TELKOMNIKA, Vol. 11, No. 10, October 2013, pp. 6202 ~ 6208
ISSN: 2302-4046

A Topology-Based Algorithm for Directed Network Alignment

Qingyu Zou, Fu Liu*, Tao Hou, Yihan Jiang
College of Communication Engineering, Jilin University, Changchun 130000, Jilin, China
Corresponding author, e-mail: qyzou11@mails.jlu.edu.cn, liufu@jlu.edu.cn

Abstract

Network alignment has brought significant advances to our understanding of complex networks, e.g., the World Wide Web, biological networks, and social networks. Triangles comprised of three nodes are the simplest subnet in the directed network. The topological distribution of triangles is an important indicator of understanding the dynamics and function of directed networks. In this paper, we present a novel alignment algorithm for direct networks only based on topology structure, which can be used for any two networks. The transcriptional regulatory networks (TRNs) of E Coli and S Crevisiae are used to evaluate the algorithm. Experimental results demonstrate that the algorithm proposed is efficient for aligning directed networks.

Keywords*: Alignment, Directed Network, Topology, Triangle, Nodes Similarity*

Copyright © 2013 Universitas Ahmad Dahlan. All rights reserved.

1. Introduction

Recently many complex systems in nature and society have been represented as networks to understand structure, dynamic, robust, and evolution [1-3]. A network is consisted of nodes and links. Nodes of the network represent the elements of the complex system and two nodes are connected by a link if they are somehow related [4]. Network alignment is the problem of finding a good mapping between nodes of a pair of different networks with similarities between the structure or topology of them. It has an enormous impact on our understanding of networks [5-11]. Despite the researches in this field grow rapidly, network alignment is still in its infancy. Thus, much research is needed to formally define the problem of network alignment, develop algorithms to solve it accurately, and design applications of network alignment to produce meaningful results. When we know a lot about some of the nodes in one network and almost nothing about their topologically similar nodes in the other network alignments of networks can be specialized knowledge about one may tell us something new about the other.

The majority of methods for network alignment have focused on local network alignments, targeting only pairwise alignments [5, 6, 12]. Local network alignments typically match a small subnetwork from one network to one subnetworks in another network. So far, many algorithms for local alignment have been developed. Kelley et al. search for high-scoring pathway alignments between pairs of node interaction paths in two different networks [13]. Koyuturk et al. develop a framework for comprehensive alignment of two networks, which is inspired by models that focus on understanding the evolution of protein interactions [14]. Berg and Lassig develop an evolutionary grounded method for the cross-species analysis of interaction networks by alignment [15]. Different from the above algorithm that primary purpose to detect conserved subnetworks, the global network alignment provides a unique, one-to-one mapping for every node in one network to exactly one node in the other network. Singh et al. use spectral graph theory to compute scores of aligning pairs of nodes from different networks [16]. Flannick et al. has been extended to allow global network alignment based on a learning algorithm. Liao et al. relies on the notion of node-specificrankings and uses a method similar to PageRank-Nibble algorithm in order to improve upon existing global alignment results [17]. Kuchaiev et al. [18] and Milenkovic et al. [19] propose a network alignment algorithm with the cost function based solely and explicitly on a strong, theoretically grounded, direct measure of network topological similarity.

Received April 19, 2013; Revised July 15, 2013; Accepted July 25, 2013

Nearly all of these methods intended for the analysis of undirected network data. However, many of the networks that we would like to study are directed including the World Wide Web, food webs, many biological networks, and even some social networks. The commonest algorithm for directed networks has been simply to ignore the link directions and apply algorithms designed for undirected networks. It is clear that we are throwing away a good deal of information about our network's structure information.In this paper, we present a novel algorithm that is only based on direct measures of network topological similarity. On account of our methods only use network topological information, they can align any two networks. Network alignment provides a fitness between two networks, but it is not obvious how to measure the "goodness" of an inexact fit. To address this problem we use a measure *LC* [18] to assess the alignment quality.

We apply our algorithm to TRNs and show that they produce by far the most complete topological alignments of TRNs to date. Our alignment of the gene-gene interaction networks of two famous model organisms, E. coli and S. cerevisiae. The consequence shows that our method works effectively in alignment for directed networks.

