Volume 3, Issue 7, July 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Efficient Method for Finding Conserved Regions in Protein Interactions Network

Smita Mujumdar
M.E Student Computer Science and Engineering,
D.Y.Patil Institute of Engineering and Technology
Pimpri, Pune, India

Ms. Jyoti Rao
Assistant Professor Computer Science and Engineering,
D.Y.Patil Institute of Engineering and Technology
Pimpri,Pune, India

Abstract - The increasing amount of available data on protein-protein interaction (PPI) networks, which considers various species like human, fly, bacteria, yeast, worm etc. across evolutionary tree put forth major challenge to biological scientists in identifying conserved network regions across species. This attracted many researchers especially for comparative analysis of protein networks, predicting the network structure, protein function as well as interaction. The researchers observed that these networks evolved at a modular level and discovery of conserved patterns in these networks became a core area of the study. The existing few proposed algorithms for aligning PPI networks have simplified structures so the main challenge is to present fast, accurate and robust algorithm for multiple network alignment. Algorithms that are able to extract conserved patterns in terms of general graphs are necessary, since conserved interactions are the parts of functional modules and protein complexes. With this motivation, we focused on discovering highly conserved protein interactions of a pair of species. In this paper, we have developed a new algorithm and investigating its efficacy in fast pairwise alignment of multiple protein networks and identification of conserved interactions. Detailed experimental results from an implementation of the proposed framework, we found that our algorithm is able to discover conserved interaction patterns very effectively, both in terms of accuracies and computational cost as compared to existing network alignment tools.

Keywords— Protein-protein interactions, pairwise alignment, Needleman wunsch algorithm, data representation, search methods.

I. INTRODUCTION

In the problem of network alignment we have to find out network regions those are conserved in their interaction patterns as well as their sequence across two or more species. The numerous methods were presented to overcome the problem of generalizing sub-graph isomorphism. One heuristic approach is to create a merged representation of the networks being compared, called a network alignment graph, facilitating the search for conserved subnet works. Protein-protein interactions (PPI) are of central importance for virtually every process in a living cell. Information about these protein interactions can improve our understanding of evolutionary trend of different organisms, diseases and provide the basis for new therapeutic approaches and in the process of drug discovery. The main aim of system biology is to find out how the proteins from the cell are interacting with each other. Biological procedures such as yeast two-hybrid and protein co-immunoprecipitation techniques are routinely employed now a days to generate large-scale protein-protein interaction networks for human and most model species [6]. As in other biological domains, a comparative study provides a powerful basis for addressing this challenge, calling for algorithms for protein network alignment. Recently there are many algorithms proposed by various researchers in order to overcome the network alignment problem. But, its extension to more than three networks proved to be difficult due to the exponential growth of the alignment graph with the number of species. Hence, this becomes a challenge to researchers and prompted them to work over network alignment problem in case of multiple protein networks [2]-[10]. Here in this paper, we are presenting and investigating a new approach for pair wise alignment of multiple protein networks, which is accurate and fast. This approach basically depends on the novel representation of the network data.

A. Network Alignments

PPI data present a valuable resource for this task. Comparative analysis is used to tackle these problems, and improve the accuracy of the predictions. A fundamental problem in molecular biology is the identification of cellular machinery that are, protein pathways and complexes. But there is a considerable challenge to interpret it due to the high noise levels in the data and the fact that no good models are available to compare pathways and complexes. Main paradigm behind comparison of PPI networks is that evolutionary conservation implies functional significance. Conservation of protein sub networks measures both in terms of protein sequence similarity, and in terms of similarity interaction topology [7]. Some basic notations that appear in many previous works that find conserved pathways or complexes in the PPI networks of different organisms describes in this section.

© 2013, IJARCSSE All Rights Reserved

A PPI network is conveniently modeled by an undirected graph G (V;E), where V denotes the set of proteins, & (u; v) E denotes an interaction between proteins u belongs to V and v belongs to V

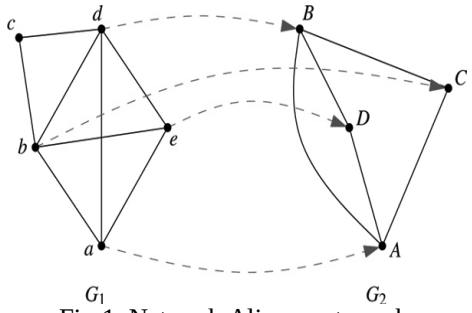

Fig 1. Network Alignment graph

Fig. 1 A dashed arrow from a node i Є V1 from the first network G1 = (V1, E1) to a node j Є V2 from the second network G2 = (V2, E2) indicates that a(i) = j. Unmapped vertices are mapped to gaps.

