# CNetA: Network alignment by combining biological and topological features

Qiang Huang, Ling-Yun Wu*, and Xiang-Sun Zhang
National Center for Mathematics and Interdisciplinary Sciences
Institute of Applied Mathematics
Academy of Mathematics and Systems Science, CAS, Beijing 100190
* Corresponding author. Email: lywu@amt.ac.cn

***Abstract*—Due to the rapid progress of high-throughput techniques in past decade, a lot of biomolecular networks are constructed and collected in various databases. However, the biological functional annotations to networks do not keep up with the pace. Network alignment is a fundamental and important bioinformatics approach for predicting functional annotations and discovering conserved functional modules. Although many methods were developed to address the network alignment problem, it is not solved satisfactorily. In this paper, we propose a novel network alignment method called CNetA, which is based on the conditional random field model. The new method is compared with other four methods on three real protein-protein interaction (PPI) network pairs by using four structural and five biological criteria. Compared with structure-dominated methods, larger biological conserved subnetworks are found, while compared with the node-dominated methods, larger connected subnetworks are found. In a word, CNetA well balances the biological and topological similarity.**

## I. Introduction

In the past decade, due to the rapidly developing high-throughput techniques, more and more biomolecular networks such as protein-protein interaction (PPI) networks, gene regulatory networks and metabolic networks are constructed and collected in various database, e.g., BIND[1], DIP[2], IntAct[3], BioGRID[4], MINT[5], MPact[6], KEGG[7]. However, the biological functional annotations to the biomolecular networks do not keep up with the pace of network data growth. There is urgent demand of efficient computational tools for network analysis and annotation. As an important bioinformatics approach for biomolecular network analysis, network alignment has extensive applications such as revealing the conserved functional modules and orthologs, predicting gene functions and new interactions, and so on. Briefly speaking, the mission of network alignment is to find the global similarity and dissimilarity among different biological networks. Network alignment is an generalization of the subgraph isomorphism problem which is known to be NP-complete. Generally network alignment is much harder than the subgraph isomorphism problem because the mutations and evolutionary events have disturbed both the network structure and biomolecule functions, as illustrated in Figure 1.

Many algorithms have been proposed to solve the network alignment problem. For example, MRF based method[8], IsoRank[9], [10], IsoRankN[11], Græmlin[12], MI-GRAAL[13]. Most methods formulate the network alignment problem as an optimization problem, and solved by greedy or heuristic algorithms such as match-and-split algorithms, the seed extend algorithms, and the graph matching algorithms, and so on. According to the major features they used, network alignment methods can be categorized into three groups: structure-dominated (mainly use the structural features of the networks), node-dominated (mainly use the biological features of the nodes in networks), mixed (comprehensively use both types of features). Although the network alignment problem has been extensively studied in literature, it is far away from being solved successfully and satisfactorily. There is a trade off between the biological similarity and the topological similarity, and it is not easy to achieve good balance. The computational complexity is another important issue when dealing with large scale networks. New approaches that can efficiently and effectively solve the problem by appropriately integrating both the biological and topological information of networks are still strongly desired.

In this paper, we propose a novel network alignment approach based on the conditional random fields (CRF) model, called CNetA. CRF is a conditional probabilistic graphical model which is an extension and generalization of hidden Markov model and maximum entropy Markov model. CNetA utilizes the biological sequence similarity and network structure features, and has the ability to integrate other information. Four structural and five biological criteria are adopted to comprehensively evaluate the performance of network alignment methods. The new method is compared with a structure-dominated method, MI-GRAAL[13], and two node-dominated methods based on BLAST. The computational experiments on the real PPI networks show CNetA make better balance between the biological similarity and the topological similarity than other methods.

## II. Methods

### A. Network alignment problem

Network alignment problem can be classified into local alignment and global alignment. There are two kinds of mapping between the nodes of two aligned networks: *one-to-one* and *many-to-many*. In this paper, we only consider the global alignment, and one-to-one mapping.

