Integrative Biology

Cite this: *Integr. Biol.*, 2012, **4**, 734–743

www.rsc.org/ibiology

PAPER

C-GRAAL: *C*ommon-neighbors-based global *GRA*ph *AL*ignment of biological networks†‡

Vesna Memišević[a] and Nataša Pržulj*[b]

Received 16th October 2011, Accepted 11th December 2011
DOI: 10.1039/c2ib00140c

Networks are an invaluable framework for modeling biological systems. Analyzing protein–protein interaction (PPI) networks can provide insight into underlying cellular processes. It is expected that comparison and alignment of biological networks will have a similar impact on our understanding of evolution, biological function, and disease as did sequence comparison and alignment. Here, we introduce a novel pairwise global alignment algorithm called Common-neighbors based GRAph ALigner (C-GRAAL) that uses heuristics for maximizing the number of aligned edges between two networks and is based solely on network topology. As such, it can be applied to any type of network, such as social, transportation, or electrical networks. We apply C-GRAAL to align PPI networks of eukaryotic and prokaryotic species, as well as inter-species PPI networks, and we demonstrate that the resulting alignments expose large connected and functionally topologically aligned regions. We use the resulting alignments to transfer biological knowledge across species, successfully validating many of the predictions. Moreover, we show that C-GRAAL can be used to align human–pathogen inter-species PPI networks and that it can identify patterns of pathogen interactions with host proteins solely from network topology.

1 Introduction

1.1 Background

Networks are used to model real-world phenomena in various research domains, including systems biology. Technological advances in high-throughput biological experimental methods for interaction detection have led to an explosion in the amount of biological network data of various types, such as protein–protein interaction (PPI) networks, metabolic networks, and transcriptional-regulation networks. We primarily focus on analyzing protein–protein interaction (PPI) networks, in which proteins are network nodes and physical interactions between them are network edges. Since proteins rarely act in isolation, but rather associate with each other to perform many biological functions, analyzing the structure of PPI networks can provide insights into the functioning of cells.

Meaningful biological network comparison is one of the foremost challenges in systems biology. Network alignment is one of the most common biological network comparison methods. In the context of PPI networks, the main goal of network alignment is to identify conserved protein subnetworks across species that are believed to represent evolutionarily conserved functional modules.[1]

These conserved subnetworks could provide transfer of knowledge between species, as well as insights into evolution,

[a] *Department of Computer Science, University of California, Irvine, Irvine, CA 92697-3435, USA*
[b] *Department of Computing, Imperial College London, London, SW7 2AZ, UK. E-mail: natasha@imperial.ac.uk*

† Published as part of an iBiology themed issue entitled "Computational Integrative Biology", Guest Editor: Prof. Jan Baumbach.

‡ Electronic supplementary information (ESI) available: Supplementary files and C–GRAAL executables. See DOI: 10.1039/c2ib00140c

Insight, innovation, integration

To address a challenge of identifying conserved protein subnetworks across species, we introduce a novel pairwise global alignment algorithm called Common-neighbors based GRAph ALigner (C-GRAAL) that uses heuristics for maximizing the number of aligned edges between two networks and is based solely on network topology. It allows the use of a node similarity measure incorporating any type or combination of biological data. However, it can also be used to align networks for which information about node similarities is not available. We show that C-GRAAL successfully aligns PPI networks of different species, exposing large connected and functionally similar topologically aligned regions and the resulting alignments can be used to successfully transfer biological knowledge across species.

protein function, and protein–protein interactions. It is expected that the network alignment will have a similar impact on our understanding of biology and evolution as did sequence alignment.

Similar to sequence alignments, there exist *local* and *global* network alignments. *Local* network alignment algorithms focus on matching small subnetworks from one network to subnetworks in another network. Many of these, such as *PathBLAST*,[2] *NetworkBLAST*,[3] *NetAlign*,[4] *MaWISh*,[5] and *Graemlin*,[6] combine network topology information with other types of biological information, *e.g.*, protein sequence similarities or phylogenetic relationships between species whose networks are being aligned, to identify small network structures that correspond to conserved pathways or protein complexes. Since these algorithms allow one node to have different pairings in different local alignments, the resulting alignments can be ambiguous.

A *global* network alignment provides a unique, one-to-one mapping for every node in one network to exactly one node in the other network, even though this can lead to suboptimal matchings in some local regions. There are several global network alignment algorithms, including *IsoRank*,[7–9] *Graemlin*,[10] *GRAAL*,[11] *H-GRAAL*,[12] and MI-GRAAL.[13] The earliest global network alignment algorithm, *IsoRank*, uses a greedy strategy to create an alignment between two networks based both on sequence similarity between nodes (*i.e.*, proteins) and on topological similarity of their neighborhoods.[7] Later, this algorithm was extended to perform multiple local and global network alignments.[8,9] *Graemlin*, originally a local network alignment algorithm, has been extended to allow global network alignment based on a set of *a priori* known protein sequence alignments and phylogenetic relationships between these proteins.[10] More recently, three global network alignment algorithms, *GRAph ALigner* (*GRAAL*),[11] *Hungarian-algorithm-based GRAAL* (*H-GRAAL*),[12] and *Matching-based Integrative GRAAL* (*MI-GRAAL*),[13] have been published. Each of these uses a cost function relying on a highly constraining measure of topological similarity between the extended network neighborhoods of nodes.[14] GRAAL is a greedy "seed and extend" approach that uses a heuristic search strategy to quickly find approximate alignments. H-GRAAL is based on the Hungarian algorithm for minimum-weight bipartite matching, and it produces an optimal alignment having the minimum total alignment cost with respect to the given cost function. MI-GRAAL algorithm combines a "seed and extend" approach with the *maximum weight bipartite matching problem* to find a pairwise alignment between two networks. Additionally, MI-GRAAL can integrate and use multiple types of similarity measures between network nodes.

1.2 Our contribution

We introduce a novel pairwise network alignment algorithm, *Common-neighbors based GRAph ALigner* (C-GRAAL). While all currently available alignment algorithms depend on one or more *a priori* defined features describing proteins, such as sequence similarities, phylogenetic relationship, or topological similarity between nodes, C-GRAAL does not require any measure of node similarity. Instead, it builds an alignment using a heuristic approach on the underlying network topology alone. As such, it can be used to align networks for which information about the node similarities is not available. However, to achieve deeper understanding of complex biological processes, we should try to use all biological data available. Our algorithm allows the use of a node similarity measure incorporating any type or combination of biological data.

We evaluate the performance of C-GRAAL and apply it to multiple tasks. First, we align high-confidence yeast and human PPI networks[15,16] and demonstrate that C-GRAAL performs as well as or better than the best currently available alignment algorithm. Our results show that C-GRAAL exposes large connected and functionally consistent aligned regions, implying that these distant species share a substantial amount of network topology.

In addition, we used C-GRAAL's alignment to predict biological characteristics, *i.e.*, the functions, processes, and localizations of proteins in prokaryotic and eukaryotic species. We validate a large number of our predictions in the literature and demonstrate that C-GRAAL can be used to transfer biological knowledge between species.

