INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 9, Number 2, Pages 326–337

A PARALLEL METHOD FOR QUERYING TARGET SUBNETWORK IN A BIOMOLECULAR NETWORK

JIANG XIE, WU ZHANG, SHIHUA ZHANG, AND TIEQIAO WEN

Abstract. Similarity analysis of biomolecular networks among different species or within one species is an efficient approach to understand evolution or disease. The more data from biological experiment, the larger networks. Sequential computational limitation on single PC or workstation have to be considered when methods are developed. The Immediate Neighbors-in-first Method is a method for querying the subnetwork which is most similar to the target in a biomolecular network. Parallel algorithm for it to treat large-scale networks is developed and the parallel performance is evaluated in this paper. Moreover, we apply the present method to two groups of tests on real biological data including protein interaction networks of Fly and Yeast and metabolic networks of Yeast and E. coli. Several conserved protein interactions and metabolic pathways are found and some new protein interactions and functions are predicted.

Key words. biomolecular network, network querying, parallel computing.

1. Introduction

Since the birth of molecular biology, a great deal of knowledge on biological molecules has been accumulated. With further in-depth research and biotechnology development, investigators pay more and more attention to interactions between molecules and networks constructed by them rather than single molecule. Various biological networks are being constructed, such as protein-protein interaction networks (PIN)[1, 2], gene regulatory networks[3, 4] and metabolic networks[5, 6] etc.. Due to the complexity of life, revealing how genes, proteins and small molecules interact to form functional cellular machinery is a major challenge in systems biology. Studies on those molecular networks provide new opportunities for understanding life science at a system-wide level[7, 8, 9, 10]. It is verified that modular structure exist in biology networks[11, 12, 13]. One of the important problems is how to impersonally and accurately define a functional module, conserved pathway or signal path as well as how to find them from a molecular network.

Network alignment and network querying are typical network comparison methods[14]. Because of evolution of species, we can expect there are some conserved sub-networks in biomolecular networks of different species. Comparison of biomolecular network between species is a promising approach to analyzing signaling pathway, looking for conserved region, discovering new biological function and understanding evolution of species. In recent years, many investigators have contributed themselves to this field and made great progress[15, 16, 17, 18, 19, 20, 21, 22]. A few querying tools have been developed, but searching a sub-network from a large network is a problem of local network comparison, involving large scale

2000 *Mathematics Subject Classification.* 35R35, 49J40, 60G40.

This research is part supported by Shanghai Leading Academic Discipline Project [J50103], Ph.d. Programs Fund of Ministry of Education of China [200802800007], the National Science Foundation of China [31070954], the Science and Technology Commission of Shanghai [09JC1406600], Innovation Fund of Shanghai University and Innovation Program of Shanghai Municipal Education Commission [11YZ03].

computation and belongs to NP hard cluster. The existing network querying tools are still at an early stage and far from perfect.

For instance, the online network comparison provided by PathBlast can only deal with some special cases because of the computational complexity, though the PathBlast family tools[15, 16, 17, 18] can implement network querying. MetaPathwayHunter[19] developed by Pinter et al. is a pathway alignment tool based on the sub-tree homeomorphism model, but the topological structure is limited to tree-like graphs. Other querying tools, such as QPath[20] that has been developed for searching linear pathways, also Netmatch [21] has been developed for one-one matching without gap, and MNAligner[22] has been developed for aligning two molecular networks. But they both have their own limitations. The bottleneck is that biomolecular networks are complex networks and querying a sub-network is computationally demanding.

To meet the demand of computational complexity and deal with large-scale biomolecular networks, an effective way is to adopt parallel computation. In this paper we adopt the Immediate Neighbors-in-first Method (INM) for biomolecular network and propose its parallel computing algorithm, and the performance of parallel computing is demonstrated by Parkinson's Disease related protein interaction network (PIN). The rest of this paper is organized as follows. Section 2 describes the INM for direct or undirected networks. Section 3 proposes the parallel computing algorithm and analyses the computational performance, including the speedup and scalability. In section 4, PIN of Fly and Yeast and metabolic networks of Yeast and E. coli are studied, some conserved protein interactions and metabolic pathways are found and some protein interactions and functions are predicted. Section 5 summarizes this paper and discusses future work.