2012 IEEE 6th International Conference on Systems Biology (ISB)
978-1-4673-4398-5/12/$31.00 ©2012 IEEE

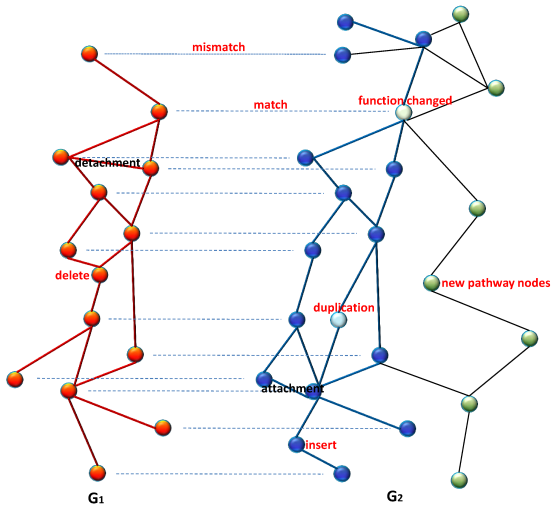



Fig. 1. An illustration of network alignment between networks $G_1$ and $G_2$. A network alignment model need to deal with the node mutations (e.g., insertion, deletion, duplication, mismatch, and also functional change) and the edge mutations (e.g., detachment, attachment).

Suppose $G = (V, E)$ and $G' = (V', E')$ are two biomolecular networks, where $V$, $V'$ are the node sets, and $E$, $E'$ are the edge sets, respectively. Network alignment problem is to find the maximum conserved subnetworks between $G$ and $G'$. A small example of network alignment problem is shown in Fig 1. The evolutionary events and mutations, including node mutations (insertion, deletion, duplication, mismatch, functional change), edge mutations (detachment, attachment), are need to handled in the computational models.

### B. Conditional random field model

The CNetA method is based on the conditional random fields model we have developed for the network querying problem[14]. Network querying problem is a special case of network alignment in which a small network is aligned with a large network. In the CRF model, network querying problem is treated as a labeling problem. The model is briefly described as follows. If we consider $G' = (V', E')$ as a label set, i.e., $V'$ is all possible labels and $E'$ is the relations between the labels, the network alignment problem can be transformed into a labeling problem. Give a network $G$ and the label set $G'$, network alignment is to find the best labels for $V$. The score of each labeling solution $Y \subseteq G'$ is computed by a conditional probability such as

$$\Pr(Y|G) = \frac{1}{Z(G)} \prod_{v_i \in V} f_N(y_i, G, i) \prod_{(v_i,v_j) \in E} f_E(y_i, y_j, G, i, j)$$

where $f_N$, $f_E$ are the feature functions, $Z(G)$ is the normalization factor. The optimal solution is the one that gives the maximal conditional probability. To deal with the insertions and deletions, we define the feature functions as follows.

$$f_N(y_i, G, i) = S(v_i, y_i),$$

$$f_E(y_i, y_j, G, i, j) = \frac{S(v_i, y_i) + S(v_j, y_j)}{2} W(y_i, y_j).$$

where $S(v_i, y_i)$ is the non-negative similarity score between nodes $v_i \in G$ and $y_i \in G'$, $W(y_i, y_j)$ is the non-negative connectivity score between nodes $y_i \in G'$ and $y_j \in G'$. In this study, $S(v_i, y_i)$ is defined based on BLAST E-value of two sequences and $W(y_i, y_j)$ is the structural measure which is reciprocal with the shortest distance between $y_i$ and $y_j$ in the network $G_2$. When $y_j$ is not reachable from $y_i$, the shortest distance is set as $L_0 \gg d_{max}$ where $d_{max}$ is the maximum distance among any connected node pairs. The details of $S(v_i, y_i)$ and $W(y_i, y_j)$ can be found in [14].

### C. Bi-directional mapping strategy

The solution obtained in the above CRF model does not guarantee one-to-one mapping. Several nodes in $G$ may be mapped to the same node in $G'$. It is not an issue in network querying since the query network is generally very small. The multiple mapping rarely occurs in the optimal solution of network querying. However, the multiple mapping problem becomes serious when the size of network increases. There are many gene duplication events in the biological evolution which results in many similar subnetworks. The larger the query network is, the higher the probability that several subnetworks of the query network are mapped onto the same subnetwork. In this paper, a bi-directional mapping strategy is proposed. This strategy can be integrated with any network querying method to obtain one-to-one network alignment.