2. Research Method

2.1. Triangle Vertexes Weightiness

In general, the simple building blocks of complex networks are a small structure of several nodes called motif [20]. Network motifs are small subgraphs that can be found in a network statistically significantly more often than in randomized networks. Among the possible motifs, the simplest one is the triangle which represents the basic unit of transitivity and redundancy in a graph, see Figure 1.

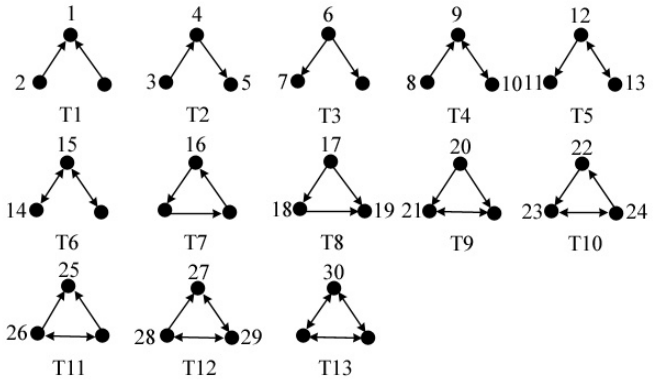

Figure 1. List of all 13 types of triangles

As shown in Figure 1, there are 13 triangle cases at most, including 39 vertexes, in an arbitrary directed network. We compare all three vertexes one another for each triangle T_i and merge the code of vertexes had the same place. Then, there are 30 special vertexes for triangles, encoded from 1 to 30 in Figure 1. We assign different weights w_i to different vertexes i, because some complex triangles contain the simple triangles, such as triangle 11 contain 1. We assign higher weights to the vertexes whose are not affected by other vertexes, and lower weights to depend on other vertexes. The w_i is calculated using a function as follows:

$$w_i = \frac{TC_i}{\max(TC_i)} \quad (1)$$

where TC_i means the number of vertexes affected by vertex i. We consider that each vertex affects itself. For instance, for vertexes 1, TC_1=2, since it affects vertexes 25 and itself; similarly, TC_6 = 3, since vertex 6 affected vertexes 17, 20 and itself. The weights of 30 vertexes as shown in Table 1.

Table 1. The weights of 30 vertexes

Vertex	Weight	Vertex	Weight	Vertex	Weight	Vertex	Weight	Vertex	Weight	Vertex	Weight
No.1	0.5	No.6	0.75	No.11	1	No.16	0.25	No.21	0.25	No.26	0.25
No.2	0.5	No.7	1	No.12	1	No.17	0.25	No.22	0.25	No.27	0.25
No.3	0.75	No.8	0.75	No.13	1	No.18	0.25	No.23	0.25	No.28	0.25
No.4	0.75	No.9	1	No.14	1	No.19	0.25	No.24	0.25	No.29	0.25
No.5	0.75	No.10	0.75	No.15	0.75	No.20	0.25	No.25	0.25	No.30	0.25

2.2. Triangle Degree

The number of triangles that the node touches is the triangle degree of it. For a node *u*, as shown Figure 2, the triangle degree values of the 30 vertexes of node *u*, are presented in the Table 2.

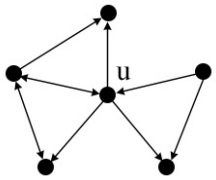

Figure 2. An illustration of a small network containing node *u*. The values of the 30 vertexes of the triangle degree are shown in Table 2

Table 2. Values of the 30 vertexes of the triangle degree of node *u* in Figure 2

Vertex	Degree	Vertex	Degree	Vertex	Degree	Vertex	Degree	Vertex	Degree	Vertex	Degree
No.1	0	No.6	3	No.11	0	No.16	0	No.21	0	No.26	1
No.2	2	No.7	1	No.12	2	No.17	0	No.22	0	No.27	0
No.3	0	No.8	2	No.13	0	No.18	0	No.23	0	No.28	1
No.4	3	No.9	1	No.14	1	No.19	0	No.24	0	No.29	0
No.5	0	No.10	3	No.15	0	No.20	0	No.25	0	No.30	0