Given two networks G1 = (V1, E1) and G2 = (V2, E2), a network alignment a: V1 → V2 U {-} maps a vertex i Є V1 to

$$a(i)=\begin{cases} j \in V_2 & a\ vertex\ j\ in\ the\ second\ network \\ - & a\ gap. \end{cases}$$

Note that, network alignments do not have to respect an inherent sequential order of the objects to align [14].

The network alignment problem: Given k different PPI networks belonging to different species, to find conserved sub networks within these networks. In order to find these conserved sub networks an alignment graph is built.

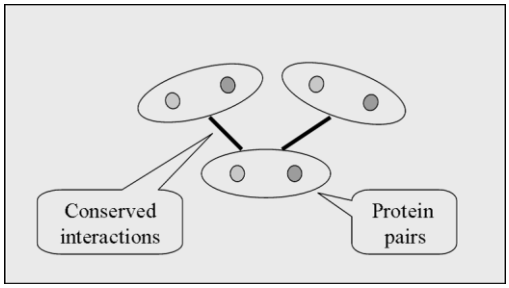

Fig. 2 Alignment graph of two species.

Nodes are constructed of pairs of proteins, one per species, which present a high level of sequence-similarity. Edges represent interactions between proteins in the original networks, which are conserved, meaning they exist in a high level of confidence in both original networks. A heuristic approach is required here since the problem of finding conserved sub networks in a group of networks is NP-Hard [2], [3], because we can reduce it to sub graph-isomorphism known as NP-Hard. Creating an alignment graph from a set of k original networks is one heuristic that enables us to search in all k PPI networks simultaneously. Other heuristics or approximation methods are applicable as well.

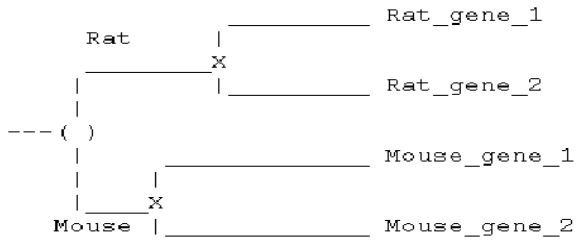

Fig. 3 Gene diversion to a mouse gene and a rat gene after speciation.

Within the mouse and the rat species the gene has been duplicated to two different genes rat gene 1 and rat gene 2 in the rat, and mouse gene 1 mouse gene 2 and in the mouse. Each pair of genes is homologous. Each pair of genes that consists of a rat gene and a mouse gene is orthologous, and each pair that consists of genes in the same species is paralogous.

1) Orthologous proteins: two proteins from different species that diverged after a speciation event. In a speciation event one species evolves into a different species (anagenesis) or one species diverges to become two or more species (cladogenesis).

2) Paralogous proteins: two proteins from the same species that diverged after a duplication event, in which part of the genome is duplicated.

3) Homologous proteins: two proteins that have common ancestry. This is often detected by checking the sequence similarity between these proteins. The proteins can be either from the same species, or from different species (either orthologous or paralogous).

© 2013, IJARCSSE All Rights Reserved

B. Needleman wunsch algorithm

The Needleman-Wunsch algorithm is an application of a best-path strategy (dynamic programming) used to find optimal sequence alignment (Needleman and Wunsch, 1970). Basically, the concept behind the Needleman-Wunsch algorithm stems from the observation that any partial sub-path that tends at a point along the true optimal path must itself be the optimal path leading up to that point. Therefore the optimal path can be determined by incremental extension of the optimal sub-paths. In a Needleman-Wunsch alignment, the optimal path must stretch from beginning to end in both sequences. The score at any position in a global alignment matrix is:

M(i,j) = MAX(Mi-1,j-1 + S(Ai, Bj) or Mi-1, j + gap or Mi,j-1 + gap)

when tracing back the alignment path, horizontal and vertical movement is a gap, and diagonal movement is a match. In order to perform a Needleman-Wunsch alignment, a matrix is created, which allows us to compare the two sequences. The score as determined through use of the above calculation is placed in the corresponding cell. This algorithm performs alignments with a time complexity of O(mn) and a space complexity of O(mn). A dynamic programming algorithm such as Needleman–Wunsch (NW) optimal global alignments considers full complete length for sequence and offers a highly accurate scoring metric but it is slow in comparison to all against all searches.

