Vol.2012-BIO-30 No.7
2012/8/9

Vertex Similarity based on Network Characteristics for Alignment of directed graphs

Hitoshi AFUSO ,†1 Takeo OKAZAKI †2
and Morikazu NAKAMURA†2

For undirected graphs, some similarity measure based on network alignment technique have been proposed. However, they cannot handle directed graph such as gene regulartory networks and it becomes difficult to define the network similarity among them. On the other hand, to capture the feature of vertices in a directed graph, network characteristics had been used in the area of social network analysis. In this paper, we proposed vertex similarity for directed graph based on network characteristics and network alignment method using it. In addition, we compared proposal network alignment method to traditional one, MI-GRAAL using protein-protein interaction(PPI) network in yeast and human. The comparison result showed that our proposed method can find larger subnetwork of yeast PPI network in human's than MI-GRAAL.

1. Background

Improvement of the technique to observe massive gene expressions in a time[1)2)] allows us to obtain various knowledge about biological networks, such as transcriptional regulatory networks(TRNs) and protein-protein interaction(PPI)networks, inside of life-form cells of various species[3)]. By such biological networks, biological function such as metabolism or cellular divison is realized. Understanding the structures of biological network and their corresponding biological function is one of the most important challenges of the post-genomic era. In the past, some researches about the structural property of biological networks, such as network motifs or scale-free property had been done *et al*[4)5)].

On the other hand, comparing the obtained biological networks among different species, it is expected that we may reveal not only the evolutional connection, but also conserved function in life-form cells between them. As major approach for comparison of biological networks, we can see the method so-called network alignment[6)7)]. Network alignment method finds the conserved subnetworks in compared networks. Up to now, various method to utilize the network alignment among biological networks had been proposed. Like as sequence alignment method, network alignment method are divided into two categories, local network alignment and global. As local network alignment methods, PathBLAST[8)] or its modification NetworkBLAST-M[9)] had been proposed to identify the conserved protein complexes in multiple species. In the early date of network alignment, local network alignment methods were considered to be of more value than global while it was believed that conserved subnetwork is small across different species. However, recently report showed that large conservation across the PPI network in yeast and human[10)] and lead to that global nework alignment draws attensions. For major instance of global network alignment, we can cite the method by Oleskii *et al*[10)] and Terada *et al*[11)]. In Oleskii *et al*[10)], authors enumerated the whole small subnetworks included in compared networks, so-colled graphlet degree[12)] to measure the similarity among vertices. Several studies reported the effectiveness of graphlet degree to capture the similarity between the vertices in PPI networks[13)]. However, some studies showed that enumeration of graphlet degree is computationally expensive operation and proposed some probabilistic approximation method[14)]. Terada *et al* proposed the network alignment method based on *"abstract graph"* that represents the rough structure of given PPI network. Although the method by Terada *et al* needs a parameter used to determine the division of given graph to abstract graph in advance, it resulted biologically plausible alignment between PPI networks in nematoda and vinegar fly.

However, these two methods assumed that given biological network is PPI network, modeled in undirected graph. Because of that assumption, it becames difficult to handle the biological networks modeled as directed graph, such as TRNs. Same as the case of PPI networks, comparing the structure of TRNs we may be able to obtain the knowledge about the functional conservation or evolutional relationships. On the other hand, some measures to capture the feature of vertices in directed graph were proposed in the area of social network analysis. Such measures are called as network characteristics and we can see

†1 Graduate School of Engineering and Science in University of the Ryukyus
†2 Faculty of Information Engineering in University of the Ryukyus

clustering coefficient[15] and closeness centrality[17] for an instance.

In this research, we proposed the new methods to utilize network alignment based on vertex similarity calculated with network characteristics. Firstly, we defined the vertex similarity using network characteritics. And next, new network alignment method were proposed based on vertex similarity. To show the effectiveness of our proposed method, we also compared it to traditional network alignment method, MI-GRAAL by Oleskii *et al* via yeast and human PPI network data. Results of the comparison showed that our proposal method can find larger subnetwork between yeast and human PPI networks than MI-GRAAL.

2. Network characteristics

In this section, we showed breif introduction of network characteristics we used. We used six network characteristics, degree, clustering coefficient, closeness centrality, eccentricity centrality, betweenness centrality and PageRank.

Degree denotes that the number of neighbors of particular vertex in a graph. In directed graph, there are two kinds of degree, in-degree and out-degee. In-degree is the number of neighbors that have in-coming edge to focal vertex and out-degree denotes the number of neighbors that have out-going edge from focal vertex respectively. In-degree and out-degree are rough measure of how focal vertex is influeced from other vertices or has effects to others respectively.

Clustering coefficient[15] is the index to measure how many neighbors of particular vertex are connected each other. This index originally had been proposed for undirected graph. Suzuki[16] extended it to the one can that handle directed graph. Same like degree, there are two kinds of clustering coefficient in directed graph. One kind is calculated focusing in-coming edge connectivety and another one is done with out-going edge connectivity. In this paper, we call them as in-coming oriented clusteirng coefficient and out-going oriented clustering coefficient respectively.

In the area of social network analysis, the idea of centrality is used to measure how much each member has "central role" in the social network[17]. There are some variation of centrality according to definition of "central role". In closeness centrality, the vertex which one can reach to from other vertices with smaller steps is considered as central. Traditional closeness centrality $Closeness(v)_G$ of vertex v in graph G was calculated with the length of shortest paths between vertices as follows:

$$Closeness(v)_G = \frac{|V|-1}{\sum_{u \in V} d_G(u,v)} \tag{1}$$

where V denotes the vertex set in graph G and $d_G(u,v)$ is the length of shortest path between vertex u and v in graph G. We assumed that $d_G(u,v)$ becomes infinite when one cannot reach from u to v according to edge direction. Although, using this definition, we cannot handle disconnected graph. To solve this problem, some extensions has been made to closeness centrality. Extended closeness centrality by Opsahl[18] is shown below.

$$Closeness(v)_G = (|V|-1) \sum_{u \in V} \frac{1}{d_G(u,v)} \tag{2}$$

In this research, we used closeness centrality by Opsahl.

In another definition of centrality, one may consider the vertex which one can reach to other vertices with smaller steps as central. According to such view about centrality, eccentricity centrality is defined. Eccentricity centrality is calculated as the maximum length of shortest path. However, using this definition, one cannot handle the disconnected graph. So, in this research we modified eccentricity centrality as follows.

$$Eccentricity(v)_G = \sum_{u \in V} \frac{d_G(v,u)}{|V|-1} \tag{3}$$

In this equation, $d_G(v,u)$ has the same value to the number of whole vertices in G when one cannot reach from vertex v to u.

Between centrality denotes how many times the shortest paths among verteices to the focal vertex. In other words, in betweeness centrality, the vertex that connectes many other vertices directly or indirectly is considered as central. This index is defined as the number of shortest paths that go thorough the focal vertex.

PageRank[19] is the network characteristic proposed by Page *et al* to measure the relative importance among Web pages. The basic concept of PageRank is that Web page that has links from other important Web pages is also considered as important. Using such recursive idea, relative importance of each Web page is calculated. PageRank can be considered as the number of visits by users that walks inside graph according to edge direction at random in enough long time. Vertex that has visit by such random walking users many times has relatively high PageRank value.

Using these network characteristics shown above, we can define the vertex similarity for directed graph.

Now, we have eight network characteritics, in-degree, out-degree, in-coming oriented clustering coefficient, out-going oriented clustering coefficient, closeness centrality, eccentricity centrality, betweenness centrality and PageRank. Using these, we can represent each vertex v in graph G as numerical vector as shown below.

$$\mathbf{f}_G(v) \rightarrow \begin{pmatrix} InDegree_G(v) \\ OutDegree_G(v) \\ InComingClustering_G(v) \\ OutGoingClusteirng_G(v) \\ Closeness_G(v) \\ Eccentricity_G(v) \\ Betweenness_G(v) \\ PageRank_G(v) \end{pmatrix} \tag{4}$$

(5)

Simply say, we represented each vertex v in given graph G as 8-dimensional vector $\mathbf{f}_G$ that each element contains the value of corresponding network characteristics.

Next, using this vector representatin of each vertex(5) we defined some vertex similarities.

3. Vertex Similarity based on Network Characteristics

In previous section, we proposed the vector representation $\mathbf{f}_G(v)$ of vertex v in graph G. In this section, we defined the vertex similarity $S(u, v)$ between vertex u and v.

Suppose that two graphs G and H are given, and vertex u and v is the element of G and H, respectively. In such case, we can calculate the vector representation $\mathbf{f}_G(u)$ and $\mathbf{f}_H(v)$ Traditionally, to measure the similarity between vectors, correlations and distances had been used. According to this convention, we defined three vertex similarity measure: $Sim_{pea}(u, v)$, $Sim_{spe}(u, v)$ and $Sim_{euc}(u, v)$ as follows:

$$Sim_{pea} = Corr_{Pearson}(\mathbf{f}_G(u), \mathbf{f}_H(v)) \tag{6}$$
$$Sim_{spe} = Corr_{Spearman}(\mathbf{f}_G(u), \mathbf{f}_H(v)) \tag{7}$$
$$Sim_{euc} = \frac{1}{(1 + Distance_{Euclid}(\mathbf{f}_G(u), \mathbf{f}_G(v)))} \tag{8}$$

where $Corre_{Pearson}$ and $Corr_{Spearson}$ denote Pearson's product-moment correlation and Spearman's rank correlation between respectively. And $Distance_{Euclid}$ is Euclid distance between given vertices.

On the other hand, Oleskii *et al* proposed similarity measure between two vertices. It was called as *"confidence score"*. In this similarity measure, each network characteritics were treated as agents that have individual opinion about the similarity between vertices. And this similarity summarizes up each agents' opinion so as to minimize the difference of each network chracteritics values. Calculation steps of confidence score consists of these steps.

(1) Calculate the difference of each network characteristics between vertices i and j in graph G and H. Arranging these results as (i, j) element in matrix, we can obtain differential matrix $D_X(G, H)$ where X denotes the network characteristics that used to calculate this matrix.

(2) Calculate $conf_X(i, j)$ from differential matrix $D_X(G, H)$. The $conf_X(i, j)$ represents the fraction of elements in the i-th row of difference matrix $D_X(G, H)$ that are strictly greather than $D_X(G, H)_{\{i,j\}}$.

(3) Sum up each difference matrix $D_X(G, H)$ to confidence matrix $Conf(G, H)$. Then, confidence matrix $Conf(G, H)$ is calculated as $Conf(G, H) = \sum_X D_X(G, H)$. In Oleskii *et al*[10], authors said that this confidence score is robust to minor error in individual difference matrix because that index is based on simple majority vote.

Using four similarity measure, Sim_{pea}, Sim_{spe}, Sim_{euc} and $Conf$, we proposed the network alignment method.

4. Network Alignment Method using Proposed Vertex Similarity

In this section, we proposed new network alignment method, called *DiAliNe*

(Digraph Aligner based on Network Characteristics) using vertex similarity defined in previous section.

Some different formulation of the global alignment problem have been proposed by Flannick *et al*[20] Liao *et al*[7] and Zaslavskiy *et al*[21]. Unlike with the sequence alignment, any reasonable formulation of this problem makes it computationally hard. The reason of this is that problems contains *subgraph isomorphism* problem as its subproblem. Given two graphs, subgraph isomorphism asks which one graph is contained as exact subgraph of the other. This problem is known to belong to NP-complete class[22]

We use the standard definition of the global alignment between two networks $G(V_G, E_G)$ and $H(V_H, E_H)$, where $|V_G| < |V_H|$, as a total injective function $f : V_G \rightarrow V_H$[6,21,13]. Function f is called as *total* if all vertices in V_G will be mapped into some vertices in V_H and *injective* if the function doesn't map different vertices in V_G to identical vertex in V_H. Hence, the alignment is global in the sense that each vertex in the smaller digraph is aligned to some vertex in the larger one.

Same to the network alignment method by Oleskii, proposed method DiAliNe is based on the seed-and-extend approach. This approach consists of two steps, selection of seed pair in given graphs and extend the alignment around seed pair. The main algorithms of DiAlNet are shown in Fig.1 and Fig.2. In Fig.1, Graph G raised to power p is defined as $G^p = (V(G), E^p)$, where $E^p = \{(u_1, u_2) : d_G(u_1, u_2) \leq p\}$. This operation is corresponding to the insertion of gaps in graphs like sequence alignment. And in Fig.2, we used the Hungarian algorithm[23] to find the maximal matching between candidates in each graph. Using this algorithm, we can utilize the network alignment even if given graphs are directed.

5. Comparison to MI-GRAAL via Yeast and Human PPI Networks

To show the effectiveness of our proposed method, we compared DiAlNet to traditional network alignment method, MI-GRAAL[10] via yeast and human PPI network data. The reason why we used MI-GRAAL for comparison is that the method is first global network alignment method in the world that revealed large conserved subnetworks across yeast and human PPI networks and their results showed us the importance and potential of global network alignment method.

```
procedure DiAliNe(directed graph G, H, similarity matrix S)
    alignedPairs ← φ
    while There are unaligned vertex in G do
        find maximal similar pair (u, v) from similarity matrix S.
        alignedPairs ← alignedPairs ∪ {(u, v)}
        newAlignedPair ←AlignLocally(u, v, G, H, S)
        alignedPairs ← newAlignedPair
        if There are still unaligned vertices in G then
            raise the graph to next power.
        end if
    end while
    return alignedPairs
end procedure
```

Fig. 1 Main procedure of DiAliNe

Species	# of Vertices	# of Edges	Average Path Length	Diameter
Yeast	2390	16127	4.819	18
Human	9141	41456	4.136	14

Table 1 Summary of PPI nework in Yeast and Human

In this experiments, we used the PPI network data in yeast from Collins *et al*[24] and also the one in human from Radivojac *et al*[25]. The outline of each PPI network were shown in Table.1.

We used two indices to measure the quality of result of network alignment, edge correctness and largest common connected subgraph same as Oleskii *et al.* Edge correctness denotes that how many edges in smaller graph were preserved in larger graph by alignment results. Edge correctness score EC of alignment m is calculated by following formula.

$$EC = \frac{|(u, v) \in E_1 \wedge (m(u), m(v)) \in E_2|}{|E_1|} \times 100\% \qquad (9)$$

where E_1 and E_2 denote the edge set in graph G and H, respectively. Note that the assumption in global network alignment. In global network alignment, all vertices in smaller graph should be mapped to some vertices in larger graph injectively. So, for calculating EC score, denominater of the formula is the edge

```
procedure ALIGNLOCALLY(vertex u0, v0, graph G, H, similarity matrix S)
    newAlignedPairs ← φ
    nextProcessingPairs ← {(u0, v0)}
    finishedFlag ← false
    while !finishedFlag do
        temporalyPairs ← φ
        finishedFlag ← true
        for all (u, v) in nextProcessingPairs do
            neighborsU ← neighbors of u that not aligned yet
            neighborsV ← neighbors of v that not aligned yet
            if neighborsU and neighborsV are not φ then
                newPairs ← FindMaximalMatching(neighborsU, neighborsV)
                temporalPairs ← temporalPairs ∪ {newPairs}
                finishedFlag ← false
            end if
        end for
        nextProcessingPairs ← temporalPairs
        newAlignedPairs ← newAlignedPairs ∪ {temporalPairs}
        if nextProcessingPairs is not φ then
            sort matchings in nextProcessingPairs by their similarity value.
        end if
    end while
    return newAlignedPairs
end procedure
```

Fig. 2 Subroutine of DiAliNe

Method	Edge Correctness	Largest Common Connected Subgraph
Sim_{pea}	20.20 %	79.70 %
Sim_{spe}	16.13 %	79.16 %
Sim_{euc}	17.42 %	80.06 %
$Conf$	13.23 %	66.31 %
MI-GRAAL	18.68 %	76.65 %

Table 2 Results in Comparison considering Gaps.

Method	Edge Correctness	Largest Common Connected Subgraph
Sim_{pea}	20.38 %	79.70 %
Sim_{spe}	16.34 %	79.16 %
Sim_{euc}	17.61 %	80.06 %
$Conf$	14.12 %	66.31 %

Table 3 Results in Comparison not considering Gaps.

set in smaller graph. To measure the topological quality of network alignment, largest common connected subgraph($LCCS$) also had been used in various study. Since it is prefer that large and contiguous subgraph is obtained by network alignment rather than small and disconnected region, greater size of $LCCS$ is desirable in network alignment result.

In this experiment, we compared alignment result with four different similarity measure shown in section.3 in two different situations, taking into account the gap or not. Because it is very difficult problem to determine the timing of gap insertion, then we simply compared two condition. The results of comparison were shown in Table.2 and Table.3. They showed the result in the case with gaps and without gaps respectively. In the Table.3, the result of MI-GRAAL is not shown because we simply compared the value in the Oleskii's paper. Results in Table.2 showed that eccept the case based on confidence score, proposed methods lead the better result than MI-GRAAL in $LCCS$ score. Especially, in the case with similarity measure by Pearson's correlation lead best EC score in compared methods. And also, comparing the results in Table.2 and Table.3, we can see that EC score had little improvement and $LCCS$ scores are same in every case. This showed that, in the proposed method, the insertion of gaps doesn't have much effect for the alignment result. From these results, DiAliNe lead the stable network alignment results for each similarity measure.

References

1) Stanford Microarray Database, http://smd.stanford.edu
2) Biolog Corp, $PhenotypeMicroArrays^{TM}$, http://www.biolog.com/pmTechDesOver.shtml
3) Insuk L, Zhihua L, Edward. M, "An Improved, Bias-Reduced Probabilistic Gene Network of Baker's Yeast, *Saccharomyces cerevisiae*", PLoS One 3:2(10), 2007
4) R.Milo, S.Shen-Orr, S.Itzkovits, N.Kashtan, D.Chklovskii and U.Alon, "Network Motifs: Simple Building Blocks of Complex Networks", Science 25:298(5594) pp.824-827, 2002
5) Raya K, Ernst W, "How Scale-Free Are Biological Networks", Journal of Computational Biology, 13(3), pp.810-818, 2006
6) Singh.R, Xu.J and Berger.B, "Global Alignment of Multiple Protein Interaction Networks", Proc. of Pacific Symposium on Biocomputing, pp.303-314, 2008
7) Lial.C, Lu.K, Baym.M, Singh.R and Berger.B, "IsoRankN: Spectral Methods for Global Alignment of Multiple Protein Networks", Bioinformatics, 25, pp.253-258, 2009
8) Kelly.B *et al*, "PathBLAST: a Tool for Alignment of Protein Interaction Networks", Nucl.Acids Res., 32, pp.83-88, 2004
9) Sharan *et al*, "Conserved Patterns of Protein Interaction in Multiple Species", PNAS, 102(6), pp.1974-1979, 2005
10) Oleksii K, Natasa P, "Integrative Network Alignment Reveals Large Regions of Global Network Similarity in Yeast and Human", ECCB Vol.00, pp.1-7, 2010
11) Aika TERADA, Jun SESE, "Global Network Alignment using Graph Summarization for Comparison Gene Function", SIG-BIO, 24(12), pp.1-7, 2011
12) Natasa Przulj, "Biological network comparison using graphlet degree distribution", ECCB Vol.23, pp.177-183, 2006
13) Kuchaiev.O, Milenkovic.T, Memisevic.V, Hayes.W and Przulj.N, "Topological Network Alignment Uncovers Biological Function and Phylogeny", ournal of the Royal Society Interface, 7, pp.1341-1354, 2010.
14) N.Przulj, D.G.Corneil and I.Jurisica, "Efficient Estimation of Graphlet Frequency Distribution in Protein-Protein Interaction Networks", Bioinformatics, 22(8), pp.974-980, 2006
15) Watts D.J, Strogatz S.H, "Collective dynamics of small world networks", Nature, Vol.393, pp.440-442, 1998
16) Tomoya SUZUKI, "Analysis for Directed and Weighted Complex Networks on the Basis of Information Flow", IPSJ journal Vol.2, pp.70-78, 2009
17) David Eppstein, Joseph Wang, "Fast Approximation of Centrality", Journal of Graph Algorithm and Applications, Vol.8(1), pp.39-45, 2004
18) Dangalchev.C, "Residual Closeness in Networks", Phisica 365, pp.556, 2006
19) Sergey B, Lawrence P, "The anatomy of a large-scale hyper textual Web search engine", Computer Networks and ISDN System, Vol.30, Issues1-7, pp.107-117, 1998
20) Flannick *et al*, "Graemlin: General and Robust Alignment of Multiple Large Interaction Networks", Genome Res, 16, pp.1169-1181, 2006
21) Zaslavskiy.M, Bach.F and Vert.J.P, "Global Alignment of Protein-Protein Interaction Networks by Graph Matching Methods", Bioinformatics, 25, pp.259-267, 2009
22) Cook.S, "The Complexity of Theorem-Proving Procedures", Proc.3rd Ann ACM Symp on Theory of Computing, pp.151-158, 1971
23) Hiroshi OZAKI and Isao SHIRAKAWA, "Theory of Graphs and Networks", CORONA publishing co.ltd, 1973.
24) Collins *et al* "Toward a Comprehensive Atlas of the Physical Interaction of Saccharomycess Cerevisiae", Molecular and Cellular Proteomics, 6, pp.439-450
25) Radivojac *et al*, "An Integrated Approach to Inferring Gene-Disease Associations in Humans", Proteins, 72, pp.1030-1037, 2008
26) Fossum *et al*, "Evolutionarily Conserved Herpesviral Protein Ineteraction Networks", PLoS Pathog, 5, e1000570, 2009