2.3. Nodes Similarity

The nodes similarity, based on the frequencies of the appearance of all 13 triangles, measure the degree of approximation of a pair of nodes originating in different networks. For a pair of nodes *u* in network *G* and *v* in network *H*, the nodes similarity of vertex *i* between them $D_i(u,v)$ is defined as:

$$D_i(u,v) = 1 - w_i \times \left| \frac{u_i}{\deg(u)} - \frac{v_i}{\deg(v)} \right| \quad (2)$$

where deg(*u*) means the degree of a node *u* in network *G*; u_i is the number of vertex *i* touched by node *u* in *G* and v_i is the number of vertex *i* touched by node *v* in *H*. Because of higher degree nodes with similar provide a tighter constraint than correspondingly similar low degree nodes. We define the nodes similarity $S(u,v)$ between nodes *u* and *v* as:

$$S(u,v) = \frac{\sum_{i=0}^{30} D_i}{\sum_{i=0}^{30} w_i} \quad (3)$$

2.4. Alignment Algorithm

The alignment algorithm aligns a pair of different networks based on the nodes similarity measure. The processes aligning two networks $G(U,E)$ and $H(V,F)$ are described in the following steps:

Stage 1: compute the nodes similarity between each node *u* in *G* and each node *v* in *H*.

Stage 2: choose the initial pair of nodes *u* and *v* that has the maximum similarity.

Stage 3: take the two sets of nodes P_G and P_H rounded *u* and *v* that the shortest path from *u* to *p* less than one third of diameter of network *G* and *H*, respectively, and greedily aligns the two sets by finding node pairs (*u*,*v*):$u \in P_G$, $v \in P_H$ with maximum nodes similarity.

Stage 4: when all around the seed have been aligned, some nodes in both networks may remain unaligned. So, repeats the same method on a pair of networks and searches for the new seed again.

Stage 5: stop the calculate process when each node from *G* is aligned to exactly one node in *H*.

2.5. Measures of the Alignment Quality

The alignment algorithm produces global alignments to find the best way to fit network *G*(*U*,*E*) into *H*(*V*,*F*), if the number of nodes in *G* less than *H*. In order to evaluate the alignment quality of a fit, a new measure called "link correctness" (*LC*) has been introduced. The *LC* could be to assess the number of aligned links–that is, the percentage of links in *G* that are aligned to links in *H*. The *LC* is defined as:

$$LC = \frac{\left|\{ f(E) \in F \} + \{ d(E) \in F \}\right|}{\left|E\right|} \quad (4)$$

where *f*(*E*) means the link alignment produced by alignment algorithm; *d*(*E*) means the direction of links produced by alignment algorithm. High *LC* means that networks *G* and *H* share similar topologies. In particular, if *LC*=100%, then the second network, *H*, contains a subnet isomorphic to *G*.

3. Results and Analysis

3.1. Evaluation of Alignment Algorithm

The TRN of Escherichia coli is one of the most elaborate reconstructions currently available. In order to evaluate our method, we used the evaluation algorithm performed by literature [18, 21] with the TRN of E. coli. We build the TRN of E. coli that were organized in RegulonDB [22]. The resulting TRN involving 1680 nodes and 4150 interactions, some main parameters are listed in Table 3. A TRN model represents the molecular regulation process by which genes regulate transcription of other genes.

Table 3. Statistics of the transcriptional regulatory networks of E. Coli and S. Cerevisiae

Network name	Regulatory genes	Regulated genes	Links
E. coli	186	1574	4150
S. cerevisiae	157	4284	12853

Imitation the assess method in literature [18]. Four expand networks obtained by random remove the nodes from the network; random remove the links from the network; random add the links to the network; and random transform the direction of links in the network. For each of the expand network, we calculate the *LC* with four percentages of change: 5%, 10%, 15% and 20%. Due to randomness, we run each experiment 30 times and average results over the 30 runs. The results as shown in Figure 2. With these experiments, we demonstrate that alignment algorithm is capable of producing high-quality alignments with *LC* of about 90%. This indicates that our algorithm is capable of discovering alignments with high *LC* if used for the high topological similarity networks.

3.2. Pairwise Alignment of S. Cerevisiae and E. Coli TRNs

We align the TRN of Escherichia coli to the TRN of S. cerevisiae reconstituted with genes interact dates in the YEASTRACT database [23], some main parameters of the TRN of S. cerevisiae are listed in Table 3. Each gene within TRNs of S. cerevisiae and E. coli was annotated with its corresponding functional class according to Monica Riley's MultiFun system

