Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Alignment of Protein Interaction Networks by Integer Quadratic Programming

Zhenping Li[1,2] Yong Wang[3] Shihua Zhang[2] Xiang-Sun Zhang[2*] Luonan Chen[3,4,5*]
[1]Beijing Wuzi University, Beijing 101149,China
[2]Academy of Mathematics and Systems Science, CAS, Beijing 100080, China
[3]Osaka Sangyo University, Osaka 574-8530, Japan
[4]Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan
[5]Institute of Systems Biology, Shanghai University, Shanghai 200444, China

***Abstract*— With more and more data on protein-protein interaction(PPI) network available, the discovery of conserved patterns in these networks becomes an increasingly important problem. In this paper, to find the conserved substructures, we develop an efficient algorithm for aligning PPI networks based on both the protein sequence similarity and the network architecture similarity, by using integer quadratic programming (IQP). Such an IQP can be relaxed into the corresponding quadratic programming (QP) which in the case of biological data sets almost always ensures the integer solution. Therefore, a QP algorithm can be adopted to efficiently solve this IQP without any approximation, thereby making PPI network alignment tractable. From the viewpoint of graph theory, the proposed method can identify similar subsets between two graphs, which allow gaps for nodes and edges.**

Keywords: systems biology, protein-protein interaction (PPI), quadratic programming, network alignment

I. INTRODUCTION

One of major challenges for post-genomic biology is to understand how genes, proteins and small molecules interact to form a functional network [1], [2], [3], [4]. In recent years, with rapid progress of biological science, several high-throughput technologies have been developed for studying interactions of molecules, such as the two-hybrid assay, co-immunoprecipitation and the chIP-chip approach, which can also be used to screen for protein-protein interaction (PPI)[5]. So far, these technologies have been adopted to derive PPI networks for some model species [5], such as bacteria, yeast, nematode worm and fruit fly.

PPI networks orchestrate the complex functions of the living cells. Various organisms differ not only because of differences of constituting proteins, but also because of architectures of the PPI networks. Hence, it is essential to address the similarities and differences in the PPI networks by comparative network analysis, which can directly be applied for analyzing signal pathways, finding conserved regions, discovering new biological functions or understanding the evolution of protein interactions. By now there are several network alignment algorithms, which are either mainly based on sequence similarities, such as PathBLAST [6] and Local graph alignment algorithm [7] or mainly based on network architecture similarities, such as pairwise local alignment algorithm [8] and heuristic graph comparison algorithm [9]. But most of the conventional approaches either restrict comparative analysis to special structures, such as pathways, or adopt heuristic algorithms to make the computation of the alignment problem tractable.

In this paper, we aim to develop an efficient algorithm for aligning general PPI networks based on both the protein sequence similarity and the network architecture similarity, by using integer quadratic programming (IQP) based on log-probability-like criterion [6] with matching terms corresponding to both nodes (or vertices) and edges. Computation and experiment on the real biological data sets show that such an IQP can be relaxed into the corresponding quadratic programming (QP) and almost always ensures the integer solution. Therefore, a QP algorithm can be adopted to efficiently solve this IQP without any approximation. In terms of computational complexity, the proposed approach makes the computation of PPI network alignment tractable. From the viewpoint of graph theory, the proposed method can identify similar subsets between two graphs, which allow gaps of nodes and edges. The similarity of two proteins (or nodes) is defined by their homology, whereas the similarity of two edges is based on their confidence ratios of interactions. We have implemented the proposed algorithm by Lingo programming which is available upon request from the authors.

II. FORMULATION OF ALIGNMENT MODEL

A PPI network is an unweighted undirected graph $G(V, E)$, where each node v in the node set V represents a protein, each edge (u, v) in the edge set E represents an interaction between nodes $u \in V$ and $v \in V$.

Given two PPI networks $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, where $V_1 = \{v_1^1, \cdots, v_m^1\}$, $V_2 = \{v_1^2, \cdots, v_n^2\}$. The adjacent matrices of G_1 and G_2 are respectively A and B, where

$$a_{ij} = \begin{cases} 1 & \text{if } (v_i^1, v_j^1) \in E_1 \\ 0 & \text{otherwise} \end{cases}$$

$$b_{ij} = \begin{cases} 1 & \text{if } (v_i^2, v_j^2) \in E_2 \\ 0 & \text{otherwise} \end{cases}$$

*Corresponding authors. E-mail address: zxs@amt.ac.cn, chen@elec.osaka-sandai.ac.jp

Instead of binary values, notice that a_{ij} and b_{ij} can be straightforward extended to real numbers between 0 and 1 to represent the confidence ratios of the interactions. In other words, the PPI network can also be formulated as a weighted graph.

For the protein similarities, we define a similarity score to measure the similarities between a pair of proteins based on their sequences. The similarity score is defined as a function $S : V_1 \times V_2 \rightarrow [0, 1]$. For any $v_i^1 \in V_1$ and $v_j^2 \in V_2$, $s_{ij} = S(v_i^1, v_j^2)$ measures the similarity between proteins v_i^1 and v_j^2. Especially, $s_{ij} = 1$ implies that the sequences of protein v_i^1 and v_j^2 are identical, whereas $s_{ij} = 0$ indicates no similarity of the sequences between v_i^1 and v_j^2.

The matching between $v_i^1 \in V_1$ and $v_j^2 \in V_2$ is represented by a binary variable x_{ij},

$$x_{ij} = \begin{cases} 1 & \text{if } v_i^1 \in V_1 \text{ matches } v_j^2 \in V_2 \\ 0 & \text{otherwise} \end{cases}$$

Clearly, depending on $X = \{x_{ij}\}$, there is a different local alignment or local matching between the two networks G_1 and G_2.

Then, similar to the log-probability score [6], the similarity between two PPI networks (G_1 and G_2) with respect to a given X can be defined by the sum score including both node and edge matching scores. Therefore, the alignment of PPI networks can be formulated as the following integer quadratic programming (IQP) by maximizing the similarity score $f(G_1, G_2)$ between networks G_1 and G_2 among all feasible combinations X.

$$\max_X \quad f(G_1, G_2) = \lambda \sum_{i=1}^{m} \sum_{j=1}^{n} s_{ij} x_{ij} + (1-\lambda) \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{m} \sum_{l=1}^{n} a_{ik} b_{jl} x_{ij} x_{kl}$$

$$s.t. \begin{cases} \sum_{j=1}^{n} x_{ij} \leq 1 & i = 1, 2, \cdots m \\ \sum_{i=1}^{m} x_{ij} \leq 1 & j = 1, 2, \cdots n \\ x_{ij} = 0, 1 & i = 1, 2, \cdots m; j = 1, 2, \cdots, n \end{cases}$$

where the coefficient λ is a scalar parameter between 0 and 1 to control the balance between node and edge socres. For instance, only the node scores are considered in the alignment for $\lambda = 1$, while only the edge scores are optimized for $\lambda = 0$. Generally, the parameter is $0 < \lambda < 1$, depending on the requirement of alignment. The first constraint implies that one node in G_1 can correspond to at most one node in G_2, while the second constraint means that each node in G_2 can match at most one node in G_1. The last condition is the integer constraint for variable X. Depending on the parameter λ, we can obtain different optimal alignment solutions.

III. MODEL COMPONENTS

In order for the PPI networks to be aligned in a biologically meaningful manner, preparation of the underlying similarity function S and adjacent matrices (A, B) is crucial.

A. Adjacent matrix

A protein-protein interaction (PPI) network is represented as an undirected graph $G(V, E)$, i.e., each node represents a protein and each edge represents existing interaction between two proteins. Let $A = (a_{ij})$ be the adjacent matrix, where $a_{ij} = 1$ if there is an edge between vertices i and j, and $a_{ij} = 0$ otherwise. Several studies have suggested methods for evaluating the reliability of protein interactions [14], [15], by which each protein interaction can be assigned a confidence value ranging from 0 to 1. Hence, the PPI network can also be further represented by a weighted adjacent matrix.

B. Similarity Function

Since proteins with similar function often have similar sequences, we measure the similarity S of proteins based on their sequences, which is assigned to any pair of proteins between two networks.