C. Normalization of alignment score

Most of the ortholog detection methods use different scoring schemes. For example, all-against-all BLAST searches offer high-speed detection of homologous sequence segments. However, BLAST often generates multiple, overlapping hits between a pair of genes. These overlapping hits represent an inaccurate estimate of overall similarity. We first find the most similar gene pairs using BLAST, then generate more accurate similarity scores of those most similar genes by pairwise Needleman Wunsch alignments assuming a BLOSUM62 scoring matrix (Henikoff and Henikoff, 1992) and using linear gap penalty scheme. Standard alignments similarity scores are normalized [13] using:

$$NSS(seq1,seq2) = \frac{NWS(seq1,seq2)/GFAL(seq1,seq2)}{\text{Min}\,(NWS(seq1,seq2)/\text{Min}(GFAL(seq1,seq1),GFAL(seq2,seq2)}$$

where NSS(seq1, seq2) is the normalized similarity score between two sequences; NWS are NW alignment scores and GAL are gap-free alignment lengths where GFAL(seq1, seq1) and GFAL(seq2, seq2) are equal to length of the two sequences. If NSS(seq1, seq2) returns a negative score, it is set to 0. With that, Equation defines globally comparable similarity scores that range from 0 to 1 (1 being most similar).

II. METHODOLOGY

A. Algorithm

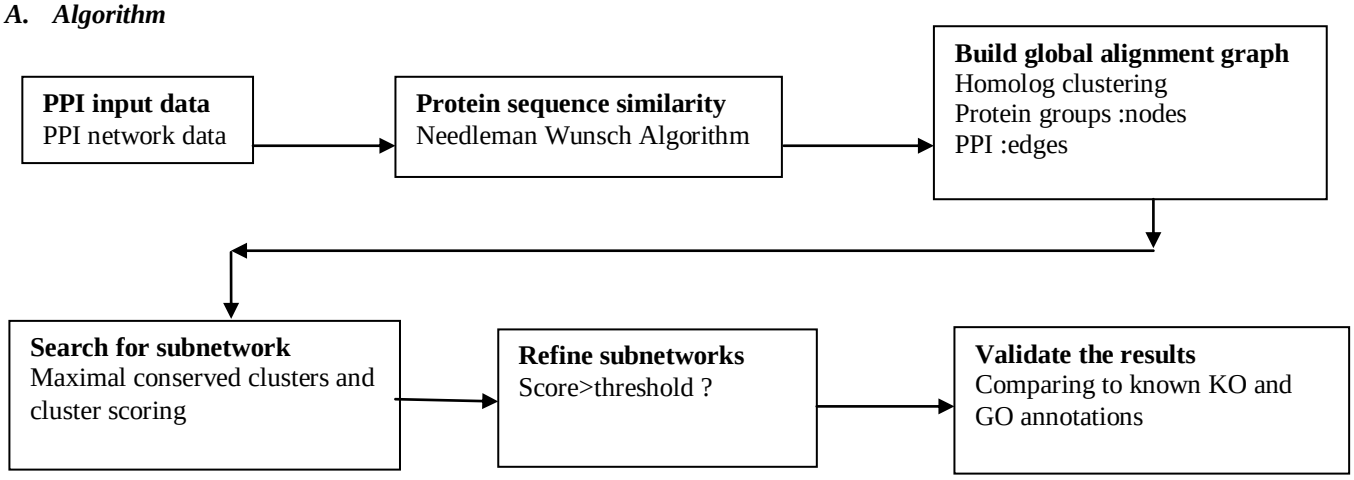

Fig. 4 The six-step flow of proposed model

Our algorithm, can be described as follows:

1: Retrieve and preprocess the Protein Protein Interaction network data from PPI network databases, such as DIP [1] andfind all protein pair that are interacting with each other in a species.

2: Find highly similar protein sequences across the species. Use homolog clustering to identify homolog groups across different species based on dynamic algorithm for global pairwise alignment. Needleman Wunsch algorithm can be used for this purpose.

3: After identifying homolog groups across the species, a network alignment graph is generated based on these groups. The node in the graph represent sets of proteins, one from each species, in the same ortholog group, and edges represent conserved protein-protein interactions across the compared species. One of the way of adding the edges between two node pairs (a_1,b_1) and (a_2,b_2) is when both (a_1,a_2) and (b_1,b_2) are directly interacting with each other in their corresponding PPI networks.