[24], available via the GeneProtEC database [25]. To measure the effectiveness of the alignment algorithm, we performed a functional concordance(*FC*) analysis for genes. Compared the functional classes of each pair of nodes which aligned by our algorithm and test the proportion of pairs which on the same functional class. *FC* was defined as:

$$FC = \frac{Ps}{Pa} \times \frac{1}{Np} \sum \frac{Psc}{Pac} \tag{3}$$

where *Ps* is the number of aligned gene pairs which on the same functional class; *Pa* is the number of all aligned genepairs; *Psc* is the number of the same functional classes contained by a gene pair; *Pac* is the number of all functional contained by a gene pair; *Np* is total of aligned gene pairs. As shown in Figure 4, the *FC* of regulatory genes is calculated in three functional hierarchies. There are 8 modules in root level, 39 modules in middle level and 78 modules in the third level. We find that our algorithm aligns network regions of Escherichia Coli and S. Cerevisiae in which a large percentage of genes perform the same biological function in both species. The result indicates that our algorithm is effective for discovering alignments with similar function.

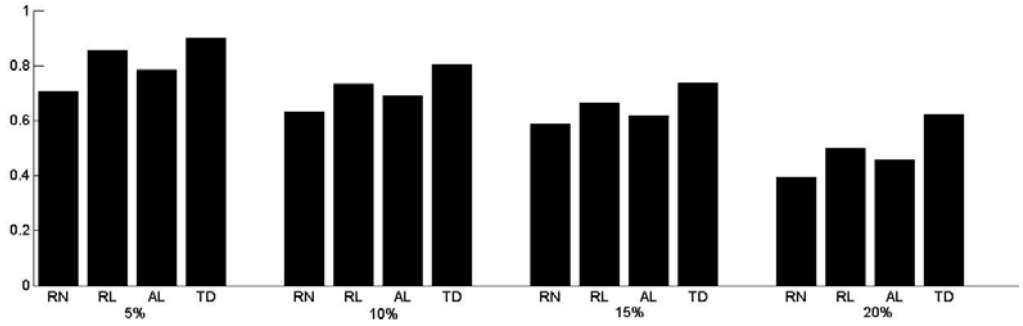

Figure 3. The *LC* of four expand networks with four percentages of change. *RN*: random remove the nodes; *RL*: random remove the links; *AL*: random add the links; *TD*: random transform the direction of links

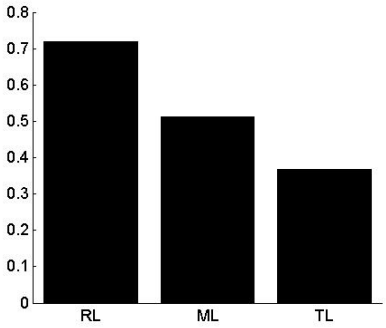

Figure 4. The functional concordance of aligned genes between Escherichia coli and S. cerevisiae TRN viaing our alignment algorithm. *RL* is the root level; *ML* is the middle level and *TL* is the third level

We extracted five functional modules from S. Cerevisiae and E. Coli TRN in accordance with the second level of the MultiFun hierarchy, respectively, as shown in Table 4.

Table 4. Statistics of five functional modules

No	Biological function	E. coli nodes	E. coli links	S. cerevisiae nodes	S. cerevisiae links
C1	Carbohydrates utilization	1408	3675	4156	10235
C2	Fatty acids utilization	1102	2891	3067	8892
C3	Amino acids utilization	1302	3399	3895	9963
C4	Amines utilization	847	2098	2569	7208
C5	Macromolecule degradation	1218	3255	3577	9631

The *LC* of five functional modules calculated by our algorithm shown in Figure 5. The efficiency of align in functional subnets is better than in whole network.

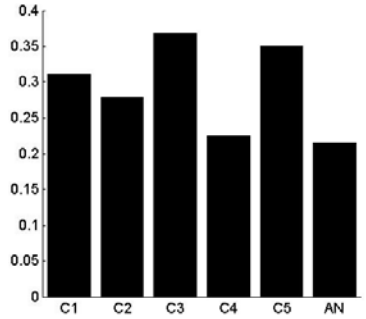

Figure 5. The *LC* of five functional modules. C1, C2, C3, C4, C5 are shown in Table 4; *AN* is the *LC* between two globe networks

4. Conclusion

In summary, we have proved that it is possible to extract function information from network topology. We introduce a new global network alignment algorithm that is based only on network topology structure. Our algorithm can be applied to any network type not just biological networks. We apply our method to align E. coli TRN and noising E. coli TRN. Experiment results demonstrate that alignment algorithm can produce high quality alignments. Additionally, we apply our method to align TRN of S. Cerevisiae and E. Coli. The results demonstrate that it produces topologically statistically significant alignments in which many aligned genes perform the same function. Network alignment has applications across an enormous span of domains, from social networks to software call graphs. In the many domains, the mass of currently available network data will only continue to increase and we believe that high-quality topological alignments can yield new and pivotal insights into function and evolution.