Moreover, networks of interactions between human proteins and bacterial pathogens are poorly understood. We show that C-GRAAL can be used to detect large conserved regions between aligned human–pathogen PPI networks.[17] Since the aligned regions are enriched with proteins with the same biological characteristics, they can be used to identify patterns of pathogen interactions with the host cell.

2 Results and discussion

C-GRAAL is a global network alignment algorithm that aligns each node in a smaller network to exactly one node in a larger network. Because of the uncertainty in the C-GRAAL algorithm, we align each pair of networks 30 times and report the average and a standard deviation of the alignment scores over all runs, as well as the best score.

2.1 Evaluation of C-GRAAL algorithm

To evaluate the performance of C-GRAAL, we apply it to the *Saccharomyces cerevisiae* PPI network (henceforth denoted as "yeast")[15] and the *Homo sapiens* PPI network (henceforth denoted as "human").[16] The yeast PPI network contains 16 127 interactions between 2390 proteins. The human PPI network contains 41 456 interactions between 9141 proteins. We compare the topological and biological quality of C-GRAAL's alignments with those obtained by the best currently published network alignment algorithm, MI-GRAAL.[13] Furthermore, we evaluate the performance of C-GRAAL on the following four bacterial PPI networks: *M. loti*, *S. sp PCC6803*,[18,19] *C. jejuni*,[20] and *E. coli*.[21] The results obtained show that C-GRAAL can be used to find large topologically similar regions in which a statistically significant number of aligned bacterial proteins perform the same biological function.

Note that we do not compare C-GRAAL to Graemlin[10] because Graemlin requires as input information about phylogenetic relationships between species whose networks are being aligned. Furthermore, we do not compare C-GRAAL to IsoRankN,[9] because IsoRankN's alignment results in many-to-many node mappings that cannot be quantified topologically with the

edge correctness (EC),[7,11–13] where EC is the percentage of edges from a smaller network that are correctly aligned to edges in a bigger network. Since many-to-many node alignments can produce exponentially many one-to-one node alignments, enumerating all of them is computationally infeasible, and thus a proper comparison with IsoRankN is infeasible.

2.1.1 Topological quality of yeast–human alignment. We define the best alignment to be the alignment with the highest edge correctness (EC) score over all node similarity measures and over all runs of the algorithm. We evaluate C-GRAAL's performances on two node similarity measures: graphlet degree vector (GDV) similarity[14] that measures topological similarity of nodes in a network, and protein sequence similarity (BLAST E-values)[22,23] (see ESI‡ for details). For each of the two similarity measures, we calculate 30 different alignments.

The best yeast–human alignment is obtained when we use only the GDV as a node similarity measure, with 3636 correctly aligned edges (EC = 22.55%). The average EC of the 30 runs is 21.07% ± 0.55%. This alignment is statistically significant, with a p-value of 7×10^{-8}, meaning that it is highly unlikely to obtain the observed EC score or better in a random alignment of these two networks (see ESI‡ for details). There are 1916 proteins involved in these correct edge alignments, representing 80.17% of all yeast proteins. The best alignment produced by MI-GRAAL using only GDV node similarity is slightly better (EC = 23.26%) than C-GRAAL. However, the average EC score for MI-GRAAL is lower and the standard deviation is higher (19.73 ± 1.39%). This implies that C-GRAAL is more consistent in producing alignments of high EC than MI-GRAAL.

Conversely, when we use only protein sequence similarity information as a node similarity measure, C-GRAAL outperforms MI-GRAAL. The best C-GRAAL alignment obtained with this node similarity measure consists of 1844 nodes involved in 3344 correctly aligned edges. The average EC of the 30 runs is 18.82% ± 0.76%. All alignment results are statistically significant, with a p-value of 10^{-8}. The best alignment produced by MI-GRAAL with only sequence similarity as node similarity has an EC of 13.73%, while the average EC is 13.30% ± 0.23%. For more details about the alignments, see Table 1.

To assess the size and connectedness of the aligned regions of our yeast–human alignment, we measure the size of the largest common connected subgraph (CCS). The largest CCS in our overall best alignment consists of 1799 nodes and 3570 edges, which is approximately 75% of nodes and 22% of edges from the yeast network (Fig. 1).

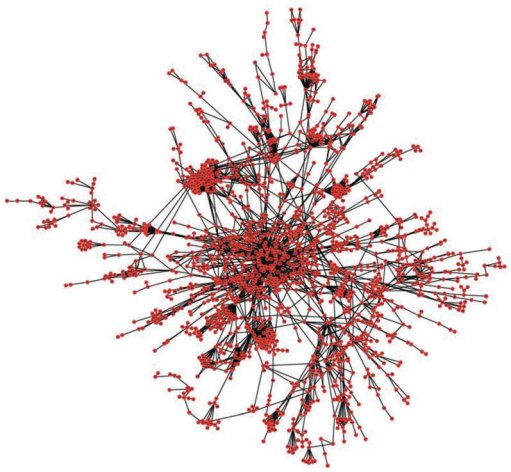


Fig. 1 The largest *common connected subgraph* of the best yeast–human alignment, consisting of 1799 nodes and 3570 edges (approximately 75% of nodes and 22% of edges from the yeast PPI network). An edge between two nodes means that an interaction exists in both species between the corresponding protein pairs.

Next, we compare the best alignments obtained for both node similarity measures (GDV and sequence similarity) and find that the overlap between these two alignments is low: only two aligned yeast–human protein pairs are present in both alignments. The small *exact* overlap between the alignments might not be surprising, given that there might exist a large number of different topological alignments, all of which may have comparable quality. For example, it is possible that topologically identical subgraphs in two networks are aligned in each run, but the actual aligned protein pairs differ for different runs; this could happen, for example, when two cliques (complete graphs having all possible edges between the nodes, *e.g.*, triangle) of the same size are aligned, since all pairings of nodes between such cliques are topologically equivalent.