4: Then identify conserved protein interaction regions in the alignment graph. The major algorithm is based on strongly connected-components (clusters) in the alignment graph. Such graph is a maximal set of vertices in which each vertex is reachable from another. We have used the Depth-first-search algorithm to find the strongly connected components.

© 2013, IJARCSSE All Rights Reserved

Strongly connected components algorithm

DFS(G): V[G] is the set of all vertices in G
(1) For each vertex u in V[G]
(2) Do color [u]:=white //unvisited
(3) Pi[u]:=null; Time:=0
(4) For each vertex u in V[G]
(5) Do if color[u]==white
(6) then DFS-Visit(u)
DFS-Visit(u)
(7) color[u] :=gray; time:= time+1; d[u]:=time
(8) for each v in Adj[u] // explore u's neighbors
(9) Do if color[v]==white
(10) then {Pi[v]:=u; DFS-Visit(v)}
(11) color[u]:=black; f[u]:=time<-time+1
Strongly-connected-components (G)
(12) call DFS(G) to compute f[u] for each vertex u; compute G^T, transpose of G
(13) call DFS(G^T), consider the vertices in decreasing order of f[u] in the main loop of DFS
(14) output the vertices of each tree in DFS forest formed

5. Score the identified clusters. For each strongly connected component i.e. each cluster, *C*, the scoring function combines sequence similarity score, interaction conservation, and functional similarity[11]. The scoring function of a cluster *C* is defined as:

$$\text{Score (cluster C)} = S(C) + I(C) + F(C) \quad (1)$$

where *S(C)* is the normalized sequence similarity score of homolog nodes in the cluster *C*, *I(C)* is the interaction conservation score of cluster *C*, and *F(C)* is the functional similarity score of cluster *C*;

Similarity scoring S(C): For each node (one protein from each species) of cluster *C* in the alignment graph, we find its ortholog confidence score. The similarity score of a cluster *C is given as* :

$$S(C) = \left(\sum_{k=1}^{k=|C|} s(k)\right) / |C| \quad (2)$$

where *s(k)* is the similarity score of node *i* in cluster *C*. For simplicity, the similarity score of a node in the alignment graph is set to 1, if it is above threshold value and validated with the proteins ortholog against annotation, such as KO group[4].

Interaction conservation scoring I(C): The conserved interactions are the number of edges connecting all nodes in the identified cluster of the alignment graph. *I(C)* is set to the portion of the direct interactions conserved in the clusters. *I(C)* is formally defined as:

$$I(C) = \frac{i(C)}{|C|(|C|-1)/2} \quad (3)$$

Where I (C) is total number of conserved interactions and |C|(|C|-1)/2 is a cliquishness measure as defined in graph theory.

Functional similarity scoring F(C): We have used the intersection over the union of the number of GO biological process[8][12] terms covered in a cluster of a local species as *F(C):*

$$F(C) = \frac{\text{Intersection of GO biological process terms in Cluster C}}{\text{Union of GO biological process terms in Cluster C}} \quad (4)$$

All the functions are normalized functions with values from the [0, 1] interval so that it is convenient to compare the scores across different clusters.

6. Validation against KO group [4] and GO biological process [12]: As we have validated our orthologs with KO annotations. If ortholog have similarity score as 1 then its tested that they have same KEGG ortholog annotation. We have calculated F (C) functional similarity score of each cluster in local species based on Gene Ontology biological process (GO terms). If the identified cluster has the score less than the threshold (0.5) then we have used GO terms as a measure to keep or remove the cluster. If F (C) is zero then we have removed cluster else we keep it.

To the best of our knowledge this is the fastest algorithm to identify the maximally conserved patterns across the species.

B. Implementation and data collection

This tool is implemented using Java 1.6.0 in a package. It will work under most operating systems running a Java Virtual Machine. Basic input to tool is a dataset containing protein protein interactions (PPI) downloaded from DIP (Database of Interacting Proteins) [1] http://dip.doe-mbi.ucla.edu/dip/Main.cgi site, for the human and mouse species. We have found out for each species the protein pairs, which are interacting based on which a graph is generated having nodes as proteins and edges as interactions between proteins. We found highly similar protein sequences across the species. We have used homolog clustering to identify homolog groups across different species based on Needleman Wunsch dynamic global alignment algorithm and normalized scores are considered to announce orthologs present between two species. Protein/ peptide sequences considered in global alignment algorithm of species Homo sapiens (human), Mus musculus (mouse) were downloaded from http://www.genome.jp/kegg/ and www.uniprot.org

© 2013, IJARCSSE All Rights Reserved

III. RESULTS

We have analyzed a part of PPI network data from Database of Interacting Proteins for two species *Homo sapien* and mouse. Protein interactions are represented as a graphical structure.