The similarity score between a pair of proteins can be measured by several methods, e.g. based on the similarity of amino acid sequences or the evolutional relation of the protein pairs. One of them is measured by detecting orthologs and in-paralogs using INPARANOID [10], which is developed for finding disjoint ortholog clusters in two species. Each orthologs cluster discovered is characterized by two main orthologs, one from each species, and possibly several other in-paralogs from both species. The main orthologs are assigned a confidence value between 0 and 1, while the in-paralogs are assigned confidence scores based on their relative similarity to the main ortholog in their own species. The similarity between two proteins u and v is defined as

$$S(u, v) = confidence(u) \times confidence(v).$$

Clearly, this score provides a normalized similarity function that takes values in interval [0,1].

Besides INPARANOID, the similarity score can also be measured by the following formula [11]

$$S(u, v) = log \frac{P_{uv}}{q_u q_v} = log(P_{uv}) - log(q_u q_v)$$

When a common ancestor exists between proteins u and v, the numerator P_{uv} is the probability that u is replaced by v, and the denominator expresses the product of the probabilities of obtaining u and v, respectively, by substitution at random (namely, the probability with which u and v are produced independently). That is, this score expresses the degree to which u and v relate evolutionarily in terms of a log-odds ratio.

Moreover, the similarity can be measured from the information of the sequence similarity, e.g., by BLAST [6].

IV. SIMULATION

We have implemented the proposed algorithm by Lingo and all simulations were performed on a PC. Although we mainly describe the proposed method for undirected

networks where all elements of adjacent matrices are positive, it can be used for the directed network, such as gene regulatory networks or metabolic networks, where the elements of adjacent matrices may take either positive or negative numbers depending on the directions of interactions. In the following section, two small examples are used to test the model.

A. Example 1: Aligning Undirected Networks

An example is taken from the tutorial files provided in the PathBLAST plugin of software Cytoscape 1.1. (`http://www.cytoscape.org/plugins1.php`) as shown in Fig 1. The adjacent matrices of the two networks and their similarity matrix S can be seen in our web (http://zhangroup.aporc.org/bioinfo/PPINA/). The results of PathBLAST are obtained by software Cytoscape 1.1 with the same data.

The comparison results firstly show that our method is consistent with PathBLAST in the ability of finding conserved pathway in the two networks. Using our method, we obtained an optimal solution with optimal objective function 4.934 for $\lambda = 0.5$. The protein matching is as follows: $A \leftrightarrow QQ$, $B \leftrightarrow CC$, $C \leftrightarrow BB$, $D \leftrightarrow JJ$ $E \leftrightarrow ZZ$, $F \leftrightarrow HH$, $G \leftrightarrow NN$, $H \leftrightarrow AA$, $I \leftrightarrow DD$, $J \leftrightarrow WW$, $K \leftrightarrow MM$, $L \leftrightarrow OO$. Analyzing this result, we find three corresponding pathways with length 3 as $C|QQ \leftrightarrow A|BB \leftrightarrow F|HH$, $J|WW \leftrightarrow I|DD \leftrightarrow L|OO$ and $H|AA \leftrightarrow G|NN \leftrightarrow B|CC$. These results are consistent very well with the results obtained by PathBLAST. By looking deep into the results of pathways with length 3 in global alignment graph by PathBLAST, surprisingly the pathway $C|QQ \leftrightarrow A|BB \leftrightarrow F|HH$ has the highest probability score, and $J|WW \leftrightarrow I|DD \leftrightarrow L|OO$ is ranked in the second place by probability score. But the pathway $H|AA \leftrightarrow G|NN \leftrightarrow B|CC$ found by us is not reported in the PathBLAST results. Similar results appear when we emphasize on the edge information by decreasing parameter λ.

Another interesting point of the comparison results is that our network alignment method can find some substructure which is not shown in PathBLAST results. For example if we further increase the weight of edge information in network alignment, the optimal objective function will be 1.684 when $\lambda = 0$. The protein matching is as follows: $A \leftrightarrow BB$, $B \leftrightarrow QQ$, $C \leftrightarrow HH$, $D \leftrightarrow OO$, $E \leftrightarrow DD$, $G \leftrightarrow CC$, $H \leftrightarrow ZZ$, $I \leftrightarrow AA$, $J \leftrightarrow JJ$, $K \leftrightarrow MM$, $L \leftrightarrow NN$. By analyzing the results we find that both the proposed method and PathBLAST pick out the conserved pathway $C|HH \leftrightarrow A|BB \leftrightarrow B|QQ \leftrightarrow I|AA \leftrightarrow L|NN$. But our method also find the substructure formed by the nodes $[C|HH, A|BB, B|QQ, G|CC, H|ZZ, I|AA, L|NN]$, which is not in pathway format so not shown in PathBALST results.

The parameter λ in our algorithm aims to balance the node similarity and the edge similarity in two networks. By adjusting this parameter, comparison results show different emphasis in network alignment.

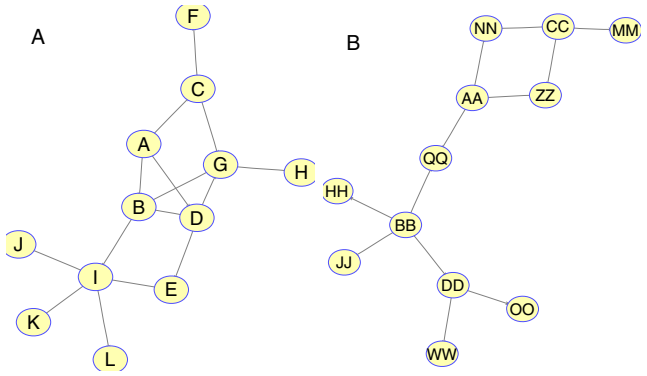

Fig. 1. An tutorial network alignment example from PathBLAST plug-in of Cytoscape software.

B. Example 2: Aligning Directed Networks

Our algorithm can also be used to the comparison of directed networks, it guarantees the feasibility of application in comparing directed biological networks such as the gene regulatory networks. For the directed networks, we have the similar results. An example of two directed networks is shown in Fig 2, where the related information including adjacent matrices and their similarity matrix can be found in our web (http://zhangroup.aporc.org/bioinfo/PPINA/). We found the optimal objective function value 8.36 with parameters $\lambda = 0.4$ using the QP model, where the optimal matching is as follows: (u_1, v_1),(u_2, v_2),(u_3, v_3),(u_4, v_4),(u_5, v_5),(u_6, v_6),(u_7, v_7), (u_8, v_{11}),(u_9, v_{10}),(u_{10}, v_9),(u_{11}, v_8),(u_{12}, v_{12}).

Furthermore, we can obtain the same corresponding result in other parameter, such as $\lambda = 0.6$, $\lambda = 0.7$.

V. CONCLUSION AND FUTURE WORKS

A. Conclusion

In contrast to PathBLAST which focuses on the search of pathways without a loop, our approach can handle a general network alignment problem, and can also find the subnetworks without any restriction. To find the conserved substructures or evaluate the similarity between two PPI networks, we developed an efficient algorithm for aligning PPI networks based on quadratic programming (QP), which allow gaps for nodes and edges. Depending on the parameter λ which balances the node and edge matching scores, the optimal result will be different. A large λ emphasizes on the node matching score, the aligned substructures generally have fewer edges but more nodes with homologous proteins. On the other hand, a small λ emphasizes on the edge matching score, and the aligned substructures generally have more edges and are also larger in size. By selecting all the nodes without gaps and constructing a minimum connected subgraph in each network, the two minimum subgraphs can be regarded as conserved patterns.