Since the yeast PPI network that we analyzed contains interactions obtained only by pull-down experiments, we also evaluate performance of C-GRAAL on networks obtained by binary, yeast-two-hybrid (Y2H), experiments. We align the same human PPI network with the yeast high-confidence PPI network by Yu *et al.*,[24] which is regarded as a gold standard binary interaction network (henceforth denoted by "B.yeast," for "Binary yeast"). The B.yeast network consists of 1263 interactions

Table 1 C-GRAAL's alignments of yeast and human. The statistics for aligning yeast and human PPI networks obtained using different node similarity measures averaged over 30 runs of the algorithm. Columns denoted by "EC (max)," "EC (avg)," and "EC (std)" represent the maximum edge correctness (EC), the average EC, and the standard deviation of EC over 30 runs, respectively. The column denoted by "p-value" represents the statistical significance of observed alignments. Columns denoted by "LCCS (nodes)" and "LCCS (edges)" represent the size of the largest common connected subgraph in terms of the number of nodes and edges, respectively, of the alignment with the maximum EC produced by C-GRAAL. Columns YH–overlap, H–nodes (overlap), and Y–nodes (overlap) represent the number of yeast–human pairs present in all 30 alignments, percentage of human proteins present in all 30 alignments, and percentage of yeast proteins present in all 30 alignments, respectively. Note that since yeast is the smaller network, all yeast proteins are present in the alignment in all 30 runs

Similarity measure	EC (max) (%)	EC (avg) (%)	EC (std) (%)	p-Value	LCCS (nodes)	LCCS (edges)	YH–overlap (%)	H-nodes (overlap) (%)	Y-nodes (overlap) (%)
GDV similarity	22.55	21.07	0.55	7×10^{-8}	1799	3570	255 (10.7%)	41.8	100
Sequence similarity	20.74	18.82	0.76	7×10^{-8}	1814	3303	290 (12.2%)	37.5	100

between 1078 proteins, so it is very sparse and of low protein coverage. The best B.yeast–human alignment is produced by C-GRAAL when only the GDV is used as a node similarity measure and this alignment contains 370 correctly aligned edges (EC = 29.29%). The average EC of 30 runs of C-GRAAL on these data is 28.49% ± 0.86% and all alignments are statistically significant. This shows that C-GRAAL performs equally well on both Y2H and pull-down data. Although B.yeast–human alignment has higher edge correctness than the yeast–human alignment, it consists of fewer than 400 edges that are common to human and yeast, so biological insight obtained from it is limited. This is because the B.yeast network is very sparse and of low coverage. Hence, we use yeast–human alignment for evaluation of biological quality of C-GRAAL's alignment of PPI networks of these species.

2.1.2 Biological quality of yeast–human alignment. We evaluate the biological quality of alignments by analyzing whether the aligned yeast–human protein pairs perform the same biological function. Specifically, for a given alignment, we count how many aligned yeast–human protein pairs share a common GO term.[25] We downloaded GO annotation data from the Gene Ontology database§ in September 2009.

We report the statistics averaged over 30 runs for both node similarity measures.

The alignments produced using the GDV similarity measure have 45.24 ± 1.19%, 13.98 ± 0.76%, and 3.62 ± 0.55% of aligned protein pairs sharing at least 1, 2, and 3 GO terms, respectively. Compared to random alignments, the *p*-values for these percentages are 8.3×10^{-5}, 1.89×10^{-3}, and 3.2×10^{-2}, respectively (see ESI‡ for details). The alignments produced using the sequence similarity measure have 57.04 ± 1.35%, 25.83 ± 1.14%, and 14.61 ± 0.84% of aligned protein pairs sharing at least 1, 2, and 3 GO terms, respectively. Compared to random alignments, the *p*-values for these percentages are 2.56×10^{-8}, 2.61×10^{-8}, and 6.03×10^{-9}, respectively. Given that all C-GRAAL alignments have similar and statistically significant GO term enrichment (as demonstrated by small standard deviations—see above), we conclude that there are many yeast–human alignments of similar biological quality. Furthermore, it is expected that sequence similarity-based alignments will result in a larger number of pairs that share a GO term, since the majority of GO terms are inferred computationally from sequence similarities. The biological quality of C-GRAAL alignments is comparable to the one obtained by MI-GRAAL.

Given that using a topologically based node similarity measure results in alignments of higher topological quality, whereas using sequence similarity results in alignments of higher biological quality (see Tables 1 and 2), it might be useful to combine the two node similarity measures to provide a balance between the topological and biological quality of an alignment. Defining such a combined measure is a subject of future research.

Finally, we find that C-GRAAL aligns 12 yeast proteins that belong to the Mediator complex and participate in transcriptional regulation of RNA polymerase II to 12 human proteins that are also mediators of RNA polymerase II transcription.

§ http://www.geneontology.org/.

Table 2 Fraction of aligned protein pairs that share at least *k* common GO terms. The fraction of aligned protein pairs from the best yeast–human alignment that share at least *k* GO terms for each node similarity measure. Numbers in parentheses represent *p*-values

k GO terms	GDV similarity	Sequence similarity
1	45.81% (0)	56.91% (1.48×10^{-8})
2	14.48% (2.19×10^{-5})	25.24% (1.62×10^{-8})
3	4.28% (6.26×10^{-4})	14.05% (0)
4	1.49% (3.84×10^{-3})	9.43% (1.39×10^{-8})
5	0.56% (0.02)	4.16% (3.28×10^{-9})

These 12 proteins participate in 45 aligned interactions (see Fig. 2). This result is encouraging, since the mediator of RNA Polymerase II transcriptional regulation is expected to be conserved from yeast to human.[26–28] Interestingly, the conservation of this complex was not detected by the best alignment produced by MI-GRAAL (using only GDV as node similarity measure), further implying differences between the alignments produced by C-GRAAL and MI-GRAAL.

2.1.3 C-GRAAL's application to protein function prediction in eukaryotes. We have demonstrated that C-GRAAL produces large, statistically significant, and biologically meaningful alignments. Hence, we expect that these alignments can be used to predict biological characteristics (*i.e.*, GO terms for molecular function (MF), biological process (BP), and cellular component (CC)) of unannotated proteins by transferring the annotation from their annotated aligned partners. We define unannotated proteins to be those that either contain no GO terms or contain only GO terms that have not been experimentally verified. To make these predictions, we use the C-GRAAL's overall best yeast–human alignment obtained using the GDV node similarity measure (as described above).

For each aligned protein pair with an unannotated protein, we check if its aligned partner is annotated with the known MF, BP, or CC GO term. If so, we transfer the annotation, *i.e.*, we assign all known MF, BP, and CC GO terms to the unannotated protein. Here, we distinguish between the complete GO annotation data set that contains all GO annotations independent of GO evidence codes and biologically-based GO annotation data set that contains GO annotations obtained by experimental evidence codes only (see The Gene Ontology Consortium[25] for details). The biological GO data set is considered to be of higher confidence than the complete GO annotation data set, since it is not biased by annotations derived from computational approaches or sequence similarity between proteins.

With respect to biologically-based GO data set, we make BP predictions for 1380 human and 1073 yeast proteins, CC predictions for 1414 human and 971 yeast proteins, and MF predictions for 1772 human and 799 yeast proteins. To validate our predictions, we check if predicted GO terms appear in the list of existing terms (including terms that have not been experimentally verified) for that protein. For human proteins, we validate 12.15% of BP predictions, 35.19% of CC predictions, and 10.09% of MF predictions. For yeast proteins, we validate 7.21% of our BP predictions, 8.53% of CC predictions, and 7.12% of MF predictions.