The bi-directional mapping strategy iteratively applies the network querying method. In the $k$-th iteration, we firstly query $G$ in $G'$ and gets a subnetwork of $G'$, say $G'_k$, which is similar to $G$. Secondly $G'_k$ is queried in $G$ to obtain a subnetwork of $G$, say $G_k$. A node pair $(x, y)$, $x \in G_k$, $y \in G'_k$, is called bi-directional matching if $x$ is mapped to $y$ in the first querying and $y$ is mapped to $x$ in the second querying. Then we fix the feature functions to ensure that $x$ can only be mapped to $y$ and vise versa. In detail, we set set $f_N(y_i, G, i) = 1$ if $(v_i, y_i)$ is bi-directional matching, otherwise, $f_N(y_i, G, i) = 0$. The iterative process terminates when the bi-directional matching pairs are not changed within two consecutive iterations. Finally, the one-to-one bi-directional matching pairs in the final iteration are extracted as the results.

### D. Evaluation measures

There are many criteria for evaluating the performance of network alignment methods. In this study, we adopt two kinds of measures to assess the alignment results from the biological and topological perspective respectively.

*Biological measures.* The mostly used biological criterion for network alignment is based on the number of shared Gene Ontology (GO)[15] or the functional similarity between the GO terms of the matching nodes. The first measure is the fraction of matching pairs that share at least $k$ GO terms (SGO)[13]. In order to investigate the effects of GO domains and depth, we further compute the GO coverage of each GO domain, which is defined as the percentage of matching


978-1-4673-4398-5/12/$31.00 ©2012 IEEE

pairs that share at least one GO term with depth $d \geq 3$. Here, the GO term depth $d$ is defined as the shortest distance from the root of GO hierarchy. The homology and pathway information are used to more deeply compare the alignment results and we use the measures in Græmlin[12]. The third measure is the number of hit pathways (HP). Hit pathways are the pathways in KEGG [7] which align at least three proteins to their counterparts in the other network. We also calculate the pathway average coverage (PAC), that is, the average fraction of proteins correctly aligned in hit pathways. Finally, to assess the homology, the number of KEGG[7] orthologous (OP) proteins in alignment results is computed which is not the same as the corresponding measure in Græmlin[12].

*Topological measures.* The first topological measure for the alignment results is the number of matching pairs (MP), i.e. the number of aligned nodes. The second measure, edge correctness (EC)[13], is the fraction of correctly aligned edges which is defined as:

$$EC = \frac{|\{(u,v) \in E \bigwedge (u',v') \in E'\}|}{|E|}$$

where $u, v \in V$, and $u', v' \in V'$ is the matching nodes of $u, v$ respectively. To take account for the partial changes of network structure, we propose an extended version of EC, edge accumulated coverage (EAC):

$$EAC(k) = \frac{|\{(u,v) \in E \bigwedge d(u',v') \leq k\}|}{|E|}$$

where $u, v \in V$, $u', v' \in V'$ is the matching nodes of $u, v$ respectively, $d(u', v')$ is the distance between $u'$ and $v'$ in $G'$, and $k = 1, 2, 3, \cdots$. Obviously, $EAC(1) = EC$. EAC is an approximate edge correctness measure considering the node insertion and deletion in network evolution. Another important indicator is the size of largest common connected subgraph (LCCS)[13] that each of the aligned networks have as an exact copy. However, due to most PPI networks in current databases are not complete, the LCCS may not reflect the real situation exactly.

## III. Results

### A. Comparison settings

In order to comprehensively investigate the capability of CNetA to integrate the biological and topological features, we compare it with two kinds of network alignment methods. For comparison with structure-dominated methods, we select MI-GRAAL[13] which can reveal large structural similarity and integrate any number and type of similarity measures. We also apply the network querying method CNetQ[14], which is based on the same CRF model, to test the effectiveness of bi-directional mapping strategy. We note that CNetQ generates multiple-to-one mapping. For comparison with node-dominated methods, we compare CNetA with two BLAST[16] based methods which only use the sequence information. The first one, BLASTQ, simply query each node of $G$ in $G'$ by BLAST. Similar to CNetQ, the results of BLASTQ may be multiple-to-one mapping. The other method is BLASTA which further integrates BLASTQ with the iterative bi-directional mapping strategy used in CNetA. In each iteration, if two nodes are bi-directional matching, the corresponding BLAST E-value are set as 0.

MI-GRAAL[13] has a random process and every run may generate different results. In this study, we use the most stable score metrics described in [13] and run five times for each alignment experiment. We choose the alignment result with maximum EC as its final result.