Acknowledgements

This research work is supported partially by the Jilin Province Science and Technology Development projects 10100505,and partially by the Jilin University Graduate innovative research projects 20121101.

References

[1] Newman MEJ. Communities, modules and large-scale structure in networks. *Nature Physics.* 2012; 8(1): 25-31.
[2] Maslov S. Complex networks - Role model for modules. *Nature Physics.* 2007; 3(1): 18-19.
[3] Strogatz SH. Exploring complex networks. *Nature*. 2001; 410(6825): 268-276.
[4] Newman MEJ. The structure and function of complex networks. *Siam Review.* 2003; 45(2): 167-256.
[5] Ciriello G, Mina M, Guzzi PH, et al. Align Nemo: A Local Network Alignment Method to Integrate Homology and Topology. *Plos One.* 2012; 7(6).
[6] Flannick J, Novak A, Balaji SS, et al. Graemlin general and robust alignment of multiple large interaction networks. *Genome Res.* 2006; 16(9): 1169-1181.

[7] Wang BB, Gao L. Seed selection strategy in global network alignment without destroying the entire structures of functional modules. *Proteome science.* 2012; 10.
[8] Pache RA, Ceol A, Aloy P. Net Aligner a network alignment server to compare complexes, pathways and whole interactomes. *Nucleic Acids Research.* 2012; 40(W1): W157-W161.
[9] Pache RA, Aloy P. A novel framework for the comparative analysis of biological networks. *Plos One.* 2012; 7(2): e31220.
[10] Fionda V, Palopoli L. Biological Network Querying Techniques: Analysis and Comparison. *Journal of Computational Biology.* 2011; 18(4): 595-625.
[11] Pagliari R, Yildiz M E, Kirti S, et al. A Simple and Scalable Algorithm for Alignment in Broadcast Networks. *Ieee Journal on Selected Areas in Communications.* 2010; 28(7): 1190-1199.
[12] Berg J, Lassig M. Local graph alignment and motif search in biological networks. *Proc Natl Acad Sci USA.* 2004; 101(41): 14689-14694.
[13] Kelley BP, Yuan B, Lewitter F, et al. PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res.* 2004; 32(Web Server issue): W83-88.
[14] Koyuturk M, Kim Y, Topkara U, et al. Pairwise alignment of protein interaction networks. *J Comput Biol*. 2006; 13(2): 182-199.
[15] Berg J, Lassig M. *Cross-species analysis of biological networks by Bayesian alignment*. Proc Natl Acad Sci U S A. 2006; 103(29): 10967-10972.
[16] Singh R, Xu J, Berger B. Pair wise global alignment of protein interaction networks by matching neighborhood topology. *Research in Computational Molecular Biology.* 2007: 16-31.
[17] Liao CS, Lu K, Baym M, et al. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*. 2009; 25(12): i253-258.
[18] Kuchaiev O, Milenkovic T, Memisevic V, et al. Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society, Interface / the Royal Society*. 2010; 7(50): 1341-1354.
[19] Milenkovic T, Przulj N. Uncovering biological network function via graph let degree signatures. *Cancer informatics*. 2008; 6: 257-273.
[20] Milo R, Shen-Orr S, Itzkovitz S, et al. Network motifs: Simple building blocks of complex networks. *Science*. 2002; 298(5594): 824-827.
[21] Salgado H, Martinez-Flores I, Lopez-Fuentes A, et al. Extracting regulatory networks of Escherichia coli from Regulon DB. *Methods Mol Biol.* 2012; 804: 179-195.
[22] Gama-Castro S, Salgado H, Peralta-Gil M, et al. Regulon DB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Research*. 2011; 39: D98-D105.
[23] Teixeira MC, Monteiro P, Jain P, et al. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Research*. 2006; 34: D446-D451.
[24] Serres MH, Riley M. MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. *Microb Comp Genomics*. 2000; 5(4): 205-222.
[25] Serres M H, Goswami S, Riley M. GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins. *Nucleic Acids Research*. 2004; 32: D300-D302.