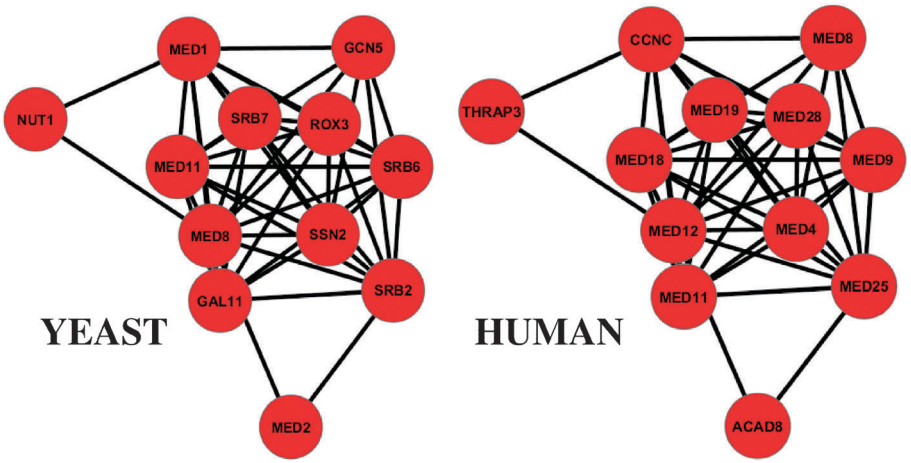


Fig. 2 C-GRAAL aligns RNA Polymerase II transcriptional regulation proteins in yeast and human.

With respect to the complete GO data set, we make BP predictions for 316 human and 133 yeast proteins, CC predictions for 274 human and 67 yeast proteins, and MF predictions for 219 human and 388 yeast proteins. To validate our predictions, we use the literature search and text mining tool CiteXplorer.[29] This tool performs an automatic search of all published articles in MEDLINE. We consider a prediction to be validated if the tool finds at least one article mentioning the protein of interest in the context of our predicted function. For human proteins, we validate 33.23% of BP predictions, 46.72% of CC predictions, and 20.09% of MF predictions. For yeast proteins, we validate 6.02% of our BP predictions, 14.93% of CC predictions, and 11.34% of MF predictions.

2.1.4 Alignment of bacterial PPI networks. We align the high-confidence part of the *C. jejuni* PPI network[20] (henceforth denoted as "CJJ") with the high-confidence PPI network of *E. coli*[21] (henceforth denoted as "ECL"). These PPI networks consist of data integrated from multiple experimental PPI data sets. The CJJ PPI network consists of 2988 interactions between 1111 proteins and the ECL PPI network consists of 3989 interactions between 1941 proteins.

Moreover, we align PPI networks of *Mesorhizobium loti* (henceforth denoted as "MZL") and *Synechocystis sp PCC6803* (henceforth denoted as "SPP"), generated by a modified high-throughput yeast two-hybrid system.[18,19] These PPI networks contain about 24% and 52% of the protein coding genes from these species, respectively. However, due to the limitations in experimental techniques, it is expected that these networks still contain false positive and false negative interactions between these proteins. The MZL PPI network consists of 3094 interactions between 1804 proteins and the SPP PPI network consists of 3102 interactions between 1920 proteins.

In MI-GRAAL, Kuchaiev *et al.* use four different combined node similarity measures, two to align CJJ and ECL, and two to align MZL and SPP PPI networks.[13] The first similarity measure, denoted as "SM-1," combines GDV similarities and degrees, the second one, denoted as "SM-2," combines GDV similarities, clustering coefficients, and BLAST *E-values*, the third one, denoted as "SM-3," combines GDV similarities, degrees, clustering coefficients, and eccentricities, and the fourth one, denoted as "SM-4," combines GDV similarities, degrees, clustering coefficients, eccentricities, and BLAST *E-values*.

To compare and evaluate C-GRAAL against MI-GRAAL on the bacterial networks, we use the same combined similarity measures.

As above, we define the best alignment to be the alignment with the highest edge correctness (EC) score over all node similarity measures and over all runs of the algorithm.

The *best* alignment between CJJ and ECL PPI networks when we use SM-1 as a node similarity measure consists of 584 edges (EC = 19.54), while the best alignment obtained for SM-2 as a node similarity measure consists of 600 edges (EC = 20.08) (for more details see Table 3). Both these alignments have the EC lower than the EC values obtained by MI-GRAAL for SM-1 (EC = 26.14) and SM-2 (EC = 24.44), respectively. However, all alignments produced by C-GRAAL are of good biological quality and contain a statistically significant fraction of protein pairs sharing at least

Table 3 C-GRAAL's alignments of bacterial PPI networks. The alignment statistics for the alignment of bacterial networks using four different combined node similarity measures: SM-1: GDV similarities and degrees; SM-2: GDV similarities, clustering coefficients, and BLAST *E*-values; SM-3: GDV similarities, degrees, clustering coefficients, and eccentricities; and SM-4: GDV similarities, degrees, clustering coefficients, eccentricities, and BLAST *E*-values. The column denoted by "Network 1–Network 2" contains names (species) of networks being aligned. The column denoted by "NSM" contains a node similarity measure used. Columns denoted by "EC (best)," "EC (avg)," and "EC (std)" represent the edge correctness (EC) of the best alignment, the average EC, and the standard deviation of EC over 30 runs, respectively. The column denoted by "p-value" represents the statistical significance of observed alignments. Columns denoted by "LCCS (nodes)" and "LCCS (edges)" represent the size of the largest common connected subgraph in terms of the number of nodes and edges, respectively, for the best observed alignments produced by C-GRAAL

Network 1–network 2	NSM	EC (best) (%)	EC (avg) (%)	EC (std) (%)	*p*-Value	LCCS (nodes)	LCCS (edges)
CJJ–ECL	SM-1	19.54	18.11	0.80	1.2×10^{-9}	441	501
CJJ–ECL	SM-2	20.08	18.43	0.72	1.2×10^{-9}	449	575
MZL–SPP	SM-3	23.79	22.97	0.39	3.6×10^{-9}	686	696
MZL–SPP	SM-4	26.02	24.71	0.74	3.6×10^{-9}	731	740

Table 4 Fraction of aligned protein pairs that share at least k common GO terms. The fraction of aligned protein pairs from the best bacterial network alignments (denoted as "best") and the average fraction and the standard deviation of aligned protein pairs over 30 runs (denoted as "avg.") that share at least 1, 2, 3, or 4 GO terms for each node similarity measure ("NSM"). Node similarity measures: SM-1: GDV similarities and degrees; SM-2: GDV similarities, clustering coefficients, and BLAST E-values; SM-3: GDV similarities, degrees, clustering coefficients, and eccentricities; and SM-4: GDV similarities, degrees, clustering coefficients, eccentricities, and BLAST E-values. Numbers in parentheses represent p-values