To fairly compare CNetA and CNetQ, we set the parameter $L_0 = 10000$ in both methods. We note that $L_0 =$ infinity leads to $f_E(y_i, y_j, G, i, j) = 0$ when $y_j$ is not reachable from $y_i$, which implies several connected components can not be matched with one single connected component of the other network. However, due to evolution and data missing, large real biomolecular network may consists of many disconnected subnetworks which should be aligned with one connected subnetwork in the other network. Therefore, we do not set $L_0$ to infinity as in [14].

### B. Experimental results

In this section, we show the computational results of several methods for aligning three real PPI networks which are used by MI-GRAAL[13]. GO[15] ontology data were obtained by Matlab Bioinformatics toolbox in November 2011. KEGG pathway and orthologous protein analysis are performed by using Matlab KEGG API web service. Local executable BLAST is version 2.2.21 which was downloaded from http://blast.ncbi.nlm.nih.gov/Blast.cgi. Yeast and human GO annotation data were downloaded from GO website in November 2011, and other species GO annotation data were downloaded from European Bioinformatics Institute (EMBL-EBI) website in May 2012. We use BP, CC, MF as the abbreviation of three GO domains biological process, cellular component, and molecular function respectively.

*1) Yeast-Human PPI network alignment:* The high-confidence *Saccharomyces cerevisiae* PPI network[17] contains 2390 proteins and 16127 interactions, while human PPI network[18] contains 9141 proteins and 41456 interactions. The sequences of yeast proteins were downloaded from Saccharomyces Genome Database (SGD, http://www.yeastgenome.org)[19] and the sequences of human proteins were got from [18]. The alignment results of five methods are shown in Table I and Figure 2.

Although MI-GRAAL gets the largest structural similar subnetwork (LCCS equals to 1467), it fails to reveal the biological similarity. Figure 2(b) shows that two matching nodes identified by MI-GRAAL have few common GO terms, i.e. the two matching nodes may be not very similar in biological sense. For example, only less than 50% pairs of matching nodes have one or more common GO terms, and less than 10% for 3 or more common terms, while the percentages for other methods are larger than 80% and 60% resepectively. MI-GRAAL gets very poor GO coverage and only one orthologous protein pairs. In KEGG[7] pathway analysis, from totally 30


978-1-4673-4398-5/12/$31.00 ©2012 IEEE

pathways which have the same definition in two species, MI-GRAAL only hits 2 pathways and covers 3.33% proteins in hit pathways, while other methods get at least 26 hit pathways and their PAC are larger than 20%. In a word, MI-GRAAL focuses more on the topological similarity than the biological similarity.

As expected, BLAST based methods get largest scores for the biological measures such as SGO (Figure 2(b)), GO coverage, HP, PAC and OP. However, their scores of topological measures are worst, for example, EC and LCCS. BLASTA ourperforms BLASTQ in terms of both biological and topological measures, which show the bi-directional mapping strategy is powerful.

Compared with MI-GRAAL, CNetA/CNetQ dramatically improve the biological similarity in the results at the cost of acceptable decline in the topological similarity. Compared with BLAST based methods, CNetA/CNetQ gets the comparable results from the biological point of view, with larger EC, LCCS and EAC, which means that CNetA/CNetQ can find large structurally conversed subnetworks preserving the biological similarity as much as possible. Compared with CNetQ, CNetA gets one more hit pathway, larger PAC, more orthologous protein pairs. With the bigger matched pairs, CNetA finds more functional similar matched pairs measured by GO coverage and SGO, which implies that the bi-directional strategy is useful to identify more orthologous proteins and functional similar proteins. The smaller EC and LCCS of CNetA may owe to the missing edges in the high-confidence PPI networks since that the EAC curves of both methods are comparable.

| Method | MI-GRAAL | CNetQ | CNetA | BLASTQ | BLASTA |
|---|---|---|---|---|---|
| MP | 2390 | 1029 | 1694 | 1297 | 1672 |
| EC | 12.88% | 15.29% | 9.25% | 4.81% | 6.52% |
| LCCS | 1467(1508) | 205(956) | 116(376) | 47(141) | 55(172) |
| GO coverage (depth $\geq$ 3) | | | | | |
| MF | 5.68% | 47.78% | 54.61% | 55.07% | 56.43% |
| BP | 3.99% | 52.01% | 53.97% | 58.55% | 58.10% |
| CC | 38.95% | 72.20% | 72.73% | 76.33% | 74.74% |
| KEGG analysis | | | | | |
| OP | 1 | 331 | 556 | 583 | 719 |
| HP | 2 | 26 | 27 | 27 | 27 |
| PAC | 3.33% | 21.80% | 32.35% | 31.66% | 35.06% |

TABLE I
YEAST-HUMAN ALIGNMENT RESULTS.

