computation of the pairwise scores as an eigenvalue problem. Denote by R the score vector to be computed (over all pairs of interspecies proteins). Let $N(v)$ denote the set of nodes adjacent to v , and let A be a stochastic matrix over pairs of inter-species proteins, where $A_{u,v),(u',v')} = 1/|N(u')||N(v')|$ if and only if $\{u, u', v, v'\}$ induce a conserved interaction (that is, (u, u') and (v, v') interact). Finally, denote by B a normalized pairwise sequence similarity vector. The goal is to find a score vector R in which the similarity score of a pair of proteins combines the prior sequence

information with a weighted average of the similarity scores of their neighbors. Thus,

$$R = \alpha AR + (1 - \alpha)B$$

where α is a parameter balancing the network-based and sequence-based terms. R can be found efficiently using the power method algorithm, which iteratively updates R according to the above equation and converges to the analytical solution $R = (I - \alpha A)^{-1}(1 - \alpha)B$.

As mentioned earlier, some manifestations of the global alignment problem can be solved to optimality using ILP. For instance, in Klau,¹⁹ a 1-1 mapping that maximizes the number of conserved edges between the two networks is sought. Interestingly, while the eigenvector solution is heuristic in nature, it performs as well as an exact solution in terms of the number of conserved edges it reveals and its correspondence to a biological ground truth. Indeed, in a recent paper,²² the performance of the heuristic approach of Singh et al. was compared to that of the exact ILP formulation. The algorithms were used to pairwise align the PPI networks of yeast, worm, and fly. Notably, for all three species pairs, the number of conserved edges in the alignment proposed by the heuristic method was equal to that of the ILP approach. As further shown in Mongiovi and Sharan,²² both approaches gave comparable results when their alignments were assessed against a gold standard database of cross-species protein matches (precisely, the HomoloGene database of clusters of orthologous genes²⁹).

Exact Approaches

In contrast to the heuristic methods highlighted here, which do not provide any guarantee of the quality of the obtained solution, exact approaches guarantee optimality at the cost of speed. Two general methodologies for efficient, yet exact, solutions have been common in network analysis: fixed parameter and ILP formulations. Fixed parameter tractable problems can be solved in time that is, typically, exponential in some carefully chosen parameter of the problem and polynomial in the size of the input networks. As we will describe, many

variants of the network querying problem are amenable to fixed parameter approaches, as the query subnetworks can be often assumed to have a small, bounded size. The other methodology we demonstrate is based on reformulating the problem at hand as an integer linear program and applying an industrial solver such as CPLEX¹⁴ to

optimize it. While arriving at a solution in a timely fashion is not guaranteed (as integer programming is NP-hard¹¹), in practice, on current networks, many of these formulations are solved in reasonable time.

We start by describing a parameterized approach—color coding—that has been extensively used in network

Querying via an Integer Linear Program

In the biological domain, where one wishes to query known protein machineries in the network of another species, oftentimes the topology of the query is not known (only the identity of the member proteins is known). One possible way to tackle this scenario, applied by Torque,⁷ is to assume that the query and, hence, the sought matches are connected. Torque is based on an ILP, which expresses the connectivity requirement by simulating a flow network, where an arbitrary node serves as a sink draining the flow generated by all the other nodes that participate in the solution. In detail, the ILP formulation uses the following variables: (i) a binary variable c_v for each node v , denoting whether it participates in the solution; (ii) a binary variable e_{uv} for each edge (u, v) , denoting whether it participates in the solution subnetwork; (iii) a pair of rational variables f_{uv}, f_{vu} for each edge (u, v) , representing both the magnitude and direction of the flow going through it; (iv) a binary variable r_v that marks the sink node; and (v) a binary variable g_{vq} for every pair of sequence-similar network and query nodes (v, q) , denoting whether q is matched with v .

The following set of constraints is immediately derived from the model and expresses the requirements that (i) the solution should span k nodes; (ii) only one node may serve as a sink; and (iii) an edge is part of the solution only when its two endpoints are:

$$\begin{aligned} \sum_{v \in V} c_v &= k \\ \sum_{v \in V} r_v &= 1 \\ e_{vu} &\leq \frac{1}{2}c_v + \frac{1}{2}c_u \quad \forall (v, u) \in E \end{aligned}$$

Let Q denote the set of query proteins, $|Q| = k$, and let $\Phi(v) \subseteq Q$ denote the (possibly empty) subset of query proteins that are sequence similar to v . To obtain an adequate match between the solution and query proteins, the following constraints are added to ensure that (i) a network protein may match at most one query protein; (ii) all query proteins are matched with exactly one network protein (assuming, for simplicity, that there are no insertions or deletions); and (iii) only nodes that are part of the solution may be matched.

$$\begin{aligned} \sum_{q \in \Phi(v)} g_{vq} &\leq 1 \quad \forall v \in V \\ \sum_{v \in V} g_{vq} &= 1 \quad \forall q \in Q \\ g_{vq} &\leq c_v \quad \forall v \in V, q \in \Phi(v) \end{aligned}$$

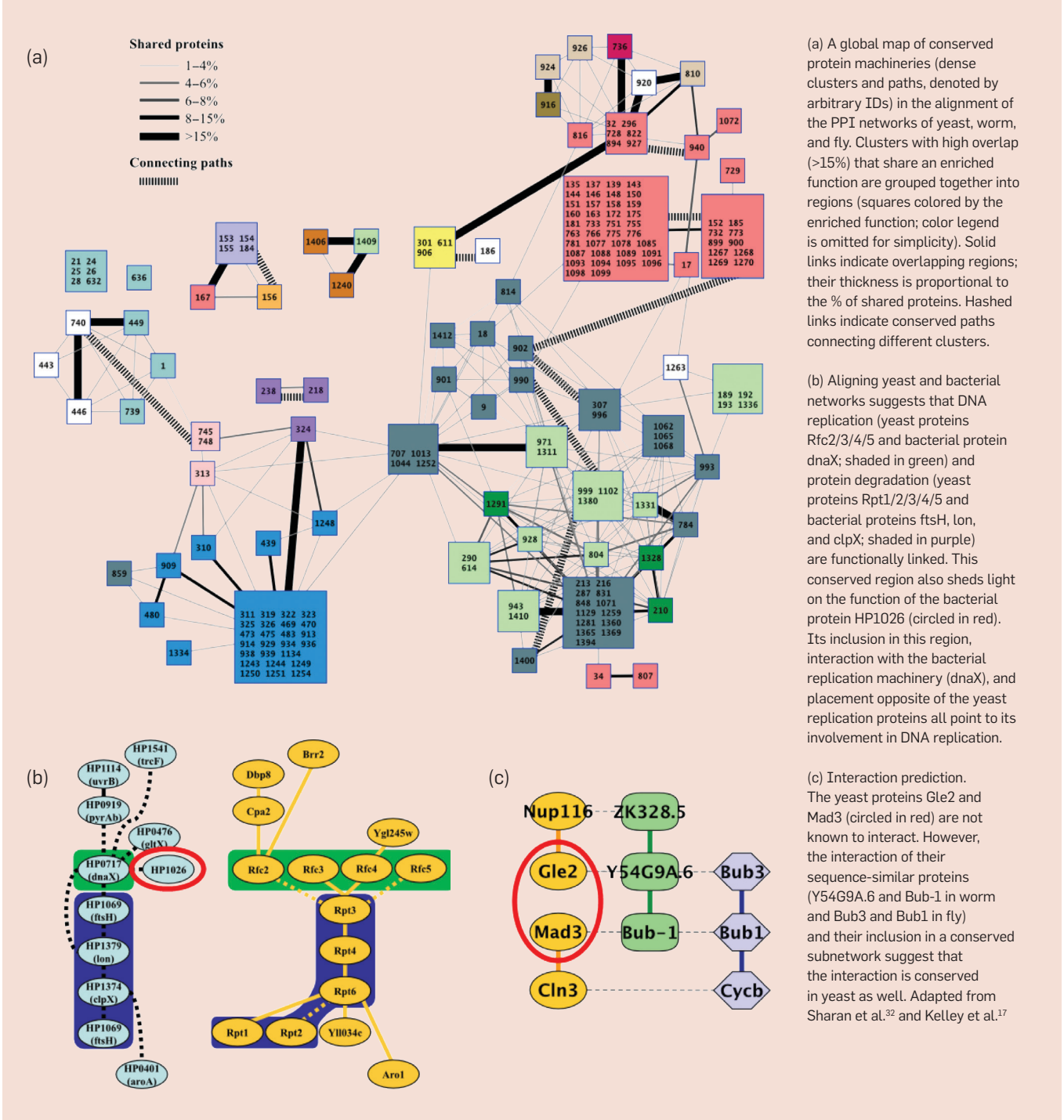
To maintain a legal flow, one also needs to ensure that (i) the pair of flow variables associated with a given edge agrees on its direction and magnitude; (ii) the flow may only pass through edges that participate in the solution; and (iii) source nodes generate flow that is drained by the sink. These conditions are formulated by the following constraints:

$$\begin{aligned} f_{vu} &= -f_{uv} \quad \forall (v, u) \in E \\ f_{vu}, f_{uv} &\leq (k-1)e_{vu} \quad \forall (v, u) \in E \\ \sum_{u \in N(v)} f_{vu} &= c_v - k \cdot r_v \quad \forall v \in V \end{aligned}$$

Together, the above constraints restrict the solutions to take the form of a connected subnetwork spanning exactly k nodes that are sequence similar to their respective matches in the query. Finally, denoting the weight of edge (u, v) by $w(u, v)$, the objective is to maximize the weight of the solution subnetwork:

$$\max \sum_{(u,v) \in E} w(u,v)e_{uv}$$

Figure 4. Insights derived from a multiple network alignment.



querying applications. Color coding was originally developed by Alon et al.² for searching for structured size- k subgraphs, such as simple paths and trees, within a graph. Its complexity is $2^{O(k)}m$, where m is the size of the searched graph. Color coding is based on the idea that by randomly assigning k distinct colors to the vertices of the graph, the task of finding a simple subgraph translates to that of finding a *colorful*

subgraph, namely, one spanning k distinct colors. For certain classes of tree-like subgraphs, the subsequent search can be efficiently implemented using dynamic programming. Since a particular subgraph need not be colorful in a specific color assignment, multiple color assignments should be considered to retrieve a desired subgraph with high probability. Precisely, the probability that a graph of size k is

colorful is $k!/k^k > e^{-k}$; hence, in expectation, e^k iterations of the algorithm suffice to detect the desired subgraph.

In the context of comparative network analysis, color coding was mainly used to tackle network querying problems, where typically the query subnetwork is small (5–15 proteins), motivating the use of this parameterized approach. One specific example is the QPath³³ method for querying

paths in a network. QPath extends the basic color coding formulation by querying weighted networks for inexact matches (see the accompanying sidebar “Finding a Hairpin in a Haystack”). The algorithm takes minutes to run with length-7 queries and up to three insertions (unaligned match vertices) and three deletions (unaligned query vertices). Efficient heuristics to color the network can be used to reduce its time even further.^{9,21} In a follow-up work,⁹ the QPath approach was extended to handle queries of bounded treewidth and a heuristic solution was offered for general queries.

Our last highlighted method uses an ILP formulation to optimally solve a different variant of the network querying problem, where the topology of the query is not known. This scenario is very common when querying for protein complexes, where the underlying interaction structure is rarely available³⁹ but the member proteins are assumed to be connected. Hence, instead of searching for a particular interaction pattern, the goal is to find a matching subgraph that is *connected*. In Bruckner et al.,⁷ an ILP solution to this problem is given. The main challenge in formulating the problem as an ILP is to express the connectivity of the solution sought. The Torque algorithm⁷ solves this problem by modeling the solution subgraph as a flow network, where one of its nodes is arbitrarily designated as a sink, capable of draining $k - 1$ units of flow, and all the other nodes are set as sources, each generating one unit of flow. A set of constraints requires that the total flow in the system is preserved. The detailed program is given in the accompanying sidebar “Querying via an Integer Linear Program.”

Notably, there is also a parameterized approach to this querying problem. The approach is based on the observation that a connected subgraph can be represented by its spanning tree, so the querying problem translates to that of finding a tree of k distinct vertices. The latter problem can be solved using the color coding technique.^{6,7} Interestingly, for most instances, the dynamic programming approach is empirically slower than running the ILP formulation through a solver, as demonstrated in Figure 3(b).

The Power of Comparative Network Analysis

The successful application of comparative network analysis approaches depends not only on their computational attributes but also on their biological relevance. Here, we give examples for the applications of several of the reviewed approaches and the biological insights they have enabled. We demonstrate the power of comparative network analysis approaches by comparing their performance with that of methods that are either sequence-based and, thus, cannot exploit the network information, or single-species-based and as a result are more prone to noise in the network data.

The most intuitive use of comparative network analysis is to gather support for computational predictions from multiple species. A prime example for this use is the inference of protein complexes or pathways from PPI data. For instance, in Sharan et al.,³¹ a cross-species analysis is used to identify yeast-bacterial conserved complexes. By comparing the inferred complexes to known complexes in yeast, it is shown that the comparative analysis increases the specificity of the predictions (compared to a yeast-only analysis) by a significant margin, albeit at the price of reduced sensitivity.

In Sharan et al.,³² the local alignment of three networks (yeast, worm, and fly) was used to identify their conserved protein machineries (Figure 4a); those were used in a systematic manner to infer protein function (Figure 4b) and interaction (Figure 4c), showing superior performance compared to that of a sequence-based analysis. In brief, NetworkBLAST was used to identify high-scoring triplets of matching subnetworks across the three networks. Whenever the proteins in a conserved subnetwork were enriched for a certain function and at least half the proteins in the subnetwork were annotated with that function, the rest of the subnetwork's proteins were predicted to have that function as well. This prediction strategy yielded 58%–63% accuracy across the three species (in a cross-validation test), versus 37%–53% accuracy for a sequence-based method that predicts the function of a protein based on its most sequence-similar protein in the other species. The conserved

subnetworks were further used to predict novel PPIs in the following manner: a pair of proteins were predicted to interact if two sequence-similar proteins were known to interact in another species (directly or via a common network neighbor) and, additionally, if the four proteins co-occurred in one of the conserved subnetworks. Remarkably, this strategy yielded >99% specificity in cross-validation. Experimental validation of 65 of these predictions gave a success rate in the range of 40%–52%. In comparison, sequence-based predictions that do not use the conserved subnetwork information yielded success rates in the range of 16%–31%.^{18,32}

The transfer of annotations across species can go beyond single proteins to whole subnetworks. For instance, in Shlomi et al.,³³ paths in the yeast network served as queries for the fly network. The resulting matches were annotated with the function of the query (whenever the query was significantly enriched with some functional annotation; see Figure 5a), and the predictions were tested versus the known functional annotations in the fly. Overall, the annotation was accurate in 64% of the cases, compared to 40% for sequence-based annotation.

Network comparison schemes can be used to gain additional insights on protein function, interaction, and evolution. Both local and global network alignments suggest aligned pairs (or, more generally, sets) of proteins as functionally similar. Accordingly, they have been used to identify proteins that have evolved from a common ancestor and retained their function (so called *functional orthologs*).^{4,38} In Bandyopadhyay et al.,⁴ it is shown that in a majority of the cases, the aligned pairs were not the highest sequence similar ones. In addition, the conserved subnetworks often connect cellular processes that may work together in a coordinated manner (Figure 4b). The evidence is particularly compelling when the link is supported by multiple species.

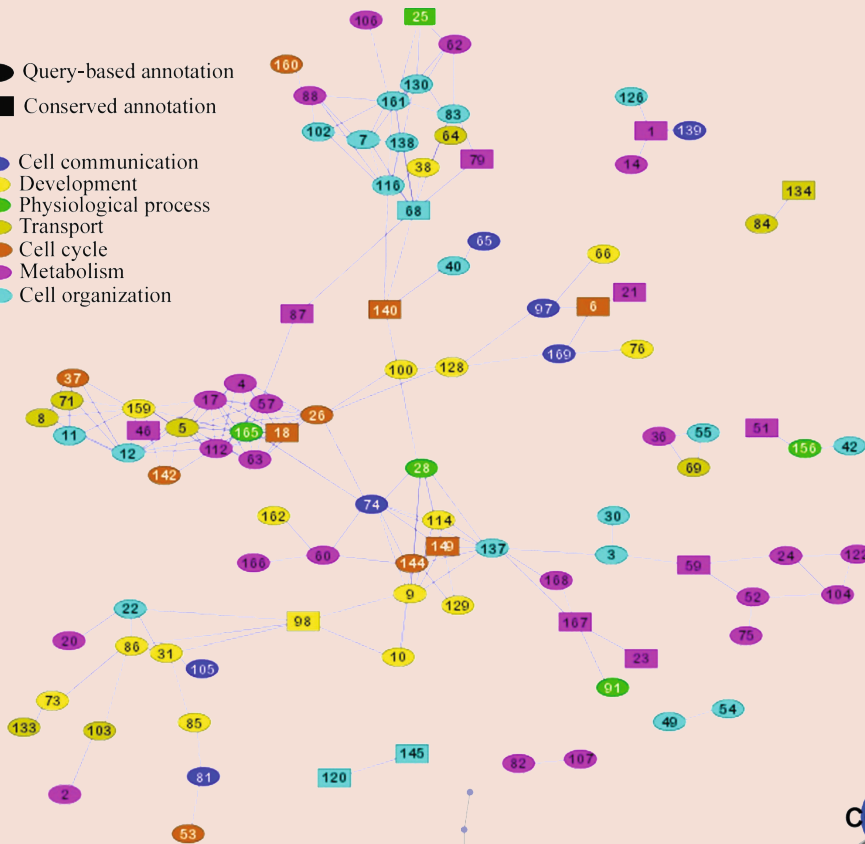
Conclusion

The explosion of molecular data in the last two decades is revolutionizing biological and, consequently, computational research. The arising computational problems require

Figure 5. Path queries.

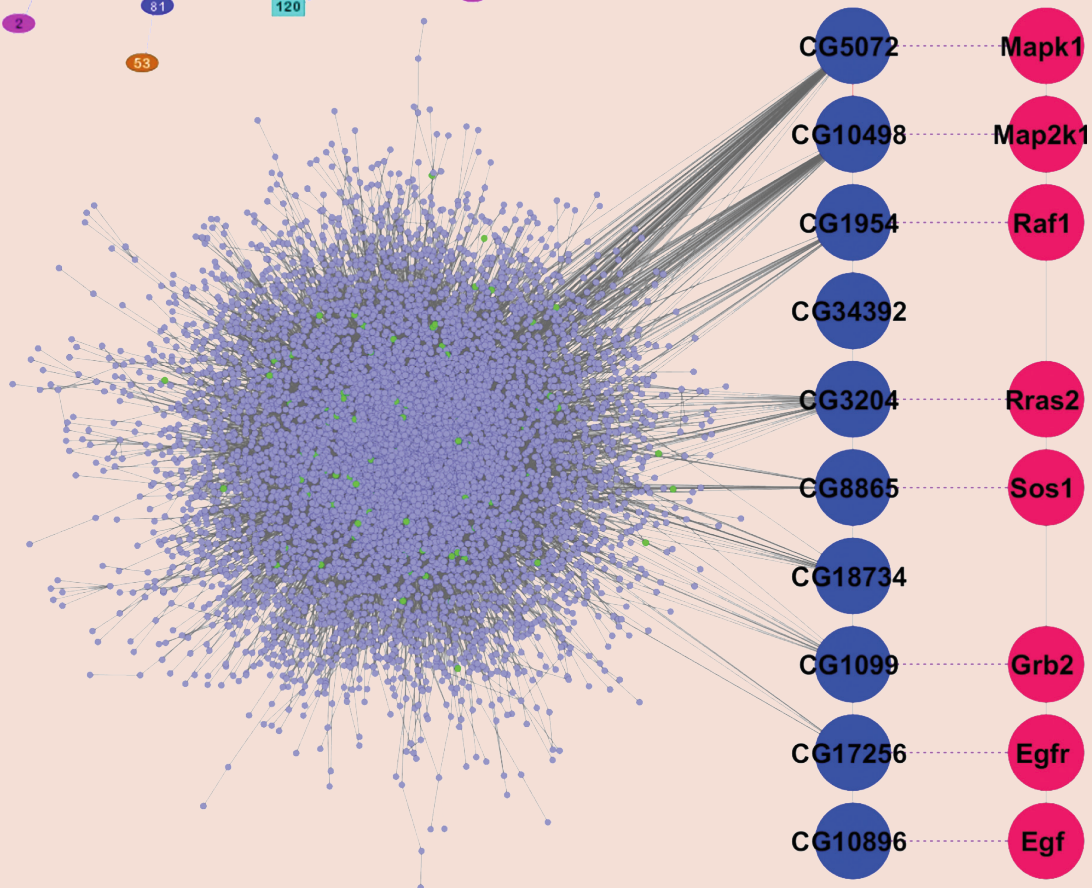
(a)

- Query-based annotation
- Conserved annotation
- Cell communication
- Development
- Physiological process
- Transport
- Cell cycle
- Metabolism
- Cell organization



(a) A yeast-fly conserved pathway map. Pathways from yeast were used to query the fly network. Nodes represent best-match pathways and are connected if they share more than two proteins. Nodes are colored according to their predicted function based on the yeast query proteins. Best-match pathways in which significantly many proteins are annotated with the predicted function appear as boxes.

(b)




(b) A querying example. QPath was applied to query the human MAPK pathway (red nodes) within the fly network (blue nodes). The best scoring pathway (dark blue nodes) contains two insertions relative to the query. Fly proteins that are similar to the query are shown in light green. Adapted from Shlomi et al.³³

practical solutions that can cope with an evergrowing scale. In addition to heuristic approaches, it is often of interest to compute exact solutions that can potentially lead to new insights about the biological problem at hand. The combination of parameterized approaches and powerful linear programming solvers has enabled the development of efficient, yet exact, methods to solve some of the key problems in comparative network analysis.

The application of comparative analysis tools to available network data has provided valuable insights on the function and interplay among molecular components in the cell. While much progress has already been made, new computational techniques will need to be developed to cope with the flood of genomic data that is expected to arrive in the coming years. These will span thousands of organisms and diverse molecular aspects. The arising challenges will involve the organization of this data into high-quality networks, data imputation through the integration of multiple information sources, multiple network alignment, and, ultimately, the propagation of curated or experimentally derived annotations through the aligned networks. Hybrid solution approaches that try to combine different techniques depending on the problem instance (see, for example, Bruckner et al.⁷) may be key to meeting those challenges.

Acknowledgments

We thank Sharon Bruckner for contributing Figure 2b. Atias was partially funded by the Edmond J. Safra Bioinformatics Program. This work was supported by a research grant from the Israel Science Foundation (Grant no. 241/11). 

References

- Aebbersold, R., Mann, M. Mass spectrometry-based proteomics. *Nature* 422 (2003), 198–207.
- Alon, N., Yuster, R., Zwick, U. Color-coding. *J. ACM* 42 (July 1995), 844–856.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* 215, 3 (Oct. 1990), 403–410.
- Bandyopadhyay, S., Sharan, R., Ideker, T. Systematic identification of functional orthologs based on protein network comparison. *Genome Res.* 16, 3 (Mar. 2006), 428–435.
- Banks, E., Nabieva, E., Peterson, R., Singh, M. Netgrep: fast network schema searches in interactomes. *Genome Biol.* 9 (2008), R138.
- Betzler, N., Fellows, M.R., Komusiewicz, C., Niedermeier, R. Parameterized algorithms and hardness results for some graph motif problems. In *Proceedings of the 19th annual symposium on Combinatorial Pattern Matching* (Berlin, Heidelberg, 2008), CPM '08, Springer-Verlag, 31–43.
- Bruckner, S., Hffner, F., Karp, R.M., Shamir, R., Sharan, R. Topology-free querying of protein interaction networks. *J. Comput. Biol.* 17, 3 (Mar. 2010), 237–252.
- Deng, M., Sun, F., Chen, T. Assessment of the reliability of protein-protein interactions and protein function prediction. In *Proceedings of the 8th Pacific Symposium on Biocomputing* (2003), 140–151.
- Dost, B., Shlomi, T., Gupta, N., Rupp, E., Bafna, V., Sharan, R. QNet: a tool for querying protein interaction networks. *J. Comput. Biol.* 15, 7 (Sep. 2008), 913–925.
- Fields, S. High-throughput two-hybrid analysis: the promise and the peril. *FEBS J.* 272 (2005), 5391–5399.
- Garey, M., Johnson, D. Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman and Co., San Francisco, 1979.
- Giot, L., Bader, J., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y., Ooi, C., Godwin, B., Vitols, E., Vijayadomodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrotta, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley, R.L., Jr, White, K., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shinkets, R., McKenna, M., Chant, J., Rothberg, J. A protein interaction map of *Drosophila melanogaster*. *Science* 302, 5651 (2003), 1727–1736.
- Hirsh, E., Sharan, R. Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics* 23 (2007), e170–e176.
- IBM. IBM ILOG CPLEX V12.1 user's manual for CPLEX, 2009.
- Ideker, T., Sharan, R. Protein networks in disease. *Genome Res.* 18 (2008), 644–652.
- Kalaev, M., Bafna, V., Sharan, R. Fast and accurate alignment of multiple protein networks. *J. Comput. Biol.* 16 (2009), 989–999.
- Kelley, B.P., Sharan, R., Karp, R.M., Sittler, T., Root, D.E., Stockwell, B.R., Ideker, T. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci. U.S.A.* 100, 20 (Sep. 2003), 11394–11399.
- Kelley, B.P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B.R., Ideker, T. PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res.* 32, Web Server issue (Jul. 2004), W83–W88.
- Klau, G.W. A new graph-based method for pairwise global network alignment. *BMC Bioinformatics* 10, Suppl 1 (2009), S59.
- Li, S., Armstrong, C., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P., Han, J.D., Chesneau, A., Hao, T., Goldberg, D., Li, N., Martinez, M., Rual, J., Lamesch, P., Xu, L., Tewari, M., Wong, S., Zhang, L., Berriz, G., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H., Elewa, A., Baumgartner, B., Rose, D., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S., Saxton, W., Strome, S., Heuvel, S.V.D., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K., Harper, J., Cusick, M., Roth, F., Hill, D., Vidal, M. A map of the interactome network of the metazoan *C. elegans*. *Science* 303, 5657 (2004), 540–543.
- Mayrose, I., Shlomi, T., Rubinstein, N.D., Gershoni, J.M., Rupp, E., Sharan, R., Pupko, T. Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm. *Nucleic Acids Res.* 35, 1 (2007), 69–78.
- Mongiovi, M., Sharan, R. Global alignment of protein-protein interaction networks. *Data Mining for Systems Biology*. H. Mamitsuka, C. DeLisi, and M. Kanehisa, eds. Springer, 2012, in press.
- 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 (2000), 4021–4028.
- Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D., Yeates, T. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.* 96 (1999), 4285–4288.
- Rain, J., Selig, L., Reuse, H.D., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A., Legrain, P. The protein-protein interaction map of *Helicobacter pylori*. *Nature* 409 (2001), 211–215.
- Ross, C.A., Tabrizi, S.J. Huntington's disease: from molecular pathogenesis to clinical treatment. *Lancet Neurol.* 10, 1 (Jan. 2011), 83–98.
- Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Lammas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P., Vidal, M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 7062 (Oct. 2005), 1173–1178.
- Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D. The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32, Database issue (Jan. 2004), D449–D451.
- Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S., Feolo, M., Fingerman, I.M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Lu, Z., Madden, T.L., Madej, T., Maglott, D.R., Marchler-Bauer, A., Miller, V., Mizrahi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Wang, Y., Wilbur, W.J., Yaschenko, E., Ye, J. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 39, Database issue (Jan. 2011), D38–D51.
- Sharan, R., Ideker, T. Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.* 24, 4 (Apr. 2006), 427–433.
- Sharan, R., Ideker, T., Kelley, B., Shamir, R., Karp, R. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J. Comput. Biol.* 12 (2005), 835–846.
- Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M., Ideker, T. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. U.S.A.* 102, 6 (Feb. 2005), 1974–1979.
- Shlomi, T., Segal, D., Rupp, E., Sharan, R. QPath: a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics* 7 (2006), 199.
- Singh, R., Xu, J., Berger, B. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl. Acad. Sci. U.S.A.* 105, 35 (Sep. 2008), 12763–12768.
- Stark, C., Breitkreutz, B.-J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Aukun, K.V., Wang, X., Shi, X., Reguly, T., Rust, J.M., Winter, A., Dolinski, K., Tyers, M. The BioGRID interaction database: 2011 update. *Nucleic Acids Res.* 39, Database issue (Jan. 2011), D698–D704.
- Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., Wanker, E. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122 (2005), 957–968.
- Uetz, P., Giot, L., Cagney, G., Mansfield, T., Judson, R., Knight, J., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadomodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 6770 (2000), 623–627.
- Yosef, N., Sharan, R., Noble, W.S. Improved network-based identification of protein orthologs. *Bioinformatics* 24, 16 (Aug. 2008), i200–i206.
- Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R.R., Simon, C., Tardivo, L., Tam, S., Szvrikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M.E., Park, J., Xin, X., Cusick, M.E., Moore, T., Boone, C., Snyder, M., Roth, F.P., Barabasi, A.-L., Tavernier, J., Hill, D.E., Vidal, M. High-quality binary protein interaction map of the yeast interactome network. *Science* 322, 5898 (Oct. 2008), 104–110.

Nir Atias (atias@post.tau.ac.il) is a Ph.D. candidate in the Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv Israel.

Roded Sharan (roded@post.tau.ac.il) is a professor in the Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv Israel.

© 2012 ACM 0001-0782/12/05 \$10.00