Alignment	NSM	1 GO term	2 GO terms	3 GO terms	4 GO terms
CJJ–ECL (best)	SM-1	33.72% (3.65×10^{-9})	11.48% (7.77×10^{-10})	3.98% (2.01×10^{-9})	1.76% (4.54×10^{-7})
CJJ–ECL (best)	SM-2	35.60% (4.31×10^{-9})	11.24% (1.54×10^{-9})	4.45% (2.29×10^{-9})	1.76% (2.82×10^{-9})
MZL–SPP (best)	SM-3	14.82% (2.13×10^{-9})	3.60% (7.16×10^{-9})	1.25% (5.45×10^{-9})	0.69% (1.32×10^{-7})
MZL–SPP (best)	SM-4	16.78% (0)	3.27% (1.17×10^{-9})	1.09% (1.24×10^{-8})	0.41% (1.03×10^{-7})
CJJ–ECL (avg.)	SM-1	$34.26 \pm 1.41\%$ (4.43×10^{-9})	$11.20 \pm 1.09\%$ (1.60×10^{-9})	$4.42 \pm 0.63\%$ (1.30×10^{-8})	$1.71 \pm 0.38\%$ (9.77×10^{-5})
CJJ–ECL (avg.)	SM-2	$33.57 \pm 1.33\%$ (4.53×10^{-9})	$10.01 \pm 0.75\%$ (1.58×10^{-9})	$3.88 \pm 0.54\%$ (2.62×10^{-9})	$1.66 \pm 0.28\%$ (6.26×10^{-6})
MZL–SPP (avg.)	SM-3	$14.46 \pm 0.96\%$ (1.27×10^{-9})	$3.28 \pm 0.38\%$ (4.20×10^{-9})	$1.03 \pm 0.25\%$ (6.87×10^{-3})	$0.20 \pm 0.21\%$ (6.30×10^{-3})
MZL–SPP (avg.)	SM-4	$15.37 \pm 1.00\%$ (1.91×10^{-9})	$3.63 \pm 0.56\%$ (1.06×10^{-7})	$1.16 \pm 0.34\%$ (3.09×10^{-5})	$0.45 \pm 0.22\%$ (9.77×10^{-5})

1, 2, 3, or 4 GO terms (see Table 4), while some of the alignments produced by MI-GRAAL for SM-1 do not contain a statistically significant fraction of protein pairs sharing GO terms.

We observe similar results in the alignment of MZL and SPP PPI networks. The *best* alignment between these PPI networks for SM-3 as a node similarity measure consists of 736 edges (EC = 23.79), while the best alignment obtained for SM-4 as a node similarity measure consists of 805 edges (EC = 26.02) (for more details see Table 3). The best alignment produced by MI-GRAAL using SM-3 node similarity has the EC of 41.79, while the best alignment for SM-4 has the EC of 39.75. Although MI-GRAAL produced alignments with higher EC values, none of its alignments obtained for SM-3 contain a statistically significant fraction of protein pairs sharing GO terms. On the other hand, all alignments produced by C-GRAAL contain a statistically significant fraction of protein pairs sharing at least 1, 2, 3, or 4 GO terms (see Table 4).

Based on these results, we can say that C-GRAAL is comparable to MI-GRAAL. Even though it does not produce alignments with better EC than MI-GRAAL, all alignments produced by C-GRAAL are of good biological quality, while some of MI-GRAAL's are not. Furthermore, these results confirm that C-GRAAL performs well on data based on Y2H experiments.

All bacterial network alignments produced by C-GRAAL are statistically significant, with a p-value of 1.2×10^{-9} (see ESI‡ for details).

2.1.5 C-GRAAL's application to protein function prediction in prokaryotes. Analogous to prediction of protein functions for yeast and human based on the yeast–human PPI network alignment, we use the best alignment between CJJ and ECL PPI networks (obtained for SM-2) and the best alignment between MZL and SPP PPI networks (obtained for SM-4) to predict function of unannotated proteins. We downloaded GO annotation data for CJJ and ECL from the European Bioinformatics Institute¶ and GO annotation data for MZL and SPP from the Kazusa DNA Research Institute‖, in March 2010.

We make BP predictions for 24 ECL and 443 CJJ proteins, CC predictions for 94 ECL and 560 CJJ proteins, and MF predictions for 61 ECL and 247 CJJ proteins. As before, we use CiteXplorer[29] to validate the predictions. For ECL proteins, we validate 17.10% of BP predictions, 55.16% of CC predictions, and 55.612% of MF predictions. For CJJ proteins, we validate 7.26% of our BP predictions, 18.72% of CC predictions, and 8.12% of MF predictions. Higher validation rates obtained for ECL can be explained by the fact that ECL (*i.e.*, *E. coli*) is a more studied organism and thus has more and better annotated proteins than CJJ.

We make BP predictions for 483 SPP and 384 MZL proteins, CC predictions for 320 SPP and 324 MZL proteins, and MF predictions for 486 SPP and 385 MZL proteins. However, we were not able to validate these predictions in the literature. One of the possible reasons for this is that these bacteria are not as well studied as ECL and CJJ, and thus, the number of articles in which they appear might be limited. Indeed, we find that MZL and SPP (the species, not their proteins) appear in about 300 PubMed** articles, while ECC and CJJ appear in over 272 000 and 5000 PubMed articles, respectively. Furthermore, CiteXplorer[29] finds an article for only 0.55% and 8.23% of MZL and SPP proteins (proteins in general, not proteins in the context of our predictions), respectively, while the same is true for 31.9% and 11.5% of ECL and CJJ proteins, respectively.

2.2 Alignment of human–pathogen PPI networks

We align three human–pathogen PPI networks: interactions of *F. tularensis* and *H. sapiens* (henceforth denoted as "FH"), *B. anthracis* and *H. sapiens* ("BH"), and *Y. pestis* and *H. sapiens* ("YH"), generated by a high-throughput yeast two-hybrid system.[17] Dyer *et al.* previously attempted to identify conserved protein interaction modules (CPIMs) amongst these three networks using existing algorithms based on homology relationship (Match-and-Split,[30] NetworkBLAST,[3] and GraphHopper[31]). Using the GraphHopper algorithm, the authors were able to identify 39, 41, and 64 CPIMs between FH and BH, FH and YH, and BH and YH PPI networks, respectively. However, when using either the Match-and-Split or NetworkBLAST algorithm, the authors were not able to identify any CPIMs.

Since various pathogenic proteins may use the same strategies to invade a human cell, we believe that by focusing mainly on homologous relationships between proteins, one may miss patterns

¶ http://www.ebi.ac.uk/.
‖ http://genome.kazusa.or.jp.

** http://www.ncbi.nlm.nih.gov/pubmed/.

Table 5 C-GRAAL's alignments of human–pathogen PPI networks. The alignment statistics for three human–pathogen PPI networks averaged over 30 runs of the algorithm. Columns denoted by "EC (max)," "EC (avg)," and "EC (std)" represent the maximum edge correctness (EC), the average EC, and the standard deviation of EC over 30 runs, respectively. Columns denoted by "LCCS (nodes)" and "LCCS (edges)" represent the size of the largest CPIMs in terms of the number of nodes and edges, respectively, of the alignment with the maximum EC produced by C-GRAAL

Alignment	EC (max) (%)	EC (avg) (%)	EC (std) (%)	LCCS (nodes) (%)	LCCS (edges) (%)
FH–BH	63.34	63.03	0.16	664	665
FH–YH	50.25	50.11	0.13	442	442
BH–YH	45.36	45.18	0.13	967	975

of human–pathogen interactions that are topologically simple to detect. Since it has been shown that sequence and topology may contain complementary information,[32] we apply the C-GRAAL algorithm using the GDV similarity measure to align these three PPI networks and to evaluate the biological information and patterns we can extract solely from topology.