2. Biomolecular Network Querying

A biomolecular network can be represented as a graph. PIN can be represented as an undirected graph, while metabolic network or gene regulatory network can be represented as a directed graph. Each node in the graph represents a molecule, and each edge represents the relationship between two molecules.

The biomolecular network querying problem that we will study in this paper, aims to discovery sub-networks that are identical or most similar to the target within or cross species in the biological sense. The characteristics of the proposed method is that it bases on attributes (such as sequences or function) of molecules themselves and increases the chance that two molecules will be matched if their neighbors have been matched. We call the algorithm Immediate Neighbors-in-first Method(INM). The INM for querying sub-networks from graph G_0 is divided in four phases here. Given the target sub-network G_t, in the first phase, the similarity score between every pair of nodes (a, b) where $a \in G_t$ and $b \in G_0$ is initialized. In the second phase, the score is updated by an iterative process. In the third phase, with immediate neighbors-in-first, the result sub-network G_s that similar to the G_t is obtained from G_0. Finally, a similarity score between G_s and G_t is computed by summing the similarity of the matched nodes and by the similarity of the edges.

2.1. Initialize the similarity scores of molecules.

Let $G_0 = (V_1, E_1)$ (undirected graph) or $G_0 = (V_1, E_1, \lambda)$ (directed graph), where $|V_1| = n_1$, and $G_t = (V_2, E_2)$ (undirected graph) or $G_t = (V_2, E_2, \lambda)$ (directed graph), where $|V_2| = n_2$. G_0 and G_t are represented by their adjacency matrix $A_1(n_1 \times n_1)$ and $A_2(n_2 \times n_2)$. $A_{n_1 \times n_2}$ is the similarity matrix S, where the entry $S(a, b)$ indicates the similarity coefficient between the node $a \in G_0$ and node $b \in G_t$. The initial value of $S(a,b)$ is $Sim(a,b)$.

In the case of the metabolic graphs, the similarity between enzymes can be defined as Tohsato[23] or Pawlowski[24] did. According to the similarity between Enzyme Classification (EC) number of the corresponding reactions, we compute the initial similarity between enzyme $a \in G_0$ and $b \in G_t$ by following formulation: $Sim(a,b) = 0.25 \times r(e_a, e_b)$ Where r indicates the number of uninterrupted and unchanged EC number. For example, $r([1.2.3.4],[1.2.3.5]) = 3$, $r([1.2.3.4],[2.1.3.4]) = 0$.

If it is the PIN or gene regulatory network, we compute the initial similarity between molecule $a \in G_0$ and $b \in G_t$ based on their sequences. E-value is computed by BLAST[25], converted into number between 0 and 1, and treated as initial value.

Regardless of the manner in which the initial value of $S(a,b)$ is obtained, $Sim(a,b)$ expresses relationship of the function or sequence of molecule a and b.

2.2. Computation of similar scores between molecules. Biomolecular networks are different from each other not only because of differences in their components, but also in their network architectures. So their similarity includes two aspects: one is nodes similarity, which means the similarity of their function or sequence, and the other is edges similarity, which means the network topological structure similarity. To take into account both of the two aspects, the topological information of network and the initial values should be put together.

Similarity of network topological structure can be described as $A1$-$A4$ and $D1$-$D4$ proposed by [26][27] for general network querying. Considering incorrectness and incompleteness of experiment data[28, 29, 30, 31], the INM describes similarity of network topological structure as terms $A1$-$A4$. Term $A1(a,b)$ represents the average similarity between the in-neighbors of a (nodes from which a has incoming edges) and the in-neighbors of b, Term $A2(a,b)$ represents the average similarity between the out-neighbors of a (nodes to which a has outgoing edges) and the out-neighbors of b, Term $A3(a,b)$ is similar to $A1(a,b)$ and represents the average similarity between the non-in-neighbors of a (nodes from which a has no incoming edges) and the non-in-neighbors of b, Term $A4(a,b)$ represents the average similarity between the non-out-neighbors of a (nodes to which a has no outgoing edges) and the non-out-neighbors of b, their mathematical definition can be found in [27]. Then the INM computes the similar coefficient in matrix S as follows:

Initialization

$$S^0(a,b) = Sim(a,b) \tag{1}$$

Iteration

A. for directed graph

$$S^{(k+1)}(a,b) = \frac{A_1^k(a,b)+A_2^k(a,b)+A_3^k(a,b)+A_4^k(a,b)}{2} \times Sim(a,b) \tag{2}$$

B. for undirected graph

$$S^{(k+1)}(a,b) = \frac{N_1^k(a,b)+N_2^k(a,b)}{2} \times Sim(a,b) \tag{3}$$

in which

$$N_1^{(k)}(a,b) = \begin{cases} \displaystyle\sum_{a_2 \leftrightarrow a, b_2 \leftrightarrow b} \frac{S^k(a_2,b_2)}{deg(a)deg(b)}, \\ \qquad if\ deg(a) \neq 0\ and\ deg(b) \neq 0 \\ \displaystyle\sum_{a_2 \in G_1, b_2 \in G_2} \frac{S^k(a_2,b_2)}{n_1 \times n_2}, \\ \qquad if\ deg(a) = deg(b) = 0 \\ 0, \quad otherwise \end{cases} \tag{4}$$

$$N_2^{(k)}(a,b) = \begin{cases} \displaystyle\sum_{a_2 \nleftrightarrow a, b_2 \nleftrightarrow b} \frac{S^k(a_2,b_2)}{(n_1 - deg(a)) \cdot (n_2 - deg(b))}, \\ \qquad if\ deg(a) \neq n_1\ and\ deg(b) \neq n_2 \\ \displaystyle\sum_{a_2 \in G_1, b_2 \in G_2} \frac{S^k(a_2,b_2)}{n_1 \times n_2}, \\ \qquad if\ (n_1 - deg(a)) = (n_2 - deg(b)) = 0 \\ 0, \quad otherwise \end{cases} \tag{5}$$

Nomorlization

$$S \leftarrow \frac{S}{\parallel S \parallel_2} \tag{6}$$

So the similarity information of network topology is added upon initial value, and the similar coefficient that involves both function and topology information of molecules is obtained by iteration.

2.3. Querying by Immediate Neighbors-in-first. After building the similarity matrix S, we are ready to start implementation of network querying. Molecules in same functional module often have similar function, take part in one molecular process, or form one signaling pathway etc., which implies that molecules that have relationship with similar molecules are probably similar to each other. In the process of querying, if the similarity of a pair of molecules with similar neighbors is invariable, the relationship between molecules cannot be captured well.

So based on NBM[32], the INM is immediate neighbors-in-first, which creases similarity of their neighbors while two nodes have been matched. According to the similarity matrix S of G_0 and G_t, every molecule in G_0 which is the most similar to molecule in G_t is found, put into queue Q in descending order of similar coefficient. In each iteration, the most matched molecule pair (a, a') that had not been matched in Q is selected, the similar coefficient of molecules that are neighbors of (a, a') and had not been matched are increased. As a result, chance that these neighboring molecules can be matched with each other will be increased in next iteration. The iteration will not finish until all molecules in A matched. Then the matched nodes and the edges between them in G_0 construct the result sub-network G_s [33].

2.4. Computing graph similarity score. As mentioned above, similarity of two biomolecular networks is not only the similarity of their molecules, but also that of the relationship between the molecules. So the similarity score of G_s and G_t is defined as follows [33]:

Let $G_s = (V_s, E_s)$ or $G_s = (V_s, E_s, \lambda_s)$, $G_t = (V_2, E_2)$ or $G_t = (V_2, E_2, \lambda_2)$, $|V_s| = n$, $v_i, v_j \in V_s$, $e_{ij} \subseteq E_s$, $(v_i, v_j) = e_{ij}$, the matched node in the G_t to $\forall v_i \in V_s$ is $\phi(i)$, and the similar coefficient between them is $Sim(i, \phi(i))$, then the scoring of G_s for G_t is:

$$Score(G_s, G_t) = \frac{1}{2}\sum_{i,j=1}^{n} Score(e_{ij}) + \sum_{i=1}^{n} Score(v_i) \tag{7}$$

in which

$$Score(e_{ij}) = \begin{cases} 1, & \exists\, e(\phi(i), \phi(j)), and\, Sim(i, \phi(i)) > 0, \\ & \qquad\qquad Sim(j, \phi(j)) > 0 \\ 0, & otherwise \end{cases}$$

$$Score(v_i) = \begin{cases} Sim(i, \phi(i)), & \sum\limits_{j=1}^{k} Score(e_{ij}) > 0 \\ 0, \quad otherwise & \end{cases}$$

3. Parallel Computing

The INM aims at studying the similarity between biomolecular networks. As bioinformatics progresses, there are a lot of professional databases, such as NCBI[34], HPRD[35], MINT[36], which supply curated data of biomolecular network for researchers. It is reported that these data is doubled every 15 months[37]. So far for some species such as Fly, Yeast and Human, magnitude of proteins that their interactions exist in databases is about 10^3, the interactions is 10^4, and these data is rising continuously. Challenge is emerging with dramatically growth of data resource. In order to meet requirements of computing large scale biomolecular network, we designed the parallel strategy of the INM.

3.1. Method. The parallel computing environment is the cluster of workstation (COW). 14 IBM HS21 blade servers and 2 x3650 servers are the computing and management nodes, each node is equipped two dual-cores CPU and 4GB memory, and connected to each other by 1KM Ethernet and 2.5G infiniBand. The storage is distributed and memory shared. The operation system of this COW is Linux, programming environment is Message Passing Interface (MPI), and the language is C/C++.

As described above, the INM is mainly composed of two parts: one is initialization of the similarity matrix S of G_t and G_0, the other is network querying. The former is $n_0 \times n_t$ iteration of matrix (where number of nodes is n_0 and n_t in G_0 and G_t respectively), and the computational complexity is $O((n_0 \times n_t)^2)$. The latter is querying the similar nodes and increasing the similarity coefficient of their neighbors, and the computational complexity is $O(n_1 \times n_2)$. So most computational time is cost for the former part, namely the initialization of similarity matrix.

Therefore, partition of parallel tasks for the INM is to decompose the initialization process of the similarity matrix S. Let entry $s_{ij} = S(a, b)$ in S indicates the similarity score between node $a \in G_0$ and $b \in G_t$, where $i = 0...n_0 - 1, j = 0...n_t - 1$, we decompose S by row to partition parallel tasks. The initialization process of S in parallel algorithm is as follows.

1. Send $S_x$initial value is $S_0$to each processor.

2. According to the number of itself, each processor judges which row in S_x will be computed in the local, the rule is: entry $s^x_{ij}(j = 0...n_t - 1)$ in row i is computed by node which number is $(imodk)$, where k is the total number of processors.

3. Each processor computes similarity between $s^x_{ij}(j = 0...n_t - 1)$ that need to process in the local and each entries in S_x, obtains $s^{x+1}_{ij}(j = 0...n_t - 1)$, and sends the result to NO.0 processor.

4. No.0 processor adjudges whether S_{x+1} is convergence, if not, then S_x is replaced by S_{x+1}, and repeat step 1-4; else stop computing.

3.2. Performance Evaluation. To study Parkinson's Disease (PD), our biological research group obtained some differentially expressed proteins in Fly model. Based on the PIN dataset in [31] including 7038 proteins and 20720 interactions, we construct the target sub-network G_t, which includes 60 proteins and 100 interactions of Fly. The G_0 is PIN of human that is obtained from HPRD[35] and involves 6340 proteins and 23591 interactions.

By different number of processor, computing time for querying sub-network G_t in G_0 is shown in Fig. 1.