B. Future Works

The IQP model is solved by the relaxed QP model, there is a need to analysis the reasonability although the

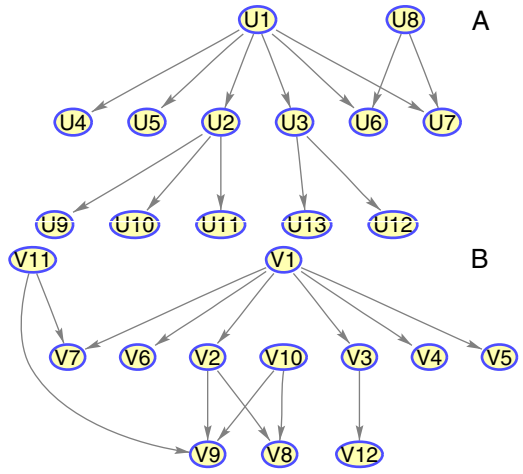

Fig. 2. The simulated example of two directed networks

computation on the real biological data sets almost always gives integer solution. Furthermore, an IQP (even the relaxed QP model) model is still intractable for large PPI networks with thousand proteins. So finding a simplified strategy to efficiently solve such a large-scale IQP is just in process. We will adopt decomposition technique, which decomposes the whole PPI networks into small overlapping subnetworks so that the proposed model can be used on these subnetwork pairs and further detect conserved patterns between two large PPI networks.

VI. ACKNOWLEDGMENTS

This work is partly supported by Important Research Direction Project of CAS "Some Important Problem in Bioinformatics", National Natural Science Foundation of China under Grant No.10471141, and K.G.Wang Education Foundation Hong Kong.

REFERENCES

[1] Wang, R., Jing, Z., and Chen, L. Modelling periodic oscillation in gene regulatory networks by cyclic feedback systems. *Bulletin of Mathematical Biology*, 67: 339-367, 2005.

[2] Wang, R., and Chen, L. Synchronizing genetic oscillators by signalling molecules. *Journal of Biological Rhythms*, 20: 257-269, 2005.

[3] Chen, L., Wang, R., Kobayashi, T., and Aihara, K. Dynamics of gene regulatory networks with cell division cycle. *Physical Review E*, 70: 011909, 2004.

[4] Chen, L., Wang, R., Zhou, T., and Aihara, K. Noise-induced cooperative behavior in a multi-cell system. *Bioinformatics*, 21: 2722-2729, 2005.

[5] Kelley, B.P., Sharan, R., Karp, R., Sittler, E.T., Root, D.E., Stockwell, B.R., and Ideker, T. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl Acad. Sci. USA.* 100:11394-11399, 2003.

[6] Kelley,P.B., Yuan,B., Lewitter,F., Sharan,R., Stockwell,B.R., and Ideker,T. PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Research* 32:83-88, 2004.

[7] Berg,J., and Lassing,M. Local graph algorithm and motif search in biological networks. *Proc. Natl. Acad. Sci. USA* 101: 14689-14694, 2004.

[8] Koyutürk,M., Grama, A., and Szpankowski, W. Pairwise local alignment of protein interaction network guided by models of evolution. *RECOM 2005, Lecture notes in bioinformatics*, 3500: 48-65, 2005.

[9] Ogata,H., Fujibuchi,W.,Goto,S., and Kanehisa,M., A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Research*, 28(20): 4021-4028, 2000.

[10] Remm M., Storm C. E.V., and Sonnhammer E.L.l., automatic clustering of orthologs and in-paralogs from pairs species comparisons, *J. Mol. Bio.*, 314:1041-1052, 2001.

[11] Durbin,R., Eddy,S.R., Krogh,A., and Mitchison, G., Biological Sequence analyses: Probabilitic Methods of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, 1998.

[12] Chen,L., Wu,L-Y., Wang,Y., Zhang,X-S. Inferring Protein Interactions from Experimental Data by Association Probabilistic Method. *Proteins*, 62: 833-837, 2006.

[13] Ala Trusina, Kim Sneppen, I. B. Dodd, K. E. Shearwin and J. B. Egan., Functional alignment of regulatory networks: A study of temerate phages. *Plos Computational Biology* **1**, 2005.

[14] Bader, J.S., Chaudhuri, A., Rothberg, J.M., Chant, J. Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.*, 22:78-85, 2004.

[15] Deng, M., Sun, F. and Chen, T. Assessment of the reliability of protein-protein interactions and protein function prediction. *Pacific Symposium on Biocomputing(PSB2003)*, 140-151, 2003.

[16] Bandyopadhyay, S., Sharan, R., Ideker, T. Systematic identification of functional orthologs based on protein network comparison *Genome Res.* doi:10.1101/gr.4526003, 2006.