The FH PPI network consists of 1345 proteins and 1383 interactions, the BH PPI network consists of 2604 proteins and 3062 interactions, and the YH PPI network consists of 3322 proteins and 4053 interactions. All networks contain uncharacterized (*i.e.*, putative) pathogenic proteins. C-GRAAL correctly aligns between 45% and 63% of edges between these three networks. All alignments are statistically significant (p-value $< 10^{-10}$). Moreover, C-GRAAL identified 39, 41, and 64 CPIMs between FH and BH, FH and YH, and BH and YH PPI networks, respectively. The largest CPIM consists of 665 aligned edges between 664 proteins in the FH–BH alignment, 422 edges between 422 proteins in the FH–YH alignment, and 975 edges between 967 proteins in the BH–YH alignment. For more details about the alignments see Table 5.

All alignments between human–pathogen PPI networks have a significant number of aligned protein pairs that share one or more GO terms (see Table 6). Furthermore, C-GRAAL detects multiple correctly aligned edges (interactions) where aligned proteins share the same GO terms. We call these interactions *conserved interactions*. C-GRAAL detects 42 connected components that consist of conserved interactions in the FH–BH alignment, 18 in the FH–YH alignment, and 38 in the BH–YH alignment, respectively. Out of 42 such components from the FH–BH alignment, 13 consist of two or more edges. Similarly, 4 out of 18 components from the FH–BH alignment and 4 out of 38 components from the BH–YH alignment also represent conserved components that consist of two or more edges. Interestingly, the majority of aligned proteins do not have statistically significant sequence similarity, and thus, such interactions could not have been discovered by methods based on homologous information.

Table 6 Fraction of aligned protein pairs that share at least k common GO terms. The fraction of aligned protein pairs from the human–pathogen alignments that share at least k common GO terms for each node similarity measure. Numbers in parentheses represent p-values

k GO terms	FH–BH	FH–YH	BH–YH
1	38.45% (0)	29.96% (0)	26.90% (0)
2	19.63% (0)	15.02% (0)	12.55% (0)
3	8.80% (0)	7.36% (0)	5.4% (0)
4	4.12% (0)	3.37% (0)	3.21% (0)
5	2.10% (0)	1.46% (0)	1.36% (7×10^{-10})

These results confirm that topology is rich in biological information and can detect biologically significant patterns. Therefore, it is expected that the results obtained by alignment of these networks can be used for annotation transfer. Analogous to prediction of protein functions for yeast and human based on the yeast–human PPI network alignment, we use the best alignment between the human–pathogen PPI networks to predict function of unannotated proteins.

We make GO term biological process predictions for 504 *H. sapiens*, 305 *F. tularensis*, 647 *B. anthracis*, and 661 *Y. pestis* proteins, cellular component predictions for 315 *H. sapiens*, 390 *F.tularensis*, 1177 *B. anthracis,* and 1182 *Y. pestis* proteins, and molecular function predictions for 485 *H. sapiens*, 270 *F.tularensis,* 620 *B. anthracis,* and 719 *Y. pestis proteins*. As before, we use CiteXplorer[29] to validate the predictions. For human proteins, we validate 16.11% of BP predictions, 17.26% of CC predictions, and 15.70% of MF predictions. However, we were not able to validate predictions for pathogenic species in the literature. One possible reason is that these pathogens are not as well studied and thus, the number of articles in which they appear might be limited. We find that CiteXplorer[29] finds an article for less than 1% *F. tularensis, B. anthracis*, and *Y. pestis* proteins (proteins in general, not proteins in the context of our predictions), respectively. We downloaded GO annotation data for all species from the European Bioinformatics Institute†† in June 2011.

3 C-GRAAL algorithm

Let $G(V,E)$ and $H(U,F)$ be two networks, where V and U are sets of nodes and E and F are sets of edges of G and H, respectively. Without loss of generality, let us assume that $|V| < |U|$ (hence $|G| < |H|$). C-GRAAL is a global network alignment algorithm that aligns each node in the smaller network G to exactly one node in the larger network H. That is, C-GRAAL's alignment of G to H is a set of ordered pairs (v,u), $v \in V$ and $u \in U$, such that no two ordered pairs share a node. We call each such ordered pair an *aligned pair*.

We denote by $\deg(v)$ the degree of a node v in network G, by $N(v)$ the set of neighbors of node v, by $N[v]$ the *closed neighborhood* of v, defined as $N[v] = N(v) \cup \{v\}$, the neighborhood density of v as:

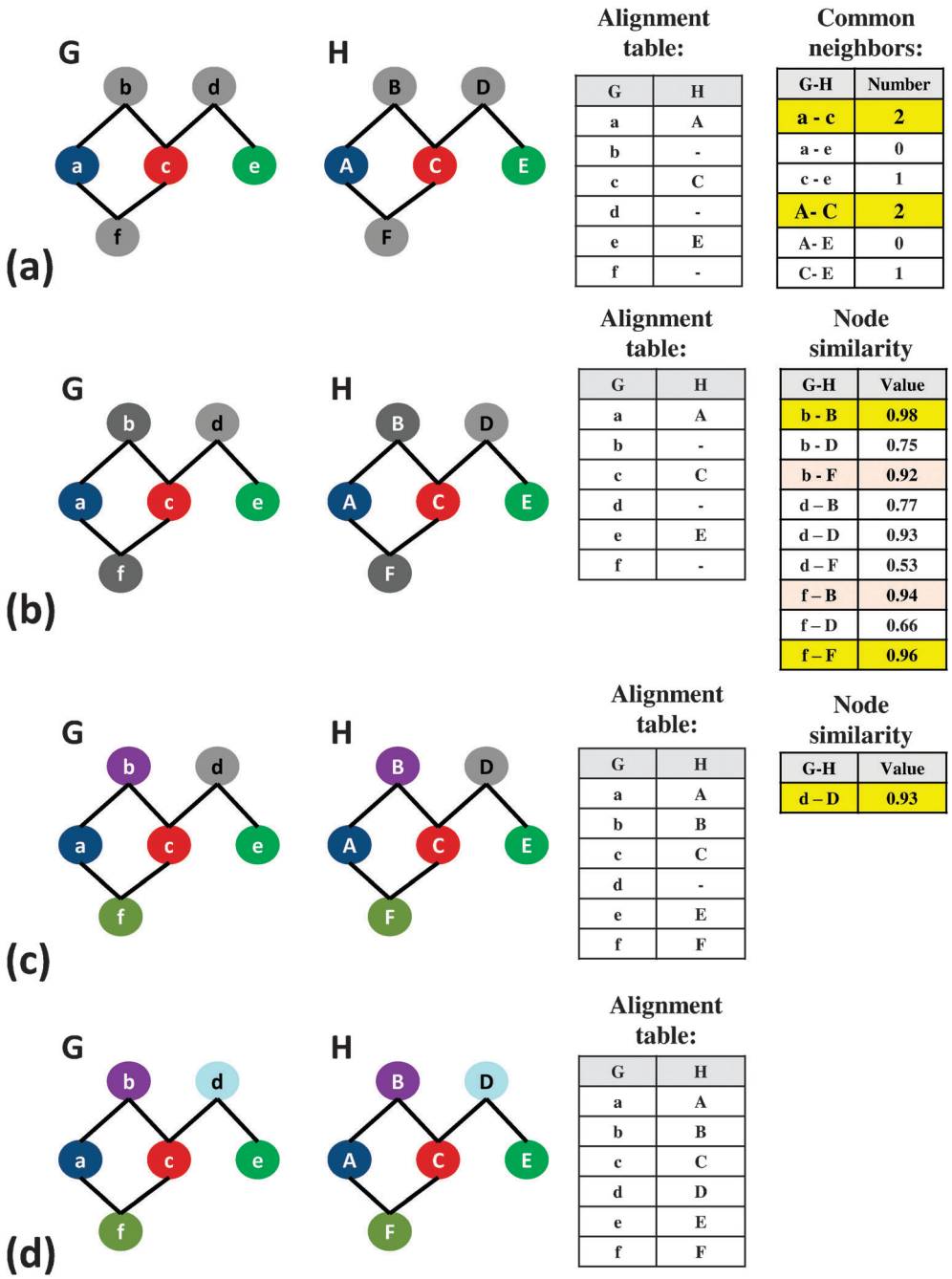
$$nd(v) = \sum_{v_k \varepsilon N[v]} \deg(v_k), \qquad (1)$$