MP: Matching pairs; EC: edge correctness; LCCS: Largest common connected subgraph; MF: Molecular function; BP: Biological process; CC: Cellular component; OP: Orthologous proteins; HP: Hit pathways; PAC: Pathway average coverage. The numbers in LCCS are the number of nodes and edges of LCCS respectively.

*2) Campylobacter jejuni-Escherichia loli PPI network alignment: C. jejuni* PPI network[20] contains 1091 proteins and 2966 interactions, and *E. coli* PPI network[21] contains 1873 proteins and 3803 interactions. The networks are not completely the same as the networks used in MI-GRAAL[13]. The sequence data were downloaded from Uniprot[22]. All results are shown in Table II and Figure 3.

The experimental results are similar as yeast-human alignment results. There are totally 12 pathways which have the same definition in two species in KEGG[7] database. CNetA and BLASTA hit 11 pathways with PAC larger than 29%, while MI-GRAAL only hits 3 pathways with PAC 9.47%. CNetQ and BLASTQ are slightly worse. In this experiment, CNetA gets much smaller topological measures than MI-GRAAL because PPI networks of *C. jejuni* and *E. coli* are not complete and include many small disconnected subnetworks. Compared with CNetQ and BLASTQ, CNetA and BLASTA get remarkable improvement in biological measures with similar topological measures respectively.

| Method | MI-GRAAL | CNetQ | CNetA | BLASTQ | BLASTA |
|---|---|---|---|---|---|
| MP | 1091 | 444 | 677 | 533 | 711 |
| EC | 23.33% | 1.69% | 1.21% | 0.37% | 0.84% |
| LCCS | 598(634) | 7(6) | 7(6) | 3(2) | 4(3) |
| GO coverage (depth $\geq$ 3) | | | | | |
| MF | 2.53% | 27.70% | 30.58% | 30.96% | 32.21% |
| BP | 0.84% | 23.87% | 26.44% | 28.33% | 30.38% |
| CC | 4.60% | 12.39% | 14.33% | 13.88% | 14.35% |
| KEGG analysis | | | | | |
| OP | 0 | 95 | 146 | 152 | 206 |
| HP | 3 | 10 | 11 | 10 | 11 |
| PAC | 9.47% | 15.40% | 29.61% | 21.91% | 36.68% |

TABLE II
C. JEJUNI-E. COLI ALIGNMENT RESULTS

MP: Matching pairs; EC: edge correctness; LCCS: Largest common connected subgraph; MF: Molecular function; BP: Biological process; CC: Cellular component; OP: Orthologous proteins; HP: Hit pathways; PAC: Pathway average coverage. The numbers in LCCS are the number of nodes and edges of LCCS respectively.

*3) Mesorhizobium-Synechocystis PPI network alignment: Mesorhizobium loti*[23] and *Synechocystis sp. PCC6803*[24] have 3094 interactions among 1804 proteins and 3102 interactions among 1920 proteins, respectively. The sequence data were downloaded from Kazusa DNA Research Institute (http://www.kazusa.or.jp/e/). All results are shown in Table III and Figure 4.

Since the orthologous proteins of two species are not well studied until now, we do not compare the OP for this experiment. There are only 2 pathways which have the same definition in KEGG database. The experimental results are similar to the above two experiments, which show that CNetA can well balance the biological similarity and topological similarity, and reveal more function similar matching protein pairs.

| Method | MI-GRAAL | CNetQ | CNetA | BLASTQ | BLASTA |
|---|---|---|---|---|---|
| MP | 1803 | 414 | 744 | 414 | 764 |
| EC | 41.69% | 2.52% | 1.55% | 0% | 0.097% |
| LCCS | 1149(1155) | 31(35) | 10(9) | 1(0) | 2(1) |
| GO coverage (depth $\geq$ 3) | | | | | |
| MF | 2.55% | 26.52% | 33.60% | 28.16% | 38.24% |
| BP | 1.36% | 23.84% | 24.56% | 24.76% | 31.67% |
| CC | 0.51% | 8.52% | 8.23% | 9.22% | 9.72% |
| KEGG analysis | | | | | |
| HP | 1 | 1 | 2 | 1 | 2 |
| PAC | 1.76% | 1.06% | 3.43% | 0.82% | 5.45% |

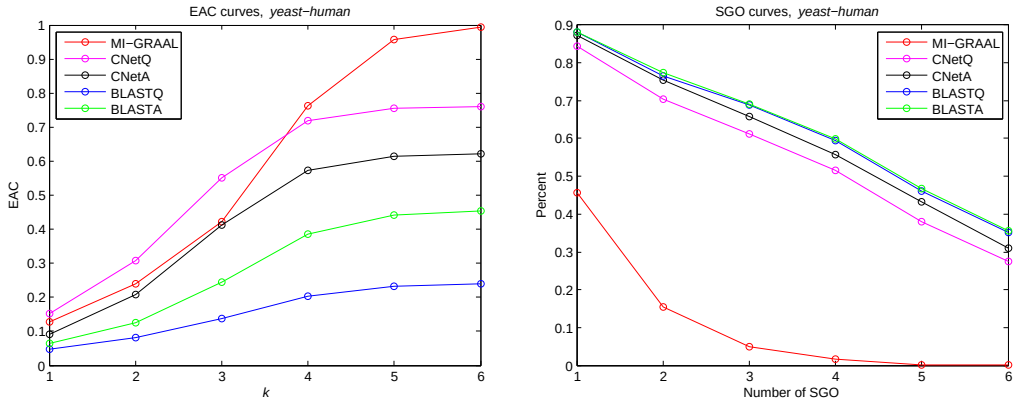
TABLE III
MESORHIZOBIUM - SYNECHOCYSTIS ALIGNMENT RESULTS

MP: Matching pairs; EC: edge correctness; LCCS: Largest common connected subgraph; MF: Molecular function; BP: Biological process; CC: Cellular component; OP: Orthologous proteins; HP: Hit pathways; PAC: Pathway average coverage. The numbers in LCCS are the number of nodes and edges of LCCS respectively.

## IV. CONCLUSION AND DISCUSSION

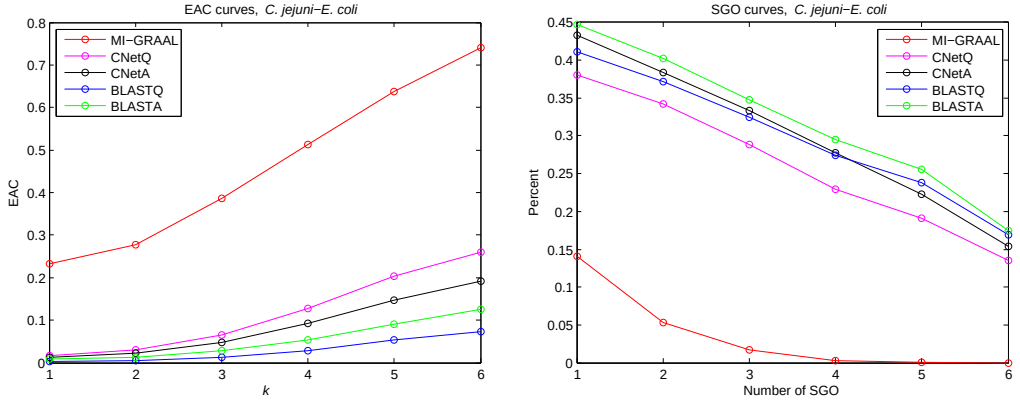
A network alignment method based on the CRF model, called CNetA, is presented in this paper. CNetA employs the iterative bi-directional mapping strategy to identify one-to-one mapping instead of multi-to-one mapping results in CNetQ, the CRF-based network querying method. The bi-directional mapping strategy also improves the biological similarity measures since the bi-directional matching proteins are more likely to be evolutional conserved. This is also confirmed by the comparison between the results of BLASTQ and BLASTA.


978-1-4673-4398-5/12/$31.00 ©2012 IEEE



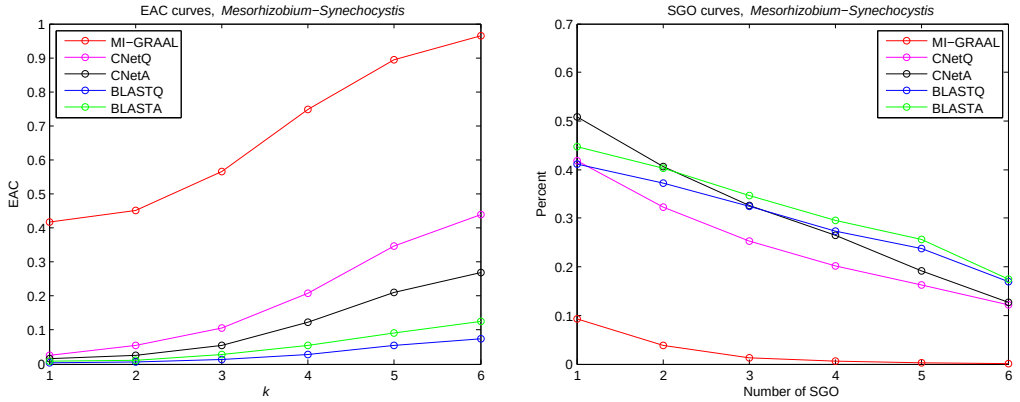

(a) (b)

Fig. 2. EAC and SGO curves for aligning yeast and human PPI networks. (a) EAC curves. The x-axis is the distance $k$ between two nodes aligned to two ends of edges. The y-axis is $EAC(k)$. The legend is the network alignment methods. (b) SGO curves. The x-axis is the number of shared GO terms. The y-axis is the percentage of matching protein pairs.




(a) (b)

Fig. 3. EAC and SGO curves for comparing *C. jejuni* and *E. coli* PPI networks. The legends are the same as Figure 2.




(a) (b)

Fig. 4. EAC and SGO curves for comparing *Mesorhizobium* and *Synechocystis* PPI networks. The legends are the same as Figure 2.


978-1-4673-4398-5/12/$31.00 ©2012 IEEE

Since there is a tradeoff between the biological similarity and topological similarity, the performance of network alignment methods can not be evaluated by a single measure. We collect several biological and topological measures from literature to access the network alignment results. Several new measures are also developed in order to better compare network alignment methods. For example, we extend the edge correctness measure to edge accumulated coverage which considers the node insertion and deletion in network evolution.

As a representative of structure-dominated methods, MI-GRAAL[13] tries to align all proteins in the small network to the large network. However, it may not be proper in the network alignment problem, since that two real networks are impossible to match perfectly. Instead, CNetA aims to find the high quality matching proteins which constitute conserved subnetworks. The network alignment results are not convincible if the functional similarities between matching proteins are too low. In other words, the biological similarity should play an equally important role as the topological similarity in network alignment, if not more important. As shown by the computational experiments on real PPI networks, CNetA can find the high quality network alignment with both biologically and topologically conserved subnetworks, which can be useful for downstream analysis such as protein function prediction.

Although the network alignment has been extensively studied in literature, there still exists many problems which are not solved completely. For example, lack of the benchmark datasets and measures for evaluating and comparing the network alignment methods. There are many datasets, including simulated and real datasets, and measures used for testing network alignment methods proposed in literature. However, there is no standard and widely accepted datasets and measures in the field of network alignment, which make the comparison of network alignment methods difficult. We note that the biomolecular databases are currently not complete which is not considered in most network alignment studies. As shown in this paper, when two networks are not complete, the true alignment may contain many disconnected pieces. In this case, if the topological similarity is emphasized too much, the biological meanings of alignment results may be reduced. Finally, the multiple network alignment is still a big challenge and rare in literature, but it is absolutely one of the most important directions in this field and need more attention from more researchers.

## ACKNOWLEDGES

*Funding*: This work is supported by Shanghai Key Laboratory of Intelligent Information Processing, China (Grant No. IIPL-2012-004).

*Conflict of interest statement*. None declared.

## REFERENCES

[1] G. Bader, D. Betel, and C. Hogue, "Bind: the biomolecular interaction network database," *Nucleic acids research*, vol. 31, no. 1, pp. 248–250, 2003.

[2] I. Xenarios, D. Rice, L. Salwinski, M. Baron, E. Marcotte, and D. Eisenberg, "Dip: the database of interacting proteins," *Nucleic acids research*, vol. 28, no. 1, pp. 289–291, 2000.

[3] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia *et al.*, "Intact: an open source molecular interaction database," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D452–D455, 2004.

[4] C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "Biogrid: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[5] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, "Mint: a molecular interaction database," *FEBS letters*, vol. 513, no. 1, pp. 135–140, 2002.

[6] U. Güldener, M. Münsterkötter, M. Oesterheld, P. Pagel, A. Ruepp, H. Mewes, and V. Stümpflen, "Mpact: the mips protein interaction resource on yeast," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D436–D441, 2006.

[7] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 27, no. 1, p. 29, 1999.

[8] S. Bandyopadhyay, R. Sharan, and T. Ideker, "Systematic identification of functional orthologs based on protein network comparison," *Genome research*, vol. 16, no. 3, pp. 428–435, 2006.

[9] R. Singh, J. Xu, and B. Berger, "Pairwise global alignment of protein interaction networks by matching neighborhood topology," in *Research in computational molecular biology*. Springer, 2007, pp. 16–31.

[10] ——, "Global alignment of multiple protein interaction networks with application to functional orthology detection," *Proceedings of the National Academy of Sciences*, vol. 105, no. 35, pp. 12 763–12 768, 2008.

[11] C. Liao, K. Lu, M. Baym, R. Singh, and B. Berger, "Isorankn: spectral methods for global alignment of multiple protein networks," *Bioinformatics*, vol. 25, no. 12, pp. i253–i258, 2009.

[12] J. Flannick, A. Novak, B. Srinivasan, H. McAdams, and S. Batzoglou, "Græmlin: general and robust alignment of multiple large interaction networks," *Genome research*, vol. 16, no. 9, pp. 1169–1181, 2006.

[13] O. Kuchaiev and N. Pržulj, "Integrative network alignment reveals large regions of global network similarity in yeast and human," *Bioinformatics*, vol. 27, no. 10, p. 1390, 2011.

[14] Q. Huang, L. Wu, and X. Zhang, "An efficient network querying method based on conditional random fields," *Bioinformatics*, vol. 27, no. 22, pp. 3173–3178, 2011.

[15] M. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall *et al.*, "The gene ontology (go) database and informatics resource," *Nucleic acids research*, vol. 32, no. Database issue, p. D258, 2004.

[16] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[17] S. Collins, P. Kemmeren, X. Zhao, J. Greenblatt, F. Spencer, F. Holstege, J. Weissman, and N. Krogan, "Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae," *Molecular & Cellular Proteomics*, vol. 6, no. 3, pp. 439–450, 2007.

[18] P. Radivojac, K. Peng, W. Clark, B. Peters, A. Mohan, S. Boyle, and S. Mooney, "An integrated approach to inferring gene–disease associations in humans," *Proteins: Structure, Function, and Bioinformatics*, vol. 72, no. 3, pp. 1030–1037, 2008.

[19] J. Cherry, E. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. Chan, K. Christie, M. Costanzo, S. Dwight, S. Engel *et al.*, "Saccharomyces genome database: the genomics resource of budding yeast," *Nucleic Acids Research*, vol. 40, no. D1, pp. D700–D705, 2012.

[20] J. Parrish, J. Yu, G. Liu, J. Hines, J. Chan, B. Mangiola, H. Zhang, S. Pacifico, F. Fotouhi, V. DiRita *et al.*, "A proteome-wide protein interaction map for campylobacter jejuni," *Genome biology*, vol. 8, no. 7, p. R130, 2007.

[21] J. Peregrín-Alvarez, X. Xiong, C. Su, and J. Parkinson, "The modular organization of protein interactions in escherichia coli," *PLoS computational biology*, vol. 5, no. 10, p. e1000523, 2009.

[22] U. Consortium *et al.*, "Reorganizing the protein space at the universal protein resource (uniprot)," *Nucleic Acids Res*, vol. 40, pp. D71–D75, 2012.

[23] Y. Shimoda, S. Shinpo, M. Kohara, Y. Nakamura, S. Tabata, and S. Sato, "A large scale analysis of protein–protein interactions in the nitrogen-fixing bacterium mesorhizobium loti," *DNA research*, vol. 15, no. 1, pp. 13–23, 2008.

[24] S. Sato, Y. Shimoda, A. Muraki, M. Kohara, Y. Nakamura, and S. Tabata, "A large-scale protein–protein interaction analysis in synechocystis sp. pcc6803," *DNA research*, vol. 14, no. 5, pp. 207–216, 2007.


978-1-4673-4398-5/12/$31.00 ©2012 IEEE