TABLE I
The PPI networks analysed in this paper

Species (short name)	# proteins	#PPIs	Source
Homo sapien(*hsa*)	154	172	DIP[1]
Muscus (*mmu*)	147	173	DIP[1]

Since this tool uses KO groups for validation of homolog clustering, it has higher sensitivity and specificity than the other tools.

TABLE II
Sensitivity results for hsa /mmu

Species	#Conserved regions	#unique proteins
Homo sapien(*hsa*)	6	24
Muscus(*Mmu*)	6	22

We use the PPI network data from DIP for two organism pairs *hsa / mmu.* The total number of conserved regions only included larger than size two for *hsa / mmu*. The values for cluster size in this tool are 2 to 6 nodes. The probability to consider the identified cluster is set to threshold (0.5) as well when the functional similarity is zero then we have removed cluster. In the results of this tool, 6 clusters with the node size larger than one were obtained from material of DIP database and applying dynamic pairwise alignment algorithm. If we increase the number of input interactions and when we consider other pairs of species we are getting very efficient results as compared with other existing network alignment tools such as NETWORKBLAST, Gramelin etc.

IV CONCLUSION

As we stated above, during this paper we have presented the new approach for network alignment by considering the multiple networks alignment problem. This approach is based on the novel representation not only single but also the multiple protein-protein interaction networks as well as the orthology relations between their proteins. A dynamic programming algorithm such as Needleman–Wunsch (NW) optimal global alignments offers a highly accurate scoring metric. Based on genome similarity across different species, interaction conservations and functional similarity, we developed a pairwise network alignment tool, to improve the speed, accuracy and generality of the alignment. This tool is fast; it is linear in terms of the number of nodes and edges in the alignment graph. This tool has higher sensitivity as we are using KEGG ortholog annotations for higher accuracy, than the other tools. This methods claims efficient and faster network alignment of multiple protein-protein networks. We have implemented the investigated approach here and showed its effectiveness through the extensive performance evaluation of proposed and existing cases.

REFERENCES

[1] DIP, http://dip.doe-mbi.ucla.edu/.
[2] Sharan, R., S. Suthram, R.M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R.M. Karp, and T. Ideker, Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A, 2005. 102(6): p. 1974-9.
[3] Singh, R., J. Xu, and B. Berger, Global alignment of multiple protein interaction networks. Pac Symp Biocomput, 2008: p. 303-14.
[4] Kanehisa, M. and S. Goto, KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000. 28(1): p. 27-30.
[5] Remm, M., C.E. Storm, and E.L. Sonnhammer, Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol, 2001. 314(5): p. 1041-52.
[6] Kelley, B.P., R. Sharan, R.M. Karp, T. Sittler, D.E. Root, B.R. Stockwell, and T. Ideker, *Conserved pathways within bacteria and yeast as revealed by global protein network alignment.* Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11394-9.
[7] Koyuturk, M., Y. Kim, U. Topkara, S. Subramaniam, W. Szpankowski, and A. Grama, *Pairwise alignment of protein interaction networks.* J Comput Biol, 2006. **13**(2): p. 182-99.
[8] Ashburner, M., 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, *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
[9] Sharan, R. and T. Ideker, *Modeling cellular machinery through biological network comparison.* Nat Biotechnol, 2006. **24**(4): p. 427-33.
[10] Kalaev, M., V. Bafna, and R. Sharan. *Fast and Accurate Alignment of Multiple Protein Networks*. in *RECOMB*. 2008.

© 2013, IJARCSSE All Rights Reserved

[11] Wenhong Tian, Nagiza F.Samatova, "Pairwise Alignment of Interaction Networks By Fast Identification of Maximal conserved patterns" Pacific Symposium on Biocomputing 14:99-110 (2009).
[12] http://www.geneontology.org/
[13] Onur Sakarya Kenneth S. Kosik and Todd H. Oakley "Reconstructing ancestral genome content based on symmetrical best alignments and Dollo parsimony" Bioinformatics Vol. 24 no. 5 2008, pages 606–612
[14] Gunnar W Klau "A new graph-based method for pairwise global network alignment" BMC Bioinformatics. 2009; 10(Suppl 1): S59.

© 2013, IJARCSSE All Rights Reserved