and the combined neighborhood density of nodes v and u ($v\varepsilon G$, $u\varepsilon H$) as:

$$cnd(v,u) = \frac{nd(v) + nd(u)}{max\_nd(G) + max\_nd(H)}, \qquad (2)$$

where $max\_nd(G)$ and $max\_nd(H)$ are maximum neighborhood densities of nodes in networks G and H, respectively, $cnd(v,u) \in [0,1]$.

†† http://www.ebi.ac.uk/.



G	H
a	A
b	-
c	C
d	-
e	E
f	-

G-H	Number
a - c	2
a - e	0
c - e	1
A- C	2
A- E	0
C- E	1

G	H
a	A
b	-
c	C
d	-
e	E
f	-

G-H	Value
b - B	0.98
b - D	0.75
b - F	0.92
d – B	0.77
d – D	0.93
d – F	0.53
f – B	0.94
f – D	0.66
f – F	0.96

G	H
a	A
b	B
c	C
d	-
e	E
f	F

G-H	Value
d – D	0.93

G	H
a	A
b	B
c	C
d	D
e	E
f	F

Fig. 3 C-GRAAL—alignment of common neighbors. Let G and H be two networks we want to align, where colored (blue, red, green) nodes are already in the alignment and gray nodes are currently unaligned. Nodes that are colored with the same color are aligned to each other, *e.g.*, a to A, c to C and e to E. (a) First, we calculate the number of common neighbors for all pairs of nodes that are already in the alignment (see table "Common neighbors"). The two pairs of aligned nodes that have the largest number of aligned neighbors are (a,A) and (c,C). (2) We want to align common neighbors of nodes a and c to neighbors of nodes A and C, *i.e.*, we want to align either node b or node f to either node B or node F. To do so, we compare node similarities between these four pairs of nodes from G and H (b-B, b-F, f-B, and f-F). From the "node similarity" table, we can see that nodes b and B have the highest similarity of 0.99, and thus we align node b to node B (and we color them purple). Next, we align remaining common neighbor of nodes a and c to remaining common neighbor of nodes A and C, *i.e.*, we align node f to node F (and color them olive-green). (c) Since there are no more aligned pairs of nodes that share common neighbors, C-GRAAL performs step (1) of the algorithm. Since aligned nodes e and E have one unaligned neighbor, the flag F_{al} is set to 1, and neighbors of these nodes are aligned based on their node similarity. (d) Final alignment between networks G and H.

We denote by "F_{al}," the flag that describes the alignment status of nodes on which we perform calculations. When this flag has a value equal to 1, we take into consideration only nodes that are already aligned, otherwise we take into consideration only nodes that are not yet aligned. Initially, the flag F_{al} is set to 0, because at that point the alignment set is empty.

We can describe the C-GRAAL algorithm in three steps:

1. Finding seed and expansion around the seed;
2. Aligning common neighbors of already aligned nodes, and
3. Finalizing the alignment.

In the first step, C-GRAAL sets the flag F_{al} to 1 if the current alignment contains at least one node v from G, such that both nodes v and u, v aligned to $u \in H$, have at least one unaligned neighbor. Next, C-GRAAL calculates all-to-all combined neighborhood densities for all pairs of nodes in the two networks being aligned. The pair of nodes (v,u), $v \in G$ and $u \in H$, that have the largest combined neighborhood density represents the seed. Next, C-GRAAL expands around the seed (v,u), $v \in G$ and $u \in H$, by greedily aligning their direct neighbors v_i and u_j ($v_i \in N(v)$, $u_j \in N(u)$), based on the given node similarity measure (see Fig. 3 for an illustration).

In the second step, given the existing alignment, C-GRAAL decreasingly orders pairs of nodes v_i and v_j ($v_i,v_j \in G$) that are already in the alignment based on the number of their common neighbors in G. Next, the node pair having the highest number of common neighbors is selected. It is checked for whether the pair of nodes from H that is aligned to the selected pair in G shares at least one common neighbor. If so, based on the given node similarity measure, C-GRAAL greedily aligns the common neighbors of the selected pair in G with the common neighbors of the corresponding aligned pair in H. If not, C-GRAAL descends the ordered list and selects the next pair from G with the highest number of common neighbors. C-GRAAL repeats step (2) while there exist pairs in the alignment that have common neighbors not in the alignment and the corresponding aligned pairs in H also have common neighbors not in the alignment. When there are no such pairs left, C-GRAAL goes back to step (1). See Fig. 3 for an illustration.

C-GRAAL repeats steps (1) and (2) while there exist aligned pairs in which both nodes have at least one neighbor that is unaligned. Otherwise, C-GRAAL proceeds with step (3), where it greedily aligns all of the remaining (unaligned) nodes in G to nodes in H based only on the node similarity measure, without taking explicitly into account any network connectivity information. Each pair of nodes is aligned one at a time (one node from each network) based on the given node similarity measure.

In case that node similarity measure is not provided, the algorithm assigns the same value of similarity to all nodes. All ties in the algorithm are broken randomly.

