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.
Searching for repetitions in biological networks

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
**DNA element towards the transcription factor (outgoing).**

A gene regulatory network may also be represented as a ‘connectivity matrix’ $M$, such that $M_{ij} = 1$ if the component associated with the node $j$ encodes a transcription factor regulating the component associated with the node $i$, and $M_{ij} = 0$ 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...
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.

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. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

Figure 3: Local alignment. The legend is analogous to the one in Figure 2. A colour version of this figure is available at BIB online: http://bib.oxfordjournals.org.

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 network to the target network.

**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**

<table>
<thead>
<tr>
<th>Method</th>
<th>Application domain</th>
<th>Category</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2002</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shen-Orr et al. [113]</td>
<td>Gene regulatory networks</td>
<td>Motif extraction (topology only)</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Validation:</strong></td>
<td>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.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Findings:</strong></td>
<td>Dense regions of overlapping interactions have been shown to exist inside the gene interaction network of <em>E. coli</em>, 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'.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2003</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PathBLAST [89]</td>
<td>PPI networks</td>
<td>Pairwise local alignment and querying (specific topology)</td>
<td>No</td>
</tr>
<tr>
<td><strong>Input/Output:</strong></td>
<td>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.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>URL:</strong></td>
<td><a href="http://www.pathblast.org/">http://www.pathblast.org/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2004</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berg and Lassig [115]</td>
<td>Gene regulatory networks</td>
<td>Motif extraction (topology only)</td>
<td>No</td>
</tr>
<tr>
<td><strong>Validation:</strong></td>
<td>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.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Findings:</strong></td>
<td>The algorithm produced well-defined motifs of maximal likelihood in the gene interaction network of <em>E. coli</em>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2005</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MetaPathwayHunter [103]</td>
<td>Metabolic networks</td>
<td>Querying (specific topology)</td>
<td>No</td>
</tr>
<tr>
<td><strong>Validation:</strong></td>
<td>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.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Findings:</strong></td>
<td>All possible alignments between 113 <em>E. coli</em> pathways and 151 <em>S. cerevisiae</em> 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 <em>E. coli</em>, and 262 in <em>S. cerevisiae</em>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>URL:</strong></td>
<td><a href="http://www.cs.technion.ac.il/olegro/metapathwayhunter/">http://www.cs.technion.ac.il/olegro/metapathwayhunter/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NetworkBlast [90]</td>
<td>PPI networks</td>
<td>Multiple local alignment</td>
<td>No</td>
</tr>
<tr>
<td><strong>Validation:</strong></td>
<td>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 &lt;0.01 are retained.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Findings:</strong></td>
<td>71 conserved subgraphs were found across <em>C. elegans</em>, <em>Drosophila melanogaster</em> and <em>S. cerevisiae</em>. For 46-45 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.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>URL:</strong></td>
<td><a href="http://www.cs.tau.ac.il/~bnet/networkblast.htm">http://www.cs.tau.ac.il/~bnet/networkblast.htm</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2 Continued

<table>
<thead>
<tr>
<th>Method</th>
<th>Application domain</th>
<th>Category</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2006</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graemlin [94]</td>
<td>PPI networks</td>
<td>Multiple local alignment</td>
<td>No</td>
</tr>
</tbody>
</table>

**Validation:** The performances of Graemlin 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:** Graemlin was applied to perform a 10-way alignment of *E. coli*, *Salmonella typhimurium*, *Vibrio cholerae*, *Caulobacter crescentus*, *Campylobacter jejuni*, *Helicobacter pylori*, *Synechocystis* interactions, *Streptomyces coelicolor*, *Mycobacterium tuberculosis* and *Streptococcus pneumoniae*. This generated ~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/](http://graemlin.stanford.edu/)

<table>
<thead>
<tr>
<th>MAWISH [92]</th>
<th>PPI networks</th>
<th>Pairwise local alignment and querying (general topology)</th>
<th>No</th>
</tr>
</thead>
</table>

**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/](http://www.cs.purdue.edu/homes/koyuturk/mawish/)

<table>
<thead>
<tr>
<th>MOTUS [118]</th>
<th>Metabolic networks</th>
<th>Motif extraction (topology and nodes)</th>
<th>Yes</th>
</tr>
</thead>
</table>

**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/](http://pbil.univ-lyon1.fr/software/motus/)

<table>
<thead>
<tr>
<th>QPath [104]</th>
<th>PPI networks</th>
<th>Querying</th>
<th>No</th>
</tr>
</thead>
</table>

**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.

**URL:** [http://ferrolab.dmi.unict.it/netmatch.html](http://ferrolab.dmi.unict.it/netmatch.html)

(continued)
<table>
<thead>
<tr>
<th>Method</th>
<th>Application domain</th>
<th>Category</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parida [119]</td>
<td>Metabolic networks</td>
<td>Motif extraction (topology and nodes)</td>
<td>Yes</td>
</tr>
<tr>
<td>QNet [105]</td>
<td>PPI networks</td>
<td>Querying (specific topology)</td>
<td>No</td>
</tr>
<tr>
<td>IsoRank [76]</td>
<td>PPI networks</td>
<td>Pairwise global alignment</td>
<td>No</td>
</tr>
<tr>
<td>SAGA [106]</td>
<td>Metabolic networks</td>
<td>Querying</td>
<td>Yes</td>
</tr>
<tr>
<td>Fionda et al. [108]</td>
<td>PPI networks</td>
<td>Querying (general topology)</td>
<td>No</td>
</tr>
<tr>
<td>Wu et al. [97]</td>
<td>PPI and disease networks</td>
<td>Pairwise local alignment</td>
<td>No</td>
</tr>
</tbody>
</table>

**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.
- 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.
- 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.
- 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://groups.csail.mit.edu/cb/mna/
- http://www.eecs.umich.edu/saga

---

(continued)
### Table 2 Continued

<table>
<thead>
<tr>
<th>Method</th>
<th>Application domain</th>
<th>Category</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Torque [109]</td>
<td>PPI networks</td>
<td>Querying (general topology)</td>
<td>No</td>
</tr>
</tbody>
</table>

**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](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, Gramlin 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/](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](http://www.mi.fu-berlin.de/w/LiSA/Natalie)

| Bi-GRAPPI~N [93]| PPI networks                            | Pairwise local alignment          | No          |

**Validation:** Comparison with known complexes.

**Findings:** It was able to solve previously unsolved functional orthologs ambiguities.

| C3Parts-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/](http://www.inrialpes.fr/helix/people/viari/lxgraph/)

| AbiNet [19]    | PPI networks                            | Pairwise global alignment         | 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/](http://siloe.deis.unical.it/ABiNet/)

| EDGEGO-BNM, EDGEBEFORENESS-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>
<thead>
<tr>
<th>Method</th>
<th>Application domain</th>
<th>Category</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI-GRAAL [80]</td>
<td>PPI networks</td>
<td>Pairwise global alignment</td>
<td>No</td>
</tr>
<tr>
<td><strong>Validation:</strong></td>
<td></td>
<td>Counting the fraction of aligned protein pairs with common Gene Ontology annotations.</td>
<td></td>
</tr>
<tr>
<td><strong>Findings:</strong></td>
<td></td>
<td>Large topological conservations between yeast and human, and between bacteria PPI networks.</td>
<td></td>
</tr>
<tr>
<td><strong>URL:</strong></td>
<td><a href="http://bio-nets.doc.ic.ac.uk/MI-GRAAL/">http://bio-nets.doc.ic.ac.uk/MI-GRAAL/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GraphCrunch 2 [81]</td>
<td>PPI networks</td>
<td>Pairwise global alignment</td>
<td>No</td>
</tr>
<tr>
<td><strong>Input/Output:</strong></td>
<td></td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td><strong>URL:</strong></td>
<td><a href="http://bio-nets.doc.ic.ac.uk/graphcrunch2/">http://bio-nets.doc.ic.ac.uk/graphcrunch2/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SubMAP [98]</td>
<td>Metabolic networks</td>
<td>Pairwise local alignment</td>
<td>No</td>
</tr>
<tr>
<td><strong>Validation:</strong></td>
<td></td>
<td>The similarity of the aligned pathways has been measured by considering the EC numbers of the enzymes catalysing the corresponding reactions.</td>
<td></td>
</tr>
<tr>
<td><strong>Findings:</strong></td>
<td></td>
<td>The metabolic pathways of 20 organisms taken from the KEGG database have been compared and new conservations have been found.</td>
<td></td>
</tr>
<tr>
<td><strong>URL:</strong></td>
<td><a href="http://bioinformatics.cise.ufl.edu/SubMAP.html">http://bioinformatics.cise.ufl.edu/SubMAP.html</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AlignNemo [100]</td>
<td>PPI networks</td>
<td>Pairwise local alignment</td>
<td>No</td>
</tr>
<tr>
<td><strong>Validation:</strong></td>
<td></td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td><strong>Findings:</strong></td>
<td></td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td><strong>URL:</strong></td>
<td><a href="http://www.bioinformatics.org/alignnemo">http://www.bioinformatics.org/alignnemo</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ay et al. [102]</td>
<td>Metabolic networks</td>
<td>Pairwise local alignment</td>
<td>No</td>
</tr>
<tr>
<td><strong>Validation:</strong></td>
<td></td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td><strong>Findings:</strong></td>
<td></td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td><strong>URL:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GraphAlignment [101]</td>
<td>PPI and Gene regulatory networks</td>
<td>Pairwise local alignment</td>
<td>No</td>
</tr>
<tr>
<td><strong>Validation:</strong></td>
<td></td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td><strong>URL:</strong></td>
<td><a href="http://www.bioconductor.org">http://www.bioconductor.org</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PINALOG [99]</td>
<td>PPI networks</td>
<td>Pairwise local alignment</td>
<td>No</td>
</tr>
<tr>
<td><strong>Validation:</strong></td>
<td></td>
<td>The results were analysed in terms of precision and recall, with respect to Gene Ontology annotations.</td>
<td></td>
</tr>
<tr>
<td><strong>Findings:</strong></td>
<td></td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td><strong>URL:</strong></td>
<td><a href="http://www.sbg.bio.ic.ac.uk/~pinalog">http://www.sbg.bio.ic.ac.uk/~pinalog</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RESQUE [110]</td>
<td>PPI networks</td>
<td>Querying (general topology)</td>
<td>No</td>
</tr>
<tr>
<td><strong>Validation:</strong></td>
<td></td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td><strong>URL:</strong></td>
<td><a href="http://www.ece.tamu.edu/~bjyoon/RESQUE/">http://www.ece.tamu.edu/~bjyoon/RESQUE/</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued)
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