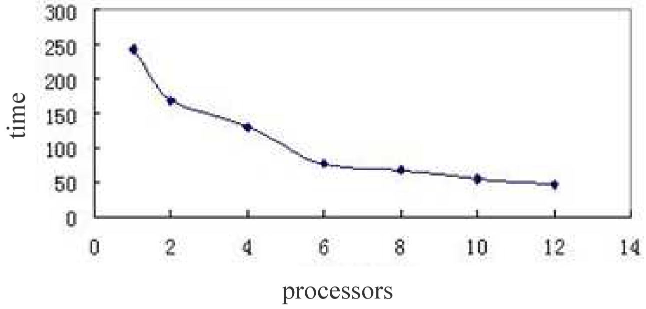

FIGURE 1. Parallel computing time and number of processor. With the increase in the number of processors, the computation time significantly reduced.

As we know, performance of parallel computing is often evaluated by two indicators, one is the speedup, and the other is the scalability. The speedup of this parallel algorithm is shown in Fig. 2, and the scalability is shown in Fig. 3.

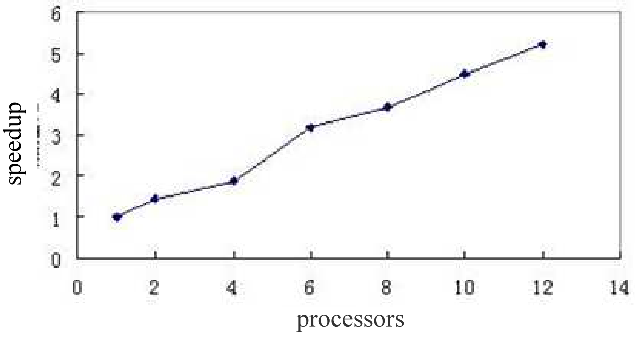

FIGURE 2. Speedup of parallel computing

As shown in Fig. 2, along with the increasing of number of processors, the multiple of speed up is increased, which demonstrates that the speedup of this

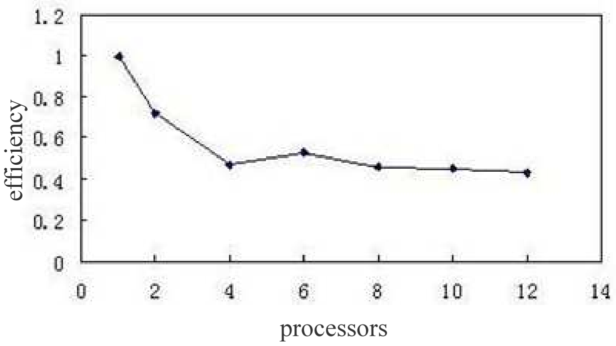

Figure 3. Efficiency of parallel computing

algorithm is good. Fig. 3 describes that there is no significant changes in the efficiency of parallel computing when the number of processor is from 4 to 12. It implies that the additional cost, such as communication, synchronization or idle time, is not improved markedly while processor grows in number. Fig. 1, Fig. 2 and Fig. 3 are indicate that the algorithm has good scalability.

4. Examples

4.1. Undirected Network.

PINs are undirected biomolecular networks. Here we study similar sub-PINs of Fly and Yeast by the INM. The PINs of Fly and Yeast are download from DIP [38] and used by Ideker group [31], which include 4389 proteins and 14319 interactions of Yeast and 7038 proteins and 20720 interactions of Fly. 236 target sub-networks G_t are complex of Yeast that come from MIPS [39]. By querying the 236 targets in PIN of Fly, scoring their relevant result sub-networks and evaluating by T-test, we obtained 34 result sub-networks with statistical significance (p value $< 10^{-22}$), as shown in table 1.

Protein interaction can be predicted by comparison of similar sub-networks in different species [16]. If protein D_a and D_b is of Fly, protein Y_a and Y_b is of Yeast, D_a and Y_a have similar sequence, D_b and Y_b have similar sequence, we predict protein interaction from two aspects:

1) If D_a interacts with D_b and Y_a interacts with Y_b, then the interactions (D_a, D_b) and (Y_a, Y_b) are conserved between Fly and Yeast, and interactions probably exist between the proteins that have similar sequence to the two pairs of proteins in other species. Therefore 19 conserved protein interactions are obtained here, as shown in table 2.