The C-GRAAL pseudocode is given in the ESI.‡

The computational complexity of C-GRAAL is quadratic in graph size, $O(|V_{G1}|x|V_{G2}| + \max(|E_{G1}|,|E_{G2}|))$. The alignment of the yeast and human PPI networks takes about 1 hour on an Intel Xenon X3350 (2.66 GHz CPU) machine. C-GRAAL's computational complexity is the same as the computational complexity of GRAAL and similar to that of MI-GRAAL, $O(|V_{G1}|x|V_{G2}| + |E_{G1}| + |E_{G2}|)$. The computational complexity of IsoRank scales exponentially with the number of aligned networks k, as $O(E^k)$, hence it is quadratic for aligning two networks. Note that if network $G(V,E)$ is sparse, $O(E) = O(V)$, while if the network is dense, then $O(E) = O(V^2)$. Hence, IsoRank's complexity for aligning two dense networks is $O(E^2) = O(V^4)$, *i.e.*, it is the fourth-power polynomial in the network input size and hence may be prohibitively computationally expensive for large networks. In contrast, C-GRAAL's, MI-GRAAL's, and GRAAL's computational complexity remains quadratic even for dense networks.

4 Conclusions

We present a novel global network alignment algorithm, C-GRAAL, that can build an alignment between two networks solely based on network topology. As such, it can be applied to the variety of other network domains, such as social, technological, or transportational networks. We show that C-GRAAL performs comparable or better than the best currently available network alignment algorithm, and that it can be used to successfully transfer biological knowledge across species. We demonstrate that C-GRAAL performs well on data of different confidence levels and sizes, and that it consistently produces topologically statistically significant alignments. We believe that with the increase of the amount and quality of biological network data, C-GRAAL algorithm will continue to prove itself as a useful tool that can provide insights into biological function.

Acknowledgements

We thank Dr Tijana Milenković for helpful suggestions and comments about the manuscript. This project was supported by ERC Starting Independent Researcher Grant 278212, NSF CDI OIA–1028394 grant, NSF CAREER IIS–0644424 grant, and the Serbian Ministry of Education and Science Project III44006.

References

1 R. Sharan and T. Ideker, *Nat. Biotechnol.*, 2006, **24**, 427–433.
2 B. P. Kelley, Y. Bingbing, F. Lewitter, R. Sharan, B. R. Stockwell and T. Ideker, *Nucleic Acids Res.*, 2004, **32**, W83–W88.
3 R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp and T. Ideker, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 1974–1979.
4 Z. Liang, M. Xu, M. Teng and L. Niu, *BMC Bioinformatics*, 2006, **7**, 457.
5 M. Koyuturk, Y. Kim, U. Topkara, S. Subramaniam, W. Szpankowski and A. Grama, *J. Comput. Biol.*, 2006, **13**, 182–199.
6 J. Flannick, A. Novak, S. Balaji, H. Harley and S. Batzglou, *Genome Res.*, 2006, **16**, 1169–1181.
7 R. Singh, J. Xu and B. Berger, *Research in Computational Molecular Biology*, Springer, 2007, pp. 16–31.
8 R. Singh, J. Xu and B. Berger, *Proc. Pac. Symp. Biocomput.*, 2008, **13**, 303–314.
9 C.-S. Liao, K. Lu, M. Baym, R. Singh and B. Berger, *Bioinformatics*, 2009, **25**, i253–258.
10 J. Flannick, A. F. Novak, C. B. Do, B. S. Srinivasan and S. Batzoglou, *RECOMB*, 2008, pp. 214–231.
11 O. Kuchaiev, T. Milenkovic, V. Memisevic, W. Hayes and N. Przulj, *J. R. Soc. Interface*, 2010, **7**, 1341–1354.
12 T. Milenković, W. L. L. Ng, W. Hayes and N. Przulj, *Cancer Inf.*, 2010, **9**, 121–137.
13 O. Kuchaiev and N. Pržulj, *Bioinformatics*, 2011, **27**, 1390–1396.
14 T. Milenković and N. Pržulj, *Cancer Inf.*, 2008, **6**, 257–273.
15 S. Collins, P. Kemmeren, X. Zhao, J. Greenblatt, F. Spencer, F. Holstege, J. Weissman and N. Krogan, *Mol. Cell. Proteomics*, 2008, **6**, 439–450.

16 P. Radivojac, K. Peng, W. T. Clark, B. J. Peters, A. Mohan, S. M. Boyle and S. D. Mooney, *Proteins: Struct., Funct., Bioinf.*, 2008, **72**, 1030–1037.
17 M. D. Dyer, C. Neff, M. Dufford, C. G. Rivera, D. Shattuck, J. Bassaganya-Riera, T. M. Murali and B. W. Sobral, *PLoS One*, 2010, **5**, e12089.
18 Y. Shimoda, S. Shinpo, M. Kohara, Y. Nakamura, S. Tabata and S. Sato, *DNA Res.*, 2008, **15**, 13–23.
19 S. Sato, Y. Shimoda, A. Muraki, M. Kohara, Y. Nakamura and S. Tabata, *DNA Res.*, 2007, **14**, 207–216.
20 J. Parrish, J. Yu, G. Liu, J. Hines, J. Chan, B. Mangiola, H. Zhang, S. Pacifico, F. Fotouhi, V. DiRita, T. Ideker, P. Andrews and R. Finley, *GenomeBiology*, 2007, **8**, R130.
21 J. M. Peregrn-Alvarez, X. Xiong, C. Su and J. Parkinson, *PLoS Comput. Biol.*, 2009, **5**, e1000523.
22 S. F. Altschul, W. Gish, W. Miller and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**, 403–410.
23 D. Mount, *Bioinformatics—Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, 2004.
24 H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabasi, J. Tavernier, D. E. Hill and M. Vidal, *Science*, 2008, **322**, 1158684.
25 M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Nat. Genet.*, 2000, **25**, 25–29.
26 A. Tóth-Petróczy, C. J. Oldfield, I. Simon, Y. Takagi, A. K. Dunker, V. N. Uversky and M. Fuxreiter, *PLoS Comput. Biol.*, 2008, **4**, e1000243.
27 F. J. Asturias, Y. W. Jiang, L. C. Myers, C. M. Gustafsson and R. D. Kornberg, *Science*, 1999, **283**, 985–987.
28 M. Boube, L. Joulia, D. L. Cribbs and H.-M. Bourbon, *Cell*, 2002, **110**, 143–151.
29 A. Labarga, F. Valentin, M. Andersson and R. Lopez, *Nucleic Acids Res.*, 2007, **35**, W6–W11.
30 M. Narayanan and R. M. Karp, *J. Comput. Biol.*, 2007, **14**, 892–907.
31 C. G. Rivera and T. M. Murali, *Proceedings of the 1st International Conference on Bioinformatics and Computational Biology*, Berlin, Heidelberg, 2009, pp. 67–78.
32 V. Memišević, T. Milenković and N. Pržulj, *J. Integr. Bioinf.*, 2010, **7**, 135.