Briefings in Bioinformatics Advance Access published December 3, 2013

doi:10.1093/bib/bbt084

Searching for repetitions in biological networks: methods, resources and tools

Simona Panni and Simona E. Rombo

Submitted: 20th August 2013; Received (in revised form): 26th October 2013

Abstract

We present here a compact overview of the data, models and methods proposed for the analysis of biological networks based on the search for significant repetitions. In particular, we concentrate on three problems widely studied in the literature: 'network alignment', 'network querying' and 'network motif extraction'. We provide (i) details of the experimental techniques used to obtain the main types of interaction data, (ii) descriptions of the models and approaches introduced to solve such problems and (iii) pointers to both the available databases and software tools. The intent is to lay out a useful roadmap for identifying suitable strategies to analyse cellular data, possibly based on the joint use of different interaction data types or analysis techniques.

Keywords: *biological networks analysis; network global alignment; network local alignment; asymmetric alignment; network querying; network motif extraction*

INTRODUCTION

The organization and functioning of cells relies on the interplay of several different factors, among which the specific association of biological components in networks has been demonstrated to be one of the most important. While in past decades attention was mainly focused on the study of single molecules, such as proteins, genes and RNA [1], a growing body of evidence suggests that cellular components cannot be analysed as independent objects when they take part in common biological processes [2]. Furthermore, studying the interactions between genes, as well as between their corresponding protein products, may help in the prediction of gene–phenotype relationships and in understanding the emergence of diseases [3–5].

The explosion of interaction data obtained by experimental and computational techniques required the proposal of efficient and effective approaches to extract useful knowledge from them. Interaction data are usually modelled by suitable graphs, called 'biological networks', such that nodes are associated with cellular components, and edges represent pairwise interactions.

Several algorithms analysing such graphs have been designed, implemented and applied to interaction data. In this article, we consider the problem of singling out 'conservation' from biological networks, although other problems have been defined over this domain (e.g. clustering [6–12] or integration [13]). Conservation here is in terms of repeated substructures of interactions occurring among different networks or across the same graph. Note that, in general, the presence of repeated substructures is often associated with relevant conservation in the biological context, as witnessed by the large attention devoted to the discovery of interesting repetitions in sequences (e.g. [14–18]), useful to model cellular components.

Our goal is to provide a general overview of the bioinformatics resources currently available for the search of repetitions in biological networks. This

Corresponding author. Simona E. Rombo, Department of Mathematics and Computer Science, University of Palermo, Italy. Tel: +39 091 23891028; Fax: +39 091 23891024; E-mail: simona.rombo@math.unipa.it

Simona Panni is an assistant professor at the DiBEST Department (Biology Ecology Earth Sciences) of the University of Calabria. Her research interests focus on protein–protein interactions with special regard to those mediated by protein binding domains.

Simona E. Rombo is an assistant professor at the Department of Mathematics and Computer Science, University of Palermo. Her research interests include algorithms and data structures, bioinformatics and data mining.

problem can be subdivided in three main subproblems, which are as follows:

- 'Network alignment', i.e. extracting 'similar' subnetworks in two or several input networks, possibly associated with different organisms, to uncover complex mechanisms at the basis of evolutionary conservations, or to infer the biological meanings of groups of interacting cellular components belonging to organisms not yet well characterized [13].
- 'Network querying', i.e. searching for the occurrences of a small network in other, larger, input networks. A typical application is to study how a specific module of a model organism differentiated in more complex organisms [19].
- 'Motif extraction', i.e. searching for repeated modules across the same network, since several studies have proved that biological networks can often be understood in terms of coalitions of basic repeated building blocks [20,21].

Previous surveys have been proposed focusing on specific aspects of the extraction of repetitions from biological networks, often dealing with only one of the above mentioned subproblems [13,22–26]. The resulting landscape is rather fragmentary, despite the correlations among several aspects characterizing such subproblems. We aim at providing a broad vision of the scope, and thus we include network alignment, network querying and motif finding in our analysis. In particular, we present an overview of the main resources that have been collected in the past few years, in terms of data types, models and available databases (Section 2), as well as problems, related approaches and available software tools (Section 3). We also provide in Section 4 a comparative discussion on the techniques presented here, by highlighting in which contexts and circumstances they can be applied. This will hopefully make this review useful to bioinformatics researchers applying the existing resources and possibly combining different interaction data types or analysis techniques, that, in our opinion, is one of the main open challenges in this context (Section 5).

RESOURCES AND MODELS FOR INTERACTION DATA

High-throughput experimental techniques [27,28] and computational methods [2,29] both contribute to the collection of cellular component interactions stored in public databases (e.g. [30,31]). To model them, suitable graphs, where interacting components are linked together, are usually used. We briefly recall some basics on graphs in the Supplementary Material—further insights can be found in, e.g. [32]. In the following, first we describe the different kinds of interaction data, and then we list the public databases where the interaction data are stored and finally we summarize the main models proposed for the analysis of interaction data.

Types of data

The main categories of biological interaction data are described below. The interested reader can find further details on the experimental methods in the Supplementary Material.

Protein–protein interactions

Protein–protein interactions (PPI) occur when two or more proteins bind together to carry out their biological function. Almost all molecular processes are carried out by protein complexes organized by specific PPI. These interactions can be detected using many different experimental approaches, a few of which can be automated to perform high-throughput experiments.

Pioneering 'interactomics' studies were performed on the model organism *Saccharomyces cerevisiae* using the two-hybrid approach [33], and to date this is one of the most used methods in the detection of binary interactions [34]. The affinity purification approach [35] instead permits isolating protein complexes using an antibody that specifically recognizes one molecule of the complex, but gives no information about direct binding [36–38]. Other approaches are based on pools of peptides of different sequences (peptide libraries) that are utilized to collect information on the binding specificity of a protein of interest [34–36].

Genes, reactions and pathways

Metabolic processes determining the physiological and biochemical properties of a cell are modelled using metabolic pathways. A metabolic pathway is a set of biochemical reactions, each catalysed by a different enzyme, which transforms the initial substrates (metabolites and chemical compounds) in the final products through a chain of subsequent modifications. Regulatory genes mediate the assembling of functional complexes to direct enzymes to their

targets and to compartmentalize molecular components [42,43]. The main metabolic pathways have been well characterized by biochemical studies and they are conserved among organisms, so that no high-throughput strategies have been introduced to produce these data.

Gene regulatory data

Transcription factors are proteins binding to particular DNA sequences to regulate the expression of specific genes. Usually each factor can activate or repress a number of different genes, but has no activity on the others. Thus, the phenotype of a cell, as well as its capability to respond to environmental signals, results from the integrated action of transcription factors on the genome. Many techniques have been developed to outline which genes are regulated by a transcription factor. Among them, two complementary approaches can be automated to obtain large-scale regulatory data: the yeast one-hybrid, where a DNA sequence is used as bait to screen binding regulatory factors [44], and the chromatin immunoprecipitation followed by DNA sequencing (ChIP-Seq), where the immunoprecipitation of a transcription factor allows the identification of bound DNA regulatory sequences [45]. Both these approaches have been used to build genome-wide regulatory maps, e.g. in yeast [46], *Caenorhabditis elegans* [47] and other organisms [48].

Disease annotation data

Identifying the genes involved in the onset of a disorder is the first step to understanding the disease mechanisms. Meta-analyses of published genetic associations, together with the new genome-wide association studies, have provided an abundance of information on 'risk alleles' and on genetic associations between genes and diseases, which are catalogued in the Online Mendelian Inheritance in Man (OMIM) database [49]. A single gene can influence many pathologies and, at the same time, human diseases are often the consequence of the perturbation of multiple cellular components. Moreover, when two or more genes are associated with the same disorder, the corresponding proteins show a high propensity to interact [50].

Available databases and benchmarks

Table 1 shows the main public interaction databases. We selected those databases currently updated and containing only experimentally validated interactions, except for some protein–protein interaction databases (e.g. [51–53]) storing both integrated interaction data and computationally predicted interactions.

Types of networks

Once the interaction data have been produced and stored, they can be analysed through the use of computational approaches. To this end, they are modelled as biological networks. The main types of biological networks are described below, and how they are usually represented is also specified.

Protein–protein interaction networks

The set of all the PPI of a given organism is its 'interactome', usually modelled by an undirected graph called a 'protein–protein interaction network' (PPI network), where nodes represent the involved proteins and edges encode their interactions (Figure 1).

Nodes can be labelled by protein names or IDs, and edges may be labelled by interaction reliability scores provided by the databases. Such scores are obtained by combining different information, such as the confidence of the techniques applied to discover a specific interaction, or the fact that the same interaction is confirmed by different experimental techniques.

Metabolic networks

A metabolic network may be modelled using a bipartite graph, where the two sets of nodes represent chemical reactions and substrates (metabolites or compounds), respectively. Alternatively, a metabolic network can also be represented using a graph where the vertices represent the substrates and information on the reactions is stored as edge labels. According to another representation, reactions are stored as vertices and information on the substrates is stored as edge labels. When directed versions of these graphs are considered, the directed edges express the reversibility/irreversibility of some reactions.

Gene regulatory networks

Gene regulatory networks describe the interactions between transcription factor proteins and the genes that they regulate. They can be represented as directed graphs, with two sets of nodes: the transcription factors and the genes that they regulate. The edges indicate the binding of the transcription factors to the gene regulatory elements and they can be directed from the transcription factor towards the DNA regulatory element (incoming) or from the

Table 1: Summary of the publicly available databases

Source	Type of data	Link
AURA [54]	GRD	http://aura.science.unitn.it/
BiGG [55]	GRP	http://bigg.ucsd.edu/bigg/home.pl
BIOcyk [56]	GRP	http://biocyc.org/
BIOGRID [57]	PPI	http://thebiogrid.org/
DIP [58]	PPI	http://dip.doe-mbi.ucla.edu/dip/Main.cgi
ENCODE [59]	GRD	http://genome.ucsc.edu/ENCODE/
GenMAPP [60]	GRP and GRD	http://www.genmapp.org/introduction.html
HAPPI [5I]	PPI	http://discern.uits.iu.edu:8340/HAPPI/
HPD [6I]	GRP	http://discern.uits.iu.edu:8340/HPD/
HPRD [62]	PPI	http://www.hprd.org/
HUPHO [63]	PPI	http://hupho.uniroma2.it/index.php
KEGG [3I]	GRP	http://www.genome.jp/kegg/
INTACT [64]	PPI	http://www.ebi.ac.uk/intact/
ITFP [65]	GRD	http://itfp.biosino.org/itfp/
Mentha [66]	PPI	http://mentha.uniroma2.it/index.php
MINT [30]	PPI	http://mint.bio.uniroma2.it/mint/
MIPS [67]	PPI	http://mips.helmholtz-muenchen.de/genre/proj/yeast/
OMIM database [49]	DAD	http://www.ncbi.nlm.nih.gov/omim
Pathway Commons 2 [68]	GRP	http://www.pathwaycommons.org/pc2/
PCDq [53]	PPI	http://h-invitational.jp/hinv/pcdq/
Reactome [69]	PPI and GRP	http://www.reactome.org/ReactomeGWT/entrypoint.html
STRING [52]	PPI	http://string-db.org/
TRED [70]	GRD	http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home
UniPathway [7I]	GRP	http://www.grenoble.prabi.fr/obiwarehouse/unipathway
UniPROBE [72]	GRD	http://the.brain.bwh.harvard.edu/uniprobe/

Note: Columns: (I) database acronym and reference; (II) type of stored data; (III) database URL.
PPI, protein–protein interactions; GRP, genes, reactions and pathways; GRD, gene regulatory data; DAD, disease annotation data.

DNA element towards the transcription factor (outgoing).

A gene regulatory network may also be represented as a 'connectivity matrix' M, such that $Mij = 1$ if the component associated with the node j encodes a transcription factor regulating the component associated with the node i, and $Mij = 0$ otherwise.

Disease networks

Disease networks may be represented by bipartite graphs, built from a set of genetic diseases and a set of disease genes [50]. According to a different representation, an undirected graph may be considered where nodes represent the diseases, while edges linking two nodes indicate that they have in common at least one gene. Symmetric representation is also used, such that nodes represent the genes and two genes are linked when they are associated with the same disorder. According to other representations [74], a metabolic disease network may be built, where disorders are linked if the corresponding mutated enzymes are involved in related pathways.

PROBLEMS AND METHODS

We now describe the main problems that have been defined in the literature, concerning the identification of repetitions in biological networks, and also present the most recent techniques proposed to solve them.

Alignment

Given two input networks, the alignment problem consists of finding a set of conserved edges across them, leading to a (not necessarily connected) conserved subgraph. In this case, the problem is also known as 'pairwise alignment'. 'Multiple alignment' is a natural extension when n networks are considered as the input; however, this is usually computationally more expensive to perform. Biological network alignment can be further distinguished into 'global alignment' and 'local alignment'.

Global alignment (Figure 2) aims at finding a unique (possibly the best) overall alignment across the input networks, in such a way that a one-to-one correspondence is found among their nodes. The result is a set of pairs (or tuples) of non-overlapping subgraphs. Local alignment (Figure 3) aims

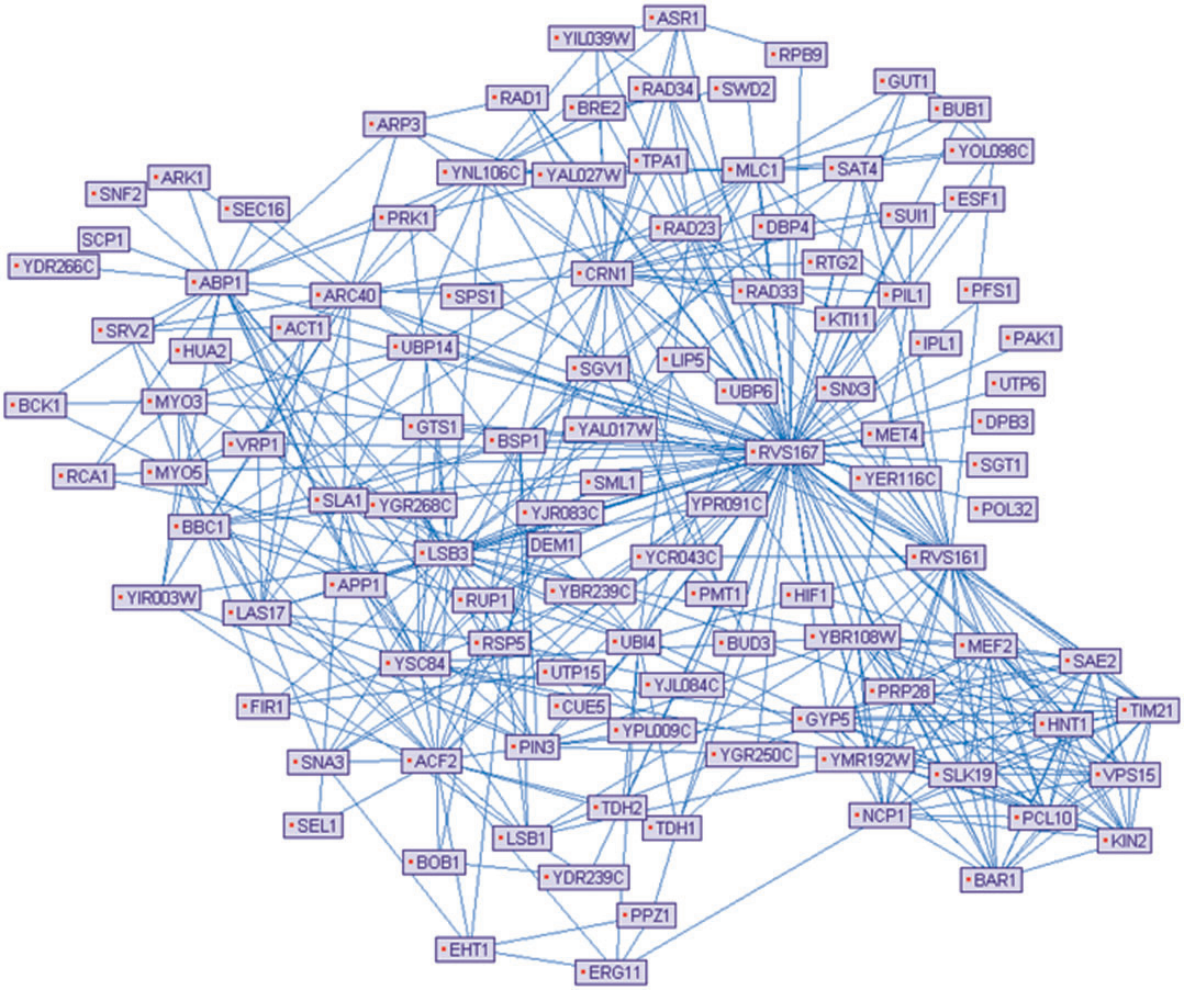

Figure 1: A small portion of the *S. cerevisiae* interactome, drawn by using PIVOT [73]. Nodes are marked by the names of the proteins.

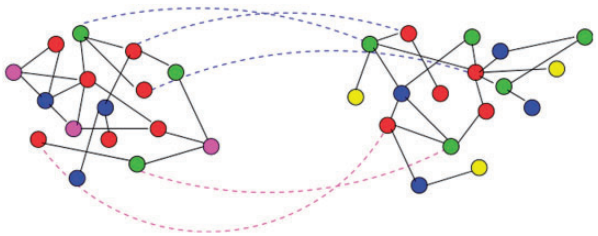

Figure 2: Global Alignment. Solid lines link nodes in the same network, dashed lines represent associated nodes in different networks, nodes with the same colour are enough similar, with respect to the considered similarity threshold.

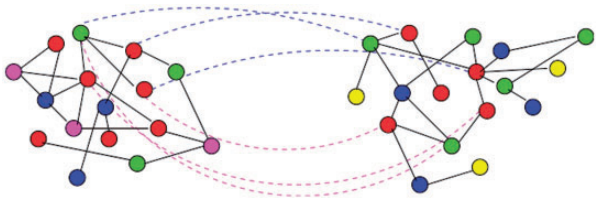

Figure 3: Local alignment. The legend is analogous to the one in Figure 2.

instead at finding multiple unrelated regions of isomorphism among the input networks, with each region implying a mapping independent of the others, where the mapping may involve overlapping subgraphs.

Network alignment can also be performed if the input networks are of different types. Usually, in this case, the input networks are merged and statistical approaches are then applied to extract the most significant subgraphs from the integrated network, often referring to motif extraction (see Section 'Motif extraction').

Most of the network alignment algorithms are also provided in input with a 'dictionary' storing the

structural (e.g. sequence) similarity values among the proteins in the input networks, to take into account both node similarity and network topology in their processing.

Techniques

Network alignment involves the problem of subgraph isomorphism, that is known an NP-complete [75] problem. Therefore, the proposed techniques are often based on approximate and heuristic algorithms.

Global alignment 'IsoRank' [76] is an algorithm for pairwise global alignment of PPI networks that first associates a score to the match between each pair of nodes in the two networks. Then, it builds a one-to-one mapping between the two networks by extracting mutually consistent matches according to a bipartite graph weighted matching performed on the two sets of nodes. 'IsoRank' has been extended in [77] to perform multiple alignment by approximate multipartite graph weighted matching. 'IsoRankN' (IsoRank-Nibble) [78] is a global multiple-network alignment tool based on spectral clustering performed on the induced graph of pairwise alignment scores. In [79], a formulation for pairwise global network alignment is introduced based on maximum structural matching, which combines a Lagrangian relaxation approach with a branch-and-bound method. 'MI-GRAAL' [80] can integrate any number and type of similarity measures between network nodes (e.g. sequence similarity, functional similarity, etc) and find a combination of similarity measures yielding the largest connected alignments. 'GraphCrunch 2' [81] performs network alignment based on the topological similarity of the associated subgraphs, and it allows also for network modelling and clustering. The notion of 'asymmetric alignment' is introduced in [19,82] to deal with the case in which the two networks have a different degree of reliability. The proposed approach relies on finite state automata and the Viterbi algorithm. Shih and Parthasarathy [83] propose a scalable algorithm for multiple alignment based on clustering and graph matching techniques that is able to both detect conserved interactions and maximize the sequence similarity of nodes. In [84], an evolutionary-based global alignment algorithm is proposed, while in [85], a greedy method is used, based on an alignment scoring matrix derived from both biological and topological information. PISwap [86] uses a local optimization heuristic approach to efficiently refine other well-studied alignment techniques. It begins with different types of network alignment approaches and then iteratively adjusts the initial alignments by incorporating network topology information, trading it off for sequence information. 'SMETANA' [87] is based on a semi-Markov random walk model to compute a probabilistic similarity measure between nodes in different networks. The estimated probabilities are enhanced by local and cross-species network similarity information, then used to predict the alignment of multiple networks based on a greedy approach. 'SPINAL' [88] computes pairwise initial similarity scores based on local neighbourhoods matching, and then it iteratively grows a locally improved solution subset. It uses bipartite graphs maximum weight matching for both phases.

Local alignment 'PathBLAST' [89] searches for high scoring pathway alignments involving two paths, one for each of the two input networks, such that proteins of the first path are paired with putative homologs occurring in the same order in the second path. An extension of 'PathBLAST' to multiple alignment is presented in [90], while it has been used in [91] to resolve ambiguous functional orthology relationships in PPI networks. In [92], 'MAWISH' is proposed based on duplication/divergence models and on efficient heuristics to solve a graph optimization problem. 'Bi-GRAPPIN' [93] is based on maximum weight matching of bipartite graphs resulting from comparing the adjacent nodes of pairs of proteins occurring in the input networks. 'Graemlin' [94] aligns an arbitrary number of networks to identify conserved functional modules, greedily assigning the aligned proteins to non-overlapping homology classes and progressively aligning the input networks. It also allows to search for different conserved topologies defined by the user. 'C3Part-M' [95] extracts connected components conserved in several networks based on a formalism that encodes correspondences in multigraphs. It was compared with 'NetworkBlast-M' [96], another technique relying on a representation of multiple networks that is only linear in their size. The approach [97] aligns heterogeneous networks, for example PPI and disease networks. The authors of 'SubMAP' [98] formulate the problem of aligning two metabolic pathways as an eigenvalue problem and solve it using an iterative technique. 'PINALOG' [99] combines information from protein sequence, function

and network topology to perform pairwise alignment. First it finds highly similar protein pairs (i.e. seeds) from highly connected subnetworks in the input networks, and then it extends the alignment to other proteins in the neighbourhoods of such seeds.

'AlignNemo' [100] builds a weighted alignment graph from the input networks, extracts all connected subgraphs of a given size from the alignment graph and uses them as seeds for the alignment solution, by expanding each seed in an iterative fashion.

'GraphAlignment' [101] incorporates information both from network vertices and network edges and it is based on an explicit evolutionary model, allowing inference of all scoring parameters directly from empirical data. In [102], an approach to align metabolic networks by first compressing them is presented. The authors provide a user-defined parameter to control the number of compression levels, which generally determines the trade-off between the quality of the alignment versus the running time.

Querying

Network querying consists of analysing an input network, called 'target network', searching for the occurrences of a 'query network' of interest (Figure 4). The query is usually much smaller than the target. Such a problem 'is aimed at transferring biological knowledge within and across species' [13], since the found subnetworks may correspond to cellular components involved in the same biological processes or performing similar functions to the components in the query.

We note that, sometimes, methods for local alignment have been applied to perform network querying ([19,88,91]), although specific techniques have been proposed to solve this task as summarized below.

Techniques

Network querying approaches may be divided in two main categories: those ones searching for efficient solutions under particular conditions, e.g. the query is not a general graph but it is a path or a tree, and other approaches where the query is a specific small graph in input, often representing a functional module of some well characterized organisms.

Specific topology 'MetaPathwayHunter' [103] queries metabolic networks by multisource trees, which are directed acyclic graphs whose corresponding undirected graphs are trees where nodes may present both incoming and outgoing edges. 'QPath' [104] queries a PPI network by a query pathway consisting of a linear chain of interacting proteins belonging to another organism. The algorithm works in analogy with sequence alignment, by aligning the query

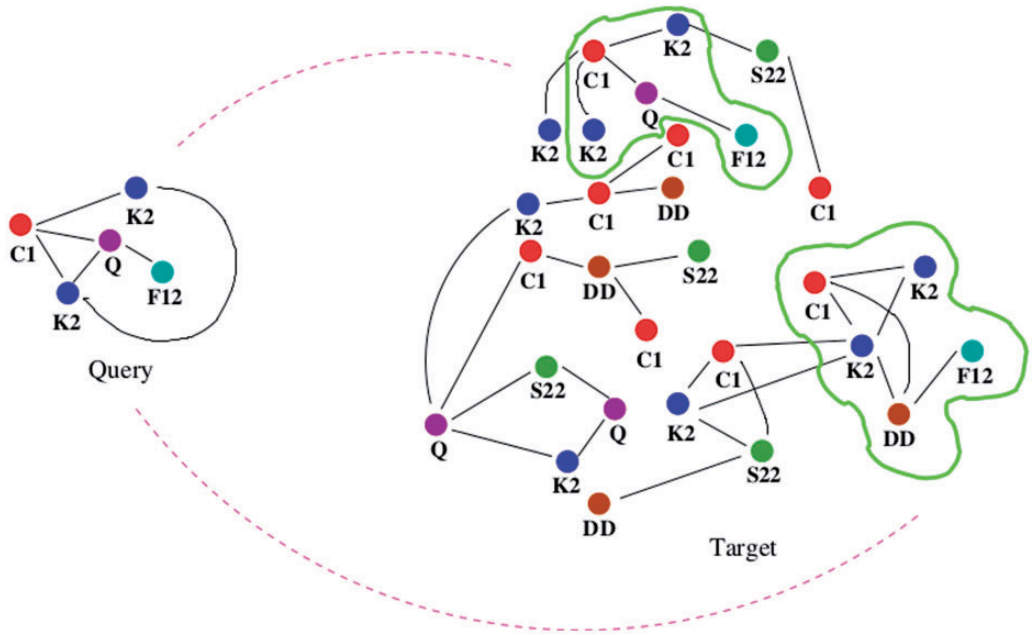

Figure 4: Network querying. Solid lines link nodes in the same network, boundaries highlight occurrences of the query in the target network and dashed lines point out the association between the query and its occurrences.

pathway to putative pathways in the target network, so that proteins in corresponding positions have similar sequences. It has been extended to trees or graphs with limited treewidth in 'QNet' [105].

General topology 'SAGA' [106] is an exact algorithm to search for subgraphs of arbitrary structure in a large graph. It groups related vertices in the target network for each vertex in the query. 'NetMatch' [107] is a Cytoscape plugin allowing for approximate queries, that is, graphs where some nodes are specified and others are wildcards (which can match an unspecified number of elements). 'NetMatch' captures the topological similarity between the query and target graphs, without taking into account any information about node similarities. In [108], a technique is proposed based on maximum weight matching of bipartite graphs. 'Torque' [109] uses both dynamic and integer linear programming to search for a matching set of proteins that are sequence-similar to the query proteins, by relaxing the topology constraints of the query. RESQUE [110] adopts a semi-Markov random walk model to probabilistically estimate the correspondence scores between nodes that belong to different networks, by iteratively reducing the target network based on such scores.

Motif extraction

Given a biological network N, a 'motif' can be defined according to its 'frequency' or to its 'statistical significance' [24]. In the first case, a motif is a subgraph appearing more than a threshold number of times in N; in the second case, it is a subgraph occurring more often than expected by chance. In particular, to measure the statistical significance of a motif, many studies compare the number of motif occurrences with those detected in a number of randomized networks [111], through the use of suitable statistical indices such as P-value and z-score [21] (see also [112] for a method to estimate the number of occurrences of a given motif in an input network).

Techniques

We can distinguish two main categories of approaches: 'topology only' and 'topology and nodes' based. The former ones relate the concept of 'motif' only to the network topology, while the latter ones also consider the biological meaning of nodes.

Topology only Shen-Orr *et al*. [113] defined 'network motifs' as 'patterns of interconnections that recur in many different parts of a network at frequencies much higher than those found in randomized networks'. They discovered three highly significant motifs composed by three/four nodes among which the most famous is the 'feed-forward loop', whose importance has been shown also in further studies [20,114]. The technique presented in [113] laid the foundations for different extensions, such as [115–117]. In [117], composite motifs consisting of two kinds of interactions are extracted by using edges of different colors in the network modelling. In particular, two types (colors) of edges are considered, representing protein–protein and transcription–regulation interactions, and algorithms are developed for detecting network motifs in networks with multiple types of edges. In [115], topological motifs derived from families of mutually similar, but not necessarily identical, patterns are discussed and extracted based on a scoring function. In [116], n-nodes 'bridge' and 'brick' motifs are searched for in complex networks by a method performing simultaneously the detection of global statistical features and local connection structures, and the location of functionally and statistically significant network motifs.

Topology and nodes As observed in [118], there are biological networks (e.g. metabolic networks) where a purely topological definition of motifs seems to be inappropriate, as similar topologies can give rise to different functions. Therefore, the authors of [118] introduce a new definition of motifs in the context of metabolic networks, such that the components of the network play the central role and the topology can be added only as a further constraint. In analogy with Lacroix *et al*. [118], Parida [119] relates the concept of motif to both graph-structure and node similarity. A three-steps exact approach is presented based on the application of the notion of maximality, used extensively in strings and arrays [18,120–127], to graphs. In [128], the two notions of 'structural' and 'biological network motifs' are distinguished, focusing on the latter one referring to biologically significant small connected subgraphs regardless of the structure. Five algorithms for the discovery of biological network motifs are introduced, each reducing the number of subgraphs to search by removing a number of edges from the

original network. At the same time, the discovery rate for biological network motifs is increased.

DISCUSSION

The techniques presented here provide interesting findings, as shown in Table 2. To cite only some examples, dense regions of overlapping interactions have been shown to exist inside the gene interaction network of *Escherichia coli* [113], and they partition it into biologically meaningful combinatorial regulation modules. In [90], Sharan *et al.* have identified 649 proteins that are conserved with high confidence among yeast, worm and fly. Many of the functions and interactions they predicted would not have been identified from sequence similarity alone, demonstrating that network comparisons provide essential biological information beyond what can be obtained from the genome. In addition, the theoretical results proposed in some of the considered studies (e.g. [95,119]) deserve attention. They aim at handling the natural combinatorial explosion caused by the necessity to deal with graph isomorphism, through special formulations of the problem.

We note that not so many exact algorithms have been proposed, and they basically handle situations where the size of the subgraphs to search for can be fixed a priori and restricted to relatively few nodes, for example, in the case of querying and motif extraction.

Table 2 allows us to draw up some conclusions on the validation of the approaches. In particular, network alignment techniques are the most difficult to validate, since there is no gold standard by which to compare the results. The most common way to test the biological quality of alignments is by evaluating the consistency of the found alignments with the Gene Ontology annotations [129]. Some approaches (e.g. [78,99]) also consider a range of metrics, based, for example, on the number of associated protein pairs belonging to the same homologous groups. Local alignment techniques may be applied using the available information on the protein complexes in one species to predict the protein complexes components in another species. Therefore, they may be validated by their agreement with known protein complexes [93,99]. However, it is worth pointing out that the analysis of the obtained alignments is often a research direction in its own right. Querying approaches can instead be easily validated, since the query is usually a known functional module of a given network, so that the results may be compared against the known modules of the target organism. The statistical significance of network motifs is often evaluated by comparison with randomized networks having the same characteristics as the real tested network [113].

As for the application contexts of the considered approaches, we observe that they involve different problems and types of data. Motif extraction has been mainly applied to metabolic and gene regulatory networks, while querying and alignment to PPI networks. We also observe that the presented techniques can be used in cascade for specific applications. As an example, one can first search for the existing motifs in a well-known network, and then query another network by the found motifs. Moreover, multiple network alignment can be reduced to motif finding if the input networks are integrated in an overall graph.

Finally, we note that in Table 2 the URL of the software implementing the presented algorithms is also provided (when publicly available).

CONCLUSIVE REMARKS

We presented a roadmap for those researchers who need to approach the analysis of interaction data, using the search of repetitions across biological networks. This compact overview may also be useful for integrative analysis, concerning both the usage of different biological interaction data (e.g. protein-protein interactions, gene regulatory data, disease annotation data), and the joint application of different types of techniques (e.g. first motif extraction and then network querying, as explained above). On the other hand, the issues addressed here also allow for the identification of some interesting open challenges.

For example, to increase the coverage of available interaction data, the methods used to unmask interactions have been automated to generate high-throughput approaches, resulting in a significant increase of false positives and a consequent reduction in the accuracy of the data [2]. An interesting task would be that of applying the automatic techniques summarized here to clean the available interaction data sets, by comparing reliable portions of networks with less reliable ones. Furthermore, increased interest has recently been generated with regard to integrated networks, which collect information from different approaches, and functional networks, such as the transcriptional profiling networks (TPN).

Table 2: Features of the considered methods

Method	Application domain	Category	Exact
		2002	
Shen-Orr *et al*. [113]	Gene regulatory networks	Motif extraction (topology only)	Yes
Validation: The statistical significance of the network motifs was evaluated by comparison with randomized networks having the same characteristics as the real tested network. The probability that a randomized network had an equal or greater number of each of the motifs than the real network was determined by enumerating the motifs found in 1000 randomized networks.			
Findings: Dense regions of overlapping interactions have been shown to exist inside the gene interaction network of *E. coli*, and they partition it into biologically meaningful combinatorial regulation modules. Three different types of motifs were discovered: the 'feedforward loop', the 'single-input module' and the 'dense overlapping regulons'.			
		2003	
PathBLAST [89]	PPI networks	Pairwise local alignment and querying (specific topology)	No
Input/Output: The query pathway is specified by entering a sequence of two to five proteins. Direct entry of FASTA sequences is useful in some cases. The target network can be specified from a pull-down menu system in the lower left-hand corner of the PathBLAST front page.			
URL: http://www.pathblast.org/			
		2004	
Berg and Lassig [115]	Gene regulatory networks	Motif extraction (topology only)	No
Validation: To quantify the statistical significance of a given number of internal links in a motif, the authors compute the probability distribution of the input network with that of a random graph generated by an unbiased sum over all graphs with the same number of nodes and the same connectivities as in the input data set.			
Findings: The algorithm produced well-defined motifs of maximal likelihood in the gene interaction network of *E. coli*.			
Yeger-Lotern *et al*. [117]	PPI and Gene regulatory networks	Motif extraction (topology only)	Yes
Validation: 1000 randomized networks were created and the statistical significance of the motifs was computed analytically by assuming a uniform distribution of TRIs over transcription factor pairs.			
Findings: A two-protein mixed-feedback loop motif, five types of three-protein motifs exhibiting co-regulation and complex formation and many motifs involving four proteins were found in an integrated data set of PPI and transcription regulation of *S. cerevisiae*.			
		2005	
MetaPathwayHunter [103]	Metabolic networks	Querying (specific topology)	No
Validation: The statistical significance of each alignment was tested by a *P*-value calculation, computed by executing the same query against 100 random pathway graphs, and counting the fraction of graphs containing an alignment that received the same score or higher. The exact binomial test was used to assess whether the number of significantly aligned pathway pairs in both inter-species and intra-species comparisons deviate significantly from the number expected by pure chance at a cut-off of 0.01.			
Findings: All possible alignments between 113 *E. coli* pathways and 151 *S. cerevisiae* pathways were performed, obtaining 610 pathway pairs that had at least one statistically significant alignment between them. The authors found that the conservation between the two species is not limited to small pathways. They also found 187 significant pathway pairs repeated in *E. coli*, and 262 in *S. cerevisiae*.			
URL: http://www.cs.technion.ac.il/olegro/metapathwayhunter/			
NetworkBlast [90]	PPI networks	Multiple local alignment	No
Validation: Conserved paths and clusters identified within the network alignment are compared with those computed from randomized data, and those at a significance level of *P*-value <0.01 are retained.			
Findings: 71 conserved subgraphs were found across *C. elegans*, *Drosophila melanogaster* and *S. cerevisiae*. For 4645 previously undescribed protein functions and 2609 previously undescribed protein interactions, statistically significant support was found. Significantly, many of the predicted functions and interactions would not have been identified from sequence similarity alone, demonstrating that network comparisons provide essential biological information beyond what is gleaned from the genome.			
URL: http://www.cs.tau.ac.il/~bnet/networkblast.htm			

(continued)

Table 2 *Continued*

Method	Application domain	Category	Exact
		2006	
Graemlin [94]	PPI networks	Multiple local alignment	No

Validation: The performances of Græmlin were evaluated by assessing its ability to align known biologically functional modules. The authors also computed the number of 'enriched' alignments, by first assigning to each protein all of its annotations from level eight or deeper in the GO hierarchy; given an alignment, they then discarded unannotated proteins and calculated its enrichment, i.e. the significant shared GO terms or parents of those GO terms, indicating what the aligned sets of proteins may have in common. They considered an alignment to be enriched if the *P*-value of its enrichment was <0.01. As a further validation, they counted the fraction of nodes that have KEGG orthologs but were aligned to any nodes other than their KEGG orthologs.

Findings: Græmlin was applied to perform a 10-way alignment of *E. coli*, *Salmonella typhimurium*, *Vibrio cholerae*, *Caulobacter crescentus*, *Campylobacter jejuni*, *Helicobacter pylori*, *Synechocystis*, *Streptomyces coelicolor*, *Mycobacterium tuberculosis* and *Streptococcus pneumoniae*. This generated $\sim$2000 significant multiple alignments, each containing all or a subset of the 10 species. Experimental evaluations showed it is able to extract more accurate alignments than both NetworkBLAST and MAWISH in some of the analysed cases.

URL: http://graemlin.stanford.edu/

MAWISH [92]	PPI networks	Pairwise local alignment and querying (general topology)	No

Validation: To evaluate the statistical significance of discovered high-scoring alignments, the authors compare them with a reference model generated by a random source. In the reference model, they assume that the interaction networks of the two organisms are independent of each other. To accurately capture the power-law nature of PPI networks, they assume that the interactions are generated randomly from a distribution characterized by a given degree sequence.

Findings: The alignment of *S. cerevisiae* and *D. melanogaster* PPI networks resulted in identification of 412 conserved subnetworks. Eighty-three conserved subnetworks were identified on *S. cerevisiae* and *C. elegans*, and 146 were identified on *C. elegans* and *D. melanogaster*, respectively.

URL: www.cs.purdue.edu/homes/koyuturk/mawish/

MOTUS [118]	Metabolic networks	Motif extraction (topology and nodes)	Yes

Validation: To demonstrate the utility of the proposed definition of network motifs, the authors show an example of application to the comparative analysis of different amino-acid biosynthesis pathways.

Findings: A new definition of network motif based on reaction labels without specifying the topology. This raises original algorithmic issues of which the complexity is discussed.

URL: http://pbil.univ-lyon1.fr/software/motus/

QPath [104]	PPI networks	Querying	No

Validation: The authors used two methods to assess the quality of the found pathways: (i) Functional enrichment, representing the tendency of the pathway's proteins to have coherent Gene Ontology (GO) functions; and (ii) Expression coherency, measuring the similarity in expression profiles of the pathway's coding genes across different experimental conditions.

Findings: Conservations were found in pathways of yeast and fly. Putatively homologous pathways across yeast, human and fly were identified.

		2007	
NetMatch [107]	PPI networks	Querying	No

Input/Output: Users can provide queries to NetMatch by (i) loading them from an existing file, (ii) importing them from the Cytoscape workspace, (iii) drawing them using the NetMatch query drawing tool. A predefined set of frequently used network motifs [107] is also provided. The matching results are shown along with images of matched subnetworks and match information. Clicking on one particular match will highlight its position in the target network in the Cytoscape main view. Any matched subnetwork can be saved and further analysed and manipulated as a separate network in the Cytoscape workspace, using standard Cytoscape features.

URL: http://ferrolab.dmi.unict.it/netmatch.html

(continued)

Table 2 *Continued*

Method	Application domain	Category	Exact
Parida [119]	Metabolic networks	Motif extraction (topology and nodes)	Yes
Findings: The natural combinatorial explosion due to isomorphisms inherent in the problem, which could result in output size being exponential in the input size, is handled by the use of compact location lists.			
QNet [105]	PPI networks	Querying (specific topology)	No
Validation: The biological plausibility of an obtained consensus matches was tested, based on functional enrichment of their member proteins with respect to the fly gene ontology (GO) process annotation. Specifically, given a set of genes in the consensus match that are annotated with a specific term, the probability of obtaining a random set of genes, of the same size as the original pathway and annotated with that term, assuming a hyper-geometric distribution is computed.			
Findings: Known yeast and human signal transduction pathways were searched for in the PPI network of fly, as well as known yeast complexes in fly. Thirty-six of the yeast complexes resulted in a consensus match with more than one protein in fly; 72% of these consensus matches were found to be significantly functionally enriched.			
IsoRank [76]	PPI networks	Pairwise global alignment	No
Validation: Predicted proteins functions were compared with what was known in the literature. Furthermore, the algorithm's error tolerance was evaluated by extracting a 200-node subgraph from one of the input networks, and then randomizing a fraction of its edges.			
Findings: The authors formulate for the first time the problem of global alignment in biological networks. They produced a global alignment between the yeast and fly PPI networks made of 1420 edges, consisting of many disconnected subgraphs, with the largest component presenting 35 edges. The found alignment was used to predict protein functions and to solve functional orthologs ambiguities.			
URL: http://groups.csail.mit.edu/cb/mna/			
SAGA [106]	Metabolic networks	Querying	Yes
Validation: The Monte Carlo simulation approach is used to assess the statistical significance of the matches. A *P*-value is computed for each match based on the frequency of obtaining such a match, or a better match, when applying SAGA with randomized data.			
Findings: Disease-associated human pathway matches were found that are significant but are not yet well studied. As an example, the authors found that T-cell receptor signaling is potentially a significant but relatively unstudied avenue for research into the etiology of *H. pylori* infection.			
URL: http://www.eecs.umich.edu/saga			
	2008		
Cheng [116]	PPI networks	Motif extraction (topology only)	No
Validation: Comparison with randomized networks.			
Findings: Brick motif similarities were found between *E. coli* and *S. cerevisiae*. The authors note that bridge motifs differentiate *C. elegans* from *Drosophila* and *sea urchin* in three types of networks. They suggest that similarities (differences) in bridge and brick motifs imply similar (different) key circuit elements in the three organisms.			
Fionda *et al*. [108]	PPI networks	Querying (general topology)	No
Validation: Comparison with known complexes and pathways.			
Findings: Yeast modules were found that are conserved in fly and *C. elegans*, and human modules were found in yeast PPI network.			
Wu *et al*. [97]	PPI and disease networks	Pairwise local alignment	No
Validation: Gene function enrichment analysis and disease category enrichment analysis.			
Findings: The authors found results confirming that phenotypic overlap is a general indicator of shared pathogenesis. Then, they performed the first heterogeneous alignment of human interactome and phenome networks, to identify pairs of matched subnetworks. The so found results suggest that the causative gene network may serve as a common pathway for the disease family.			

(continued)

Table 2 *Continued*

Method	Application domain	Category	Exact
		2009	
Torque [109]	PPI networks	Querying (general topology)	No
Input/Output: Inputs of Torque are provided in simple text format and consists of (i) a query set of proteins, stored as a comma-delimited or whitespace-delimited list; (ii) their protein sequences, in the standard FASTA format; (iii) a PPI network, where each row represents an interaction and contains the IDs of the interacting pair and a confidence value for it in the range [0, 1]; (iv) the sequences of the network proteins. The web server generates a web page with the image of the top-scoring match for the query in the target network, as well as an auxiliary file that can be viewed using Cytoscape.			
URL: http://www.cs.tau.ac.il/bnet/torque.html			
IsoRankN [78]	PPI networks	Multiple global alignment	No
Validation: Coverage and consistency were considered. Coverage is the set of genes for which the algorithm makes non-trivial predictions. Consistency measures the functional uniformity of genes in each cluster. The authors tested within-cluster consistency of GO/KEGG annotation on the reasoning that predicted orthologs in an orthology should likely have similar function. Then they tested coverage, on the reasoning that an ideal alignment should assign most proteins to a cluster.			
Findings: IsoRankN was compared with IsoRank, Græmlin 2.0 and NetworkBLAST-M on the five available eukaryotic networks of human, mouse, fly, worm and yeast. It outperformed the other methods in terms of number of clusters predicted, within-cluster consistency and GO/KEGG enrichment.			
URL: http://groups.csail.mit.edu/cb/mna/			
NATALIE [79]	PPI and metabolic networks	Pairwise global alignment	No
Findings: The proposed algorithm computes provably optimal network alignments, presenting advantages over pure heuristics approaches.			
URL: http://www.mi.fu-berlin.de/w/LiSA/Natalie			
Bi-GRAPPIN [93]	PPI networks	Pairwise local alignment	No
Validation: Comparison with known complexes.			
Findings: It was able to solve previously unsolved functional orthologs ambiguities.			
C3Part-M [95]	PPI networks	Multiple local alignment	Yes
Validation: Comparison with NetworkBlastN.			
Findings: The authors use the notion of maximality allowing the use of exact algorithms instead of heuristic ones to enumerate the vertices in a connected multigraph obtained from the input networks.			
URL: http://www.inrialpes.fr/helix/people/viari/lxgraph/			
		2011	
AbiNet [19]	PPI networks	Pairwise global alignment and querying (general topology)	No
Validation: Counting the percentage of associated proteins corresponding to the same Gene Ontology annotations.			
Findings: Conservations across different species have been found that were not discovered before, due to the asymmetric nature of the approach.			
URL: http://siloe.deis.unical.it/ABiNet/			
EDGEGO-BNM, EDGEBETWEENNESS-BNM, NMF-BNM, NMFGO-BNM, VOLTAGE-BNM, [120]	PPI networks	Motif extraction (topology only)	No
Validation: Several evaluation measures were used concerning 'motifs included in complexes', 'motifs included in functional modules' and 'GO term clustering score'.			

(continued)

Table 2 *Continued*

Method	Application domain	Category	Exact
MI-GRAAL [80]	PPI networks	Pairwise global alignment	No
Validation: Counting the fraction of aligned protein pairs with common Gene Ontology annotations.			
Findings: Large topological conservations between yeast and human, and between bacteria PPI networks.			
URL: http://bio-nets.doc.ic.ac.uk/MI-GRAAL/			
GraphCrunch 2 [81]	PPI networks	Pairwise global alignment	No
Input/Output: GraphCrunch 2 may receive input networks in the LEDA graph format or as a text file containing the edge list stored as pairs of nodes. The output can be saved in comma-separated or tab-separated formats.			
URL: http://bio-nets.doc.ic.ac.uk/graphcrunch2/			
SubMAP [98]	Metabolic networks	Pairwise local alignment	No
Validation: The similarity of the aligned pathways has been measured by considering the EC numbers of the enzymes catalysing the corresponding reactions.			
Findings: The metabolic pathways of 20 organisms taken from the KEGG database have been compared and new conservations have been found.			
URL: http://bioinformatics.cise.ufl.edu/SubMAP.html			
2012			
AlignNemo [100]	PPI networks	Pairwise local alignment	No
Validation: The found alignments have been validated by evaluating the agreement of the modules found by each method with known complexes. The biological relevance of the discovered mappings was also assessed in terms of functional similarity, by using the set of annotations from the Biological Process (BP) and Molecular Function (MF) vocabularies in the Gene Ontology.			
URL: http://www.bioinformatics.org/alignnemo			
Ay *et al*. [102]	Metabolic networks	Pairwise local alignment	No
Validation: Since the method is based on the compression of the input network, to evaluate the accuracy of the obtained alignments the authors calculated the correlation between the scores of each possible mapping in compressed domain and the scores that they obtained for these mappings without any compression.			
Findings: The proposed compression method reduces the number of reactions by almost half at each level of compression, and the alignment obtained by only one level of compression benefits from a significant performance gain while capturing the original alignment results with high accuracy.			
GraphAlignment [101]	PPI and Gene regulatory networks	Pairwise local alignment	No
Validation: The authors assessed the computational cost and accuracy in three different scenarios. In all them, they constructed pairs of networks that contain 80% of orthologous vertices and 50% of all possible edges. They introduced measures of sensitivity and coverage to determine the quality of the resultant alignments.			
URL: http://www.bioconductor.org			
PINALOG [99]	PPI networks	Pairwise local alignment	No
Validation: The results were analysed in terms of precision and recall, with respect to Gene Ontology annotations.			
Findings: Alignment of human and yeast PPINs revealed several conserved subnetworks between them that participate in similar biological processes, notably the proteasome and transcription related processes.			
URL: http://www.sbg.bio.ic.ac.uk/~pinalog			
RESQUE [110]	PPI networks	Querying (general topology)	No
Validation: To evaluate the accuracy of the querying algorithms, the authors considered the relative number of hits with significant functional coherence, assessed by the Gene Ontology annotations, and the relative number of hits that significantly overlap with a known protein complex.			
URL: http://www.ece.tamu.edu/~bjyoon/RESQUE/			

(continued)

Table 2 *Continued*

Method	Application domain	Category	Exact
Shih & Parthasarathy [83]	PPI networks	Multiple global alignment	No
Validation: The authors used *P*-value and the number of enriched GO terms to evaluate the functional consistency of the generated alignments.			
2013			
Mongiov/ & Sharan [84]	PPI networks	Pairwise global alignment	No
Validation: The authors used existing metrics to assess the obtained results.			
NETAL [85]	PPI networks	Pairwise global alignment	No
Validation: The authors used existing metrics to assess the obtained results.			
URL: http://www.bioinf.cs.ipm.ir/software/netal			
PISwap [86]	PPI networks	Multiple global alignment	No
Validation: The authors used existing metrics to assess the obtained results.			
URL: http://piswap.csail.mit.edu/			
SMETANA [87]	PPI networks	Multiple global alignment	No
Validation: To measure the overall accuracy of the predicted alignments, the authors used measures related to the equivalence class of known functional groups.			
Findings: Conserved subnetwork regions in the three-way alignment of *D. melanogaster*, *Homo sapiens* and *S. cerevisiae* were found.			
URL: http://www.ece.tamu.edu/,bjyoon/SMETANA/			
SPINAL [88]	PPI networks	Pairwise global alignment	No
Validation: The consistency of the aligned pairs of proteins with the Gene Ontology annotations was tested.			
URL: http://code.google.com/p/spinal/			

Note: Columns: (I) Method acronym and reference. (II) Number of networks it can receive in input. (III) Application domain. (IV) Method category. (V) Yes if the method is based on an exact algorithm. Further details (e.g. type of validation, main findings, input/output) are summarized and the URL of the software implementing the method is provided, when they are available.

TPN are not the representation of physical interactions that occur in the cells, but instead they are obtained by experimental evidences that nodes are somehow linked. In TPN, nodes are the genes and they are linked by edges if they have similar expression patterns (i.e. if they are co-expressed) [5,130]. The chemical characteristics of the mRNA molecules make the expression profile data much easier and less expensive to collect than the interaction data [131]. Moreover, while interaction databases offer a snapshot of the whole body of interactions that occur in cells as a static set, which rarely represents an appropriate rendering of the actual physiological situation, transcriptional profiles can be collected for each cell type of a single organism, and even for one tissue of a single patient: for example, several tumour samples have been characterized this way. A large amount of expression data are gathered in specialized databases (e.g. [132–134]), which demands automatic tools to make crude co-expression profiles easily transformable into suitable networks. Advances in this direction would provide new chances to investigate the molecular complexity of diseases and on the differences among individuals and/or cell types, by applying the techniques discussed here to analyse transcriptional profiles.

Finally, most of the alignment and querying approaches compute the similarity between pairs of cellular components (e.g. proteins) only based on sequence information (e.g. protein sequences). Improvements could be achieved by taking into account information on the molecular structures of such components, possibly predicted by the available computational techniques (e.g. [135–137]).

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- The analysis of biological networks is important to understand complex mechanisms of the cell physiology.
- Several approaches exist to search for repetitions in biological networks.
- We provide a compact overview of available resources and approaches.
- We describe the main types of interaction data, models and techniques to find repetitions in biological networks.
- We provide a list of the databases and software tools publicly available.

Acknowledgments

We are grateful to the Reviewers, Raffaele Giancarlo and Luigi Palopoli, whose valuable comments and suggestions allowed us to notably improve the quality of this manuscript. S. E. Rombo was partially supported by Progetto di Ateneo dell'Università degli Studi di Palermo 2012-ATE-0298 'Metodi Formali e Algoritmici per la Bioinformatica su Scala Genomica' and by the Project 'Approcci composizionali per la caratterizzazione e il mining di dati omici' financed by the Italian Ministry of Education, Universities and Research.

References

1. Posada D. *Bioinformatics for DNA Sequence Analysis*. New York: Humana Press, 2009.
2. von Mering D, Krause C, Snel B, *et al*. Comparative assessment of a large-scale data sets of protein-protein interactions. *Nature* 2002;**417**(6887):399–403.
3. Kann MG. Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform* 2007;**8**(5):333–46.
4. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011;**12**(1):56–68.
5. Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. *Cell* 2011;**144**(6):986–98.
6. Pizzuti C, Rombo SE. PINCoC: a Co-Clustering based Method to analyse Protein-Protein Interaction Networks. In: *Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2007), Birmingham, UK*, LNCS Vol. 4881. Springer-Verlag Berlin Heidelberg, 2007, 821–30.
7. Wang J, Li M, Deng Y, *et al*. Recent advances in clustering methods for protein interaction networks. BMC Genomics, 2010;**11**:S10.
8. Becker E, Robisson B, Chapple CE, *et al*. Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics* 2012;**28**(1):84–90.
9. Pizzuti C, Rombo SE. A coclustering approach for mining large protein-protein interaction networks. *IEEE/ACM Trans Comput Biol Bioinform* 2012;**9**(3):717–30.
10. Pizzuti C, Rombo SE. Experimental evaluation of topological-based fitness functions to detect complexes in PPI networks. In: *Genetic and Evolutionary Computation Conference (GECCO'12)*. New York: ACM, 2012;193–200.
11. Pizzuti C, Rombo SE, Marchiori E. Complex detection in protein-protein interaction networks: a compact overview for researchers and practitioners. In: *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Computational Biology (EvoBio'12), Málaga, Spain*, LNCS Vol. 7246. Springer-Verlag Berlin Heidelberg, 2012, 211–23.
12. Pizzuti C, Rombo SE. Restricted neighborhood search clustering revisited: An evolutionary computation perspective. In: *8th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2013), Nice, France*, LNCS Vol. 7986. Springer-Verlag Berlin Heidelberg, 2013, 59–68.
13. Sharan R, Ideker T. Modeling cellular machinery through biological network comparison. *Nat Biotechnol* 2006;**24**(4):427–33.
14. Palopoli L, Rombo SE, Terracina G. Flexible pattern discovery with (extended) disjunctive logic programming. In: *15th International Symposium on Foundations of Intelligent Systems (ISMIS'05)*. Springer-Verlag Berlin Heidelberg, 2005, 504–13.
15. Carvalho AM, Freitas AT, Oliveira AL, *et al*. An efficient algorithm for the identification of structured motifs in DNA promoter sequences. *IEEE/ACM Trans Comput Biol Bioinform* 2006;**3**(2):126–40.
16. Fassetti F, Leone O, Palopoli L, *et al*. Ip6k gene identification in plant genomes by tag searching. *BMC Proc* 2011;**5(Suppl 2)**:S1.
17. Grossi R, Pietracaprina A, Pisanti N, *et al*. MADMX: a strategy for maximal dense motif extraction. *J Comp Biol* 2011;**18**(4):535–45.
18. Rombo SE. Extracting string motif bases for quorum higher than two. *Theor Comput Sci* 2012;**460**:94–103.
19. Ferraro N, Palopoli L, Panni S, *et al*. Asymmetric comparison and querying of biological networks. *IEEE/ACM Trans Comput Biol Bioinform* 2011;**8**:876–89.
20. Mangan S, Itzkovitz S, Zaslaver A, *et al*. The incoherent feed-forward loop accelerates the response-time of the *gal* system of *Escherichia coli*. *J Mol Biol* 2005;**356**(5):1073–81.
21. Milo R, Shen-Orr S, Itzkovitz S, *et al*. Network motifs: simple building blocks of complex networks. *Science* 2002;**298**(5594):824–7.
22. Zhang S, Zhang XS, Chen L. Biomolecular network querying: a promising approach in systems biology. *BMC Syst Biol* 2008;**2**:5.
23. Alon U. Network motifs: theory and experimental approaches. *Nature* 2007;**8**:450–61.
24. Ciriello G, Guerra C. A review on models and algorithms for motif discovery in protein-protein interaction network. *Brief Funct Genomic Proteomic* 2008;**7**(2):147–56.
25. Fionda V, Palopoli L. Biological network querying techniques: analysis and comparison. *J Comput Biol* 2011;**18**(4):595–625.
26. Wong E, Baur B, Quader S, *et al*. Biological network motif detection: principles and practice. *Brief Bioinformatics* 2012;**13**(2):202.
27. Ito T, Chiba T, Ozawa R, *et al*. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* 2001;**98**(8):4569–74.

28. Krogan NJ, Cagney G, Zhong G, *et al*. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 2006;**440**(7084):637–43.
29. Miller JP, Lo RS, Ben-Hur A, *et al*. Large-scale identification of yeast integral membrane protein interactions. *Proc Natl Acad Sci USA* 2005;**102**(34):12123–8.
30. Ceol A, Chatr Aryamontri A, Licata L, *et al*. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res* 2010;**38**:D532–9.
31. Kanehisa M, Goto S, Sato Y, *et al*. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;**40**:D109–14.
32. Gross JL, Jay Y. *Graph Theory and Its Applications*. London: Chapman & Hall/CRC, 2005.
33. Walhout AJ, Boulton SJ, Vidal M. Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast* 2000;**17**(2):88–94.
34. Uetz P, Giot L, Cagney G, *et al*. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000;**403**(6770):623–7.
35. Rigaut G, Shevchenko A, Rutz B, *et al*. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol* 1999;**17**(10):1030–2.
36. Gavin AC, Bosche M, Krause R, *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002;**415**(6868):141–7.
37. Gavin AC, Aloy P, Grandi P, *et al*. Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006;**440**(7084):631–6.
38. Ho Y, Gruhler A, Heilbut A, *et al*. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 2002;**415**(6868):180–3.
39. Landgraf C, Panni S, Montecchi-Palazzi L, *et al*. Protein interaction networks by proteome peptide scanning. *PLoS Biol* 2004;**2**(1):E14.
40. Tong AH, Drees B, Nardelli G, *et al*. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 2002;**295**(5553):321–4.
41. Tonikian R, Xiaofeng X, Christopher P, *et al*. Bayesian modeling of the yeast sh3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. *PLoS Biol* 2009;**7**(10):e1000218.
42. Francke C, Siezen RJ, Teusink B. Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol* 2005;**13**(11):550–8.
43. Stelling J, Klamt S, Bettenbrock K, *et al*. Metabolic network structure determines key aspects of functionality and regulation. *Nature* 2002;**420**:190–3.
44. Reece-Hoyes JS, Marian Walhout AJ. Yeast one-hybrid assays: a historical and technical perspective. *Methods* 2012;**57**(4):441–7.
45. Furey TS. Chip-seq and beyond: new and improved methodologies to detect and characterize protein-dna interactions. *Nat Rev Genet* 2012;**13**(12):840–52.
46. Lee TI, Rinaldi NJ, Robert F, *et al*. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 2002;**298**(5594):799–804.
47. Vermeirssen V, Barrasa MI, Hidalgo CA, *et al*. Transcription factor modularity in a gene-centered *C. elegans* core neuronal protein-DNA interaction network. *Genome Res* 2007;**17**(7):1061–71.
48. Boyle AP, Song L, Lee BK, *et al*. High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells. *Genome Res* 2011;**21**(3):456–64.
49. Hamosh A, Scott AF, Amberger J, *et al*. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2002;**30**(1):52–5.
50. Goh KI, Cusick ME, Valle D, *et al*. The human disease network. *Proc Natl Acad Sci USA* 2007;**104**(21):8685–90.
51. Chen JY, Mamidipalli S, Huan T. HAPPI: an online database of comprehensive human annotated and predicted protein interactions. *BMC Genomcs* 2009;**10(Suppl 1)**:S16.
52. Szklarczyk D, Franceschini A, Kuhn M, *et al*. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 2011;**39**:D561–8.
53. Kikugawa S, Nishikata K, Murakami K. PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from h-invitational protein-protein interactions integrative dataset. *BMC Systems Biol* 2012;**6(Suppl 2)**:S7.
54. Dassi E, Malossini A, Re A, *et al*. AURA: atlas of UTR regulatory activity. *Bioinformatics* 2012;**28**(1):142–4.
55. Schellenberger J, Park JO, Conrad TM, *et al*. BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *Nucleic Acids Res* 2011;**39**:D691–7.
56. Karp PD, Ouzounis CA, Moore-Kochlacs C, *et al*. Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 2005;**33**(19):6083–9.
57. ChatrAryamontri A, Breitkreutz BJ, Heinicke S, *et al*. The BioGRID interaction database: 2013 update. *Nucleic Acids Res* 2013;**41**:D816–23.
58. Salwinski L, Miller CS, Smith AJ, *et al*. The database of interacting proteins: 2004 update. *Nucleic Acids Res* 2004;**32**:D449–51.
59. Encode Project Consortium, Myers RM, Stamatoyannopoulos J, Snyder M, *et al*. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 2011;**9**(4):e1001046.
60. Salomonis N, Hanspers K, Zambon AC, *et al*. GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics* 2007;**8**:217.
61. Chowbina SR, Wu X, Zhang F, *et al*. HPD: an online integrated human pathway database enabling systems biology studies. *BMC Bioinformatics* 2009;**10(Suppl 11)**:S5.
62. Keshava Prasad TS, Goel R, Kandasamy K, *et al*. Human protein reference database–2009 update. *Nucleic Acids Res* 2009;**37**:D767–72.
63. Liberti S, Sacco F, Calderone A, *et al*. HuPho: the human phosphatase portal. *FEBS J* 2013;**280**:379–87.
64. Kerrien S, Aranda B, Breuza L, *et al*. The IntAct molecular interaction database. *Nucleic Acids Res* 2012;**40**:D841–6.
65. Zheng G, Tu K, Yang Q, *et al*. ITFP: an integrated platform of mammalian transcription factors. *Bioinformatics* 2008;**24**(20):2416–7.

66. Calderone A, Castagnoli L, Cesareni G. Mentha: a resource for browsing integrated protein-interaction networks. *Nat Methods* 2013;**10**(8):690–1.
67. Mewes HW, Frishman D, Mayer KF, *et al*. MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res* 2006;**34**:D169–72.
68. Cerami EG, Gross BE, Demir E, *et al*. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res* 2011;**39**:D685–90.
69. Croft D, O'Kelly G, Wu G, *et al*. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 2011;**39**:D691–7.
70. Jiang C, Xuan Z, Zhao F, Zhang MQ. TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res* 2007;**35**:D137–40.
71. Morgat A, Coissac E, Coudert E, *et al*. UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res* 2012;**40**:D761–9.
72. Robasky K, Bulyk ML. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res* 2011;**39**:D124–8.
73. Orlev N, Shamir R, Shiloh Y. PIVOT: protein interacions visualization tool. *Bioinformatics* 2004;**20**(3):424–5.
74. Lee DS, Park J, Kay KA, *et al*. The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci USA* 2008;**105**(29):9880–5.
75. Garey M, Johnson D. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York: Freeman, 1979.
76. Singh R, Xu J, Berger B. Pairwise global alignment of protein interaction networks by matching neighborhood topology. *Proceedings of 11th Annual International Conference RECOMB, Oakland, CA, USA*, LNCS Vol. 4453. Springer-Verlag Berlin Heidelberg, 2007, 16–31.
77. Singh R, Xu J, Berger B. Global alignment of multiple protein interaction networks. *Proc Natl Acad SciUSA* 2008; **105**(35):12763–12768.
78. Liao CS, Lu K, Baym M, *et al*. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics* 2009;**25**:i253–8.
79. Klau GW. A new graph-based method for pairwise global network alignment. *BMC Bioinformatics* 2009;**10(Suppl 1)**: S59.
80. Kuchaiev O, Przulj N. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics* 2011;**27**(10):1390–6.
81. Kuchaiev O, Stevanović A, Hayes W, *et al*. GraphCrunch 2: software tool for network modeling, alignment and clustering. *BMC Bioinformatics* 2011;**12**:24.
82. Ferraro N, Palopoli L, Panni S, *et al*. Master-slave biological network alignment. In: *6th International symposium on Bioinformatics Research and Applications (ISBRA 2010), Connecticut, USA*, LNBI/LNCS Vol. 6053. Springer-Verlag Berlin Heidelberg, 2010, 215–29.
83. Shih YK, Parthasarathy S. Scalable global alignment for multiple biological networks. *BMC Bioinformatics* 2012; **13(Suppl 3)**:S11.
84. Mongiovì M, Sharan R. Global alignment of protein-protein interaction networks. In: Mamitsuka H, DeLisi C, Kanehisa M (eds). In: *Data Mining for Systems Biology, Volume 939 of Methods in Molecular Biology*. New York: Humana Press, 2013, pp. 21–34.
85. Neyshabur B, Khadem1 A, Hashemifar S, *et al*. NETAL: a new graph-based method for global alignment of protein? Protein interaction networks. *Bioinformatics* 2013;**29**(13): 11654–62.
86. Chindelevitch L, Ma CY, Liao CS, *et al*. Optimizing a global alignment of protein interaction networks. *Bioinformatics* 2013;**29**(21):2765–73.
87. Sahraeian SM, Yoon BJ. SMETANA: accurate and scalable algorithm for probabilistic alignment of large-scale biological networks. *PloS One* 2013;**8**:e67995.
88. Aladag AE, Erten C. SPINAL: scalable protein interaction network alignment. *Bioinformatics* 2013;**29**:917–24.
89. Kelley BP, Yuan B, Lewitter F, *et al*. PathBlast: a tool for alignment of protein interaction networks. *Nucleic Acid Res* 2004;**32**:W83–8.
90. Sharan R, Suthram S, Kelley RM, *et al*. From the cover: conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA* 2005;**102**(6):1974–9.
91. Bandyopadhyay S, Sharan R, Ideker T. Systematic identification of functional orthologs based on protein network comparison. *Genome Res* 2006;**16**(3):428–35.
92. Koyuturk M, Kim Y, Topkara U, *et al*. Pairwise alignment of protein interaction networks. *J Comput Biol* 2006;**13**(2): 182–99.
93. Fionda V, Panni S, Palopoli L, *et al*. A technique to search functional similarities in ppi networks. *Int J Data Min Bioin* 2009;**3**(4):431–53.
94. Flannick J, Novak A, Srinivasan BS, *et al*. Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res* 2006;**16**(9):1169–81.
95. Denielou YP, Boyer F, Viari A, *et al*. Multiple alignment of biological networks: a flexible approach. In: *Proceedings of Combinatorial Pattern Matching (CPM 2009)*, LNCS Vol. 5577. Springer-Verlag Berlin Heidelberg, 2009, 263–273.
96. Kalaev M, Bafna V, Sharan R. Fast and accurate alignment of multiple protein networks. In: *Proceedings of 12th Annual International Conference RECOMB, Singapore*, LNCS Vol. 4955. Springer-Verlag Berlin Heidelberg, 2008, 246–256.
97. Wu X, Liu Q, Jiang R. Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics* 2009;**25**(1):98–104.
98. Ay F, Kellis M, Kahveci T. SubMAP: aligning metabolic pathways with subnetwork mappings. *J Comput Biol* 2011; **18**:219–35.
99. Phan HTT, Sternberg MJE. PINALOG: a novel approach to align protein interaction networks-implications for complex detection and function prediction. *Bioinformatics* 2012; **28**:1239–45.
100. Ciriello G, Mina M, Guzzi PH, *et al*. AlignNemo: a local network alignment method to integrate homology and topology. *PloS One* 2012;**7**(6):e38107.
101. Kolář M, Meier J, Mustonen V, *et al*. GraphAlignment: bayesian pairwise alignment of biological networks. *BMC Syst Biol* 2012;**6**:144.
102. Ay F, Dang M, Kahveci T. Metabolic network alignment in large scale by network compression. *BMC Bioinformatics* 2012;**13**:S2.

103. Pinter R, Rokhlenko O, Yeger-Lotem E, *et al*. Alignment of metabolic pathways. *Bioinformatics* 2005;**21**(16):3401–08.
104. Shlomi T, Segal D, Ruppin E, *et al*. QPath: a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics* 2006;**7**:199.
105. Dost B, *et al*. QNet: A tool for querying protein interaction networks. In: *International Conference on Research in Computational Molecular Biology (RECOMB'07), Oakland, CA, USA*, LNCS Vol. 4453. Springer-Verlag Berlin Heidelberg, 2007, 1–15.
106. Yang Q, Sze SH. Saga: a subgraph matching tool for biological graphs. *J Comp Biol* 2007;**14**(1):56–67.
107. Ferro A, Giugno R, Pigola G, *et al*. NetMatch: a cytoscape plugin for searching biological networks. *Bioinformatics* 2007;**23**:910–12.
108. Fionda V, Palopoli L, Panni S, *et al*. Protein-protein interaction network querying by a "focus and zoom" approach. In: *Bioinformatics Research and Development (BRID'08)*, CCIS Vol. 13. Springer-Verlag Berlin Heidelberg, 2008, 331–46.
109. Bruckner S, Hüffner F, Karp RM, *et al*. Torque: topology-free querying of protein interaction networks. *Nucleic Acids Res* 2009;**37**(Web-Server-Issue):106–08.
110. Sahraeian SME, Yoon BJ. RESQUE: network reduction using semi-markov random walk scores for efficient querying of biological networks. *Bioinformatics* 2012;**28**:2129–36.
111. Erdos P, Renyi A. On the evolution of random graphs. *Publ Math Inst Hung Acad Sci* 1960;**5**:17–61.
112. Tran NH, Choi KP, Zhang L. Counting motifs in the human interactome. *Nat Commun* 2013;**4**:2241.
113. Shen-Orr SS, Milo R, Mangan S, *et al*. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature* 2002;**31**:64–8.
114. Mangan S, Alon U. Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci USA* 2003; **100**(21):11980–5.
115. Berg J, Lassig M. Local graph alignment and motif search in biological networks. *Proc Natl Acad Sci USA* 2004;**101**(41): 14689–94.
116. Cheng CY, Huang CY, Sun CT. Mining bridge and brick motifs from complex biological networks for functionally and statistically significant discovery. *IEEE Trans Syst Man Cybern B Cybern* 2008;**38**(1):17–24.
117. Yeger-Lotem E, Sattath S, Kashtan N, *et al*. Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. *Proc Natl Acad Sci USA* 2004;**101**(16):5934–9.
118. Lacroix V, Fernandes CG, Sagot MF. Motif search in graphs: application to metabolic networks. *IEEE/ACM Trans Comput Biol Bioinformatics* 2006;**3**(4):360–8.
119. Parida L. Discovering topological motifs using a compact notation. *J Comp Biol* 2007;**14**(3):46–69.
120. Apostolico A, Parida L. Incremental paradigms of motif discovery. *J Comp Biol* 2004;**11**(1):15–25.
121. Grossi R, Pisanti N, Crochemore M, *et al*. Bases of motifs for generating repeated patterns with wild cards. *IEEE/ACM Trans Comp Biol Bioinf* 2000;**2**(3):159–77.
122. Apostolico A, Parida L, Rombo SE. Motif patterns in 2D. *Theor Comput Sci* 2008;**390**(1):40–55.
123. Rombo SE. Optimal extraction of motif patterns in 2D. *Inf Process Lett* 2009;**109**(17):1015–20.
124. Amelio A, Apostolico A, Rombo SE. Image compression by 2D motif basis. In: *Data Compression Conference (DCC'11), Snowbird, UT, USA*. Washington, DC: IEEE CS Press, 2011, 153–62.
125. Parida L, Pizzi C, Rombo SE. Characterization and extraction of irredundant tandem motifs. In: *String Processing and Information Retrieval (SPIRE'12), Cartagena de Indias, Colombia*, LNCS Vol. 7608. Springer-Verlag Berlin Heidelberg, 2012, 385–97.
126. Parida L, Pizzi C, Rombo SE. Irredundant tandem motifs. *Theor Comput Sci* 2013. http://dx.doi.org/10.1016/j.tcs.2013.08.012.
127. Groccia MC, Furfaro A, Rombo SE. Image classification based on 2D feature motifs. In: *Flexible Query Answering Systems (FQAS 2013), Granada, Spain*, LNCS Vol. 8132. Springer-Verlag Berlin Heidelberg, 2013, 340–51.
128. Kim W, Li M, Wang J, *et al*. Biological network motif detection and evaluation. *BMC Syst Biol* 2011; **5(Suppl 3)**:S5.
129. Asburner S, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet* 2000;**25**:25–9.
130. Stuart JM, Segal E, Koller D, *et al*. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003;**302**(5643):249–55.
131. Martin JA, Wang Z. Next-generation transcriptome assembly. *Nat Rev Genet* 2011;**12**(10):671–82.
132. Krupp M, Marquardt J, Sahin U, *et al*. RNA-Seq Atlas—a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics* 2012; **28**:1184–5.
133. Kapushesky M, Adamusiak T, Burdett T, *et al*. Gene Expression Atlas update—a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acid Res* 2012;**40**:D1077–81.
134. Barrett T, Troup DB, Wilhite S, *et al*. NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acid Res* 2011;**39**:D1005–10.
135. Palopoli L, Rombo SE, Terracina G, *et al*. Improving protein secondary structure predictions by prediction fusion. *Inf Fusion* 2009;**10**(3): 217–32.
136. Raman S, Vernon R, Thompson J, *et al*. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 2009;**77(Suppl 9)**:89–99.
137. Mitra P, Shultis D, Zhang Y. EvoDesign: de novo protein design based on structural and evolutionary profiles. *Nucleic Acids Res* 2013;**41**:W273–80.