2) If D_a interacts with D_b while Y_a does not interact with Y_b, it implies that Y_a potentially interacts with Y_b, and it is same in the opposite case. So 5 protein interactions are predicted in table 3. By now these 5 protein interactions have not been in DIP [38], BioGrid[40], FlyBase [41] or MINT [36].

Functions of protein can be predicted by comparison of similar sub-networks in different species and Gene Ontology (GO) [16]. In a sub-PIN, if most proteins have same function, then the remainder proteins in this PIN are likely to have the same function. Table 4 lists the function prediction of some proteins, here we only

Table 1. Querying results that have statistical significance

Complex NO.	p-value	Complex NO.	p-value
510	1.47e-22	140	8.218849e-23
140.20	8.218849e-23	510.190	3.771932e-26
440	1.576521e-28	510.190.10	1.576521e-28
510.190.10.20.10	1.576521e-28	500.20	7.171947e-30
500.20.10	7.171947e-030	500-1	1.355895e-31
500.40.10	1.355895e-31	140.30	3.249916e-34
500-2	3.160313e-35	140.20.20	6.510381e-36
360.10	6.510381e-36	440.30	6.510381e-36
510.190.10.10	6.510381e-36	360.10.10	1.309831e-37
260	2.773203e-40	500.10	5.818282e-42
100	2.716906e-43	177	2.716906e-43
180.30	2.716906e-43	260.30	2.716906e-43
260.30.10	2.716906e-43	360.10.20	2.716906e-43
440.30.10	2.716906e-43	445	2.716906e-43
445.10	2.716906e-43	510.50	2.716906e-43
140.30.20	1.306746e-52	410	3.358057e-54
410.40	3.358057e-54	410.40.30	3.358057e-54

Table 2. conserved protein interactions

No.	Yeast protein		Fly protein	
1	YIL034C	YKL007W	CG17158	CG10540
2	YKL190W	YML057W	CG4209	CG1455
3	YBL078C	YNL223W	CG12334	CG4428
4	YLR200W	YML094W	CG7770	CG7048
5	YJL031C	YPR176C	CG12007	CG18627
6	YDL145C	YIL076W	CG7961	CG9543
7	YLR170C	YPR029C	CG5864	CG9113
8	YFR004W	YOR261C	CG18174	CG3416
9	YMR314W	YOR362C	CG4904	CG1519
10	YJR068W	YNL290W	CG8142	CG5313
11	YNL290W	YOL094C	CG5313	CG14999
12	YDL030W	YJL203W	CG2925	CG16941
13	YDR328C	YFL009W	CG16983	CG15010
14	YDL081C	YLR340W	CG4087	CG7490
15	YDR211W	YOR260W	CG3806	CG8190
16	YAL003W	YKL081W	CG6341	CG11901
17	YKL028W	YKR062W	CG10415	CG1276
18	YKL058W	YOR194C	CG5163	CG5930
19	YDR448W	YGR252W	CG9638	CG4107

consider the PINs that p value $< 10^{-22}$ and more than 50% proteins in the PIN have the same function.

4.2. Directed Network. Take metabolic network as an example, we study directed biomolecular network querying between E.coli and Yeast by the INM. Metabolic network is the data that used by Pinter et al. [19], which include 113 metabolic

TABLE 3. protein interaction prediction

No.	protein	protein
1	CG9638	CG31973
2	CG3195	CG4087
3	CG8142	CG14999
4	CG9327	CG3416
5	CG4428	CG9277

TABLE 4. GOid of predicted protein function

No.	protein	GOid predicted
1	CG11154	0006470, 0005955, 0004723, 0007269, 0016192, 0008021
2	CG12334	0004197
3	CG12576	0016272, 0006457, 0051082
4	CG31135	0016272, 0006457, 0051082
5	CG7961	0008270
6	CG8942	0008270
7	CG17945	0008270
8	CG9543	0008270
9	CG12470	0006260, 0003677, 0005663, 0005524, 0003689, 0005634
10	CG6196	0006412, 0003735, 0022625, 0005811
11	CG14818	0006412, 0003735, 0022625, 0005811
12	CG10255	0006412, 0003735, 0022625, 0005811
13	CG14011	0005829, 0003746, 0005853, 0006414, 0005811
14	CG10654	0005829, 0003746, 0005853, 0006414, 0005811
15	CG5163	0031177

pathways of E.coli and 151 metabolic pathways of Yeast. There are 49 pathways are matched well, it demonstrates that lots of conserved pathways exist between the two species. Here 3 matched pathways are listed in A, B, C of Fig. 4.

In A of Fig. 4, enzyme 1.1.1.8 and 1.1.1.94 are different, but they are both belong to the 1.1.1.- classification, and their neighbors are identical. It makes us known that function of this metabolic pathway is conserved during evolution. The p value is 7.89×10^{-24}. In B of Fig.4, the upper one can be considered to be sub functions of the lower one. The p value is 8.05×10^{-22}. In C of Fig. 4, though the structure of the two graph are different, they both include the similar sub-pathway, which maybe occur because of evolution that organism have to adapt to the changed environment. The p value is 2.65×10^{-15}.

5. Conclusions and Future Work

Similarity between biomolecular networks is of great significance in species evolution and diseases investigation. With expansion of biomolecular networks, the

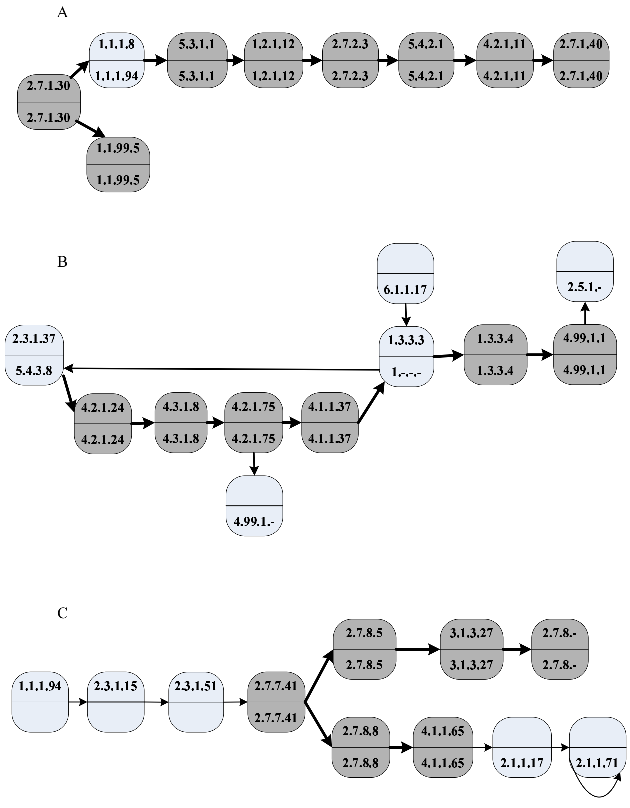

Figure 4. Metabolic pathway of yeast and E.coli. Each node represents a match: the upper part represents the enzyme of Yeast and the lower part represents the enzyme of E.coli. Color shades reflect enzyme matched well, and the thick arrows imply that the pathway exists both in Yeast and E.coli. (A)Metabolic pathway of yeast aerobic glycerol catabolism and E.coli glycerol degradationl. (B) Metabolic pathway of yeast heme biosynth and E.coli proto siroheme biosynth. (C) Metabolic pathway of yeast phosphatidic biosynth and E.coli phospholipids biosynthl.

computing scale of conventional sequential algorithms gradually cannot meet requirement of bioinformatics. The INM is developed to achieve biomolecular network querying, hence sub networks that are identical or similar to the target network within or cross species in biological sense can be discovered. It can process both undirected networks, such as PINs, and directed networks, such as metabolic networks. The parallel algorithm of the INM is developed to treat with large-scale networks. Experimental results demonstrate that its speedup and scalability are promissing.

To keep pace with the development of bioinformatics, some challenging problems should be considered in the future studies. The iteration process of computing similarity matrix involves all molecules in the network; heuristic algorithm should be taken into account to reduce correlation between molecules during iteration, and new parallel strategies should be adopted to further improve parallel efficiency. Moreover, computational results should be labored to mine more information in biomolecular networks.

References

[1] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, FH Brembeck, H. Goehler, M. Stroedicke, M. Zekner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, EE Wanker. A human protein-protein interaction network: a resource for annotating the proteome. Cell, 122(6): 957-968, 2005.

[2] R. Wang, Y. Wang, L. Wu, X. Zhang, L. Chen. Analysis on Multi-domain Cooperation for Predicting protein-protein interactions. BMC Bioinformatics, 8(1): 391, 2007.

[3] K. Basso and et al. Reverse engineering of regulatory networks in human B cells. Nat Genet, 37: 382-390, 2005.

[4] Y. Wang, J. Joshi, D. Xu, X. Zhang, L. Chen. Inferring gene regulatory networks from multiple microarray datasets. Bioinformatics, 22: 2413-2420, 2006.

[5] R. Albert. Scale-free networks in cell biology. J Cell Sci, 118: 4947-4957, 2005.

[6] A. Barabasi, Z. Oltvai. Network biology: understanding the cell's functional organization. Nature Rev Gen, 5: 101-113, 2004.

[7] S. Zhang, X. Zhang, L. Chen. Biomolecular network querying: a promising approach in systems biology. BMC Systems Biology, 2: 5, 2008.

[8] X. Zhao, L. Chen. Domain-Domain Interaction Identification with a Feature Selection Approach. attern Recognition in Bioinformatics (Lecture Notes in Computer Science), 178-186, 2008.

[9] X. Zhao, R. Wang, L. Chen, Kazuyuki Aihara. Uncovering signal transduction networks from high-throughput data by integer linear programming. Nucleic Acids Research, 36(9): e48, 2008.

[10] Z. Wu, X. Zhao, L. Chen. Identifying responsive functional modules from protein-protein interaction network. Mol Cells, 27(3): 271-277, 2009.

[11] M. Girvan, M. Newman. Community structure in social and biological networks. Proc. Natl. Acad. Sci., 99: 7821-7826, 2002.

[12] A. Rives, T. Galitski. Modular organization of cellular networks. Proc. Natl. Acad. Sci., 100: 1128-1133, 2003.

[13] V. Spirin, L. Mirny. Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci., 100: 12123-12128, 2003.

[14] Roded Sharan, Trey Ideker. Modeling cellular machinery through biological network comparison. Nature Biotechnology, 24(4): 427-433, 2006.

[15] Brian P. Kelley, Bingbing Yuan, Fran Lewitter, Roded Sharan, Brent R. Stockwell, Trey Ideker. PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Research, 32: w83-w88, 2004.

[16] Roded Sharan, Silpa Suthram, Ryan M. Kelley, Tanja Kuhn, Scott McCuine, Peter Uetz, Taylor Sittler, Richard M. Karp, Trey Ideker. Conserved patterns of protein interaction in multiple species. PNAS, 102(6): 1974-1979, 2005.

[17] Brian P. Kelley, Roded Sharan, Richard M. karp, Taylor Sittler, David E. Root, Brent R. Stockwell, Trey Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. PNAS, 100(20): 11394-11399, 2003.

[18] Silpa Suthram, Taylor Sittler, Trey Ideker. The Plasmodium protein network diverges from those of other eukaryotes. Nature, 438: 108-112, 2005.

[19] Ron Y. Pinter, Oleg Rokhlenko, Esti Yeger-Lotem, Michal Ziv-Ukelson. Alignment of metabolic pathways. Bioinformatics, 21(16): 3401-3408, 2005.

[20] Tomer Shlomi, Daniel Segal, Eytan Ruppin, Roded Sharan. QPath: a method for querying pathways in a protein-protein interaction network. BMC Bioinformatics, 7: 199, 2006.

[21] A. Ferro, R. Giugno, G. Pigola, A. Pulvirenti, D. Skripin, G. D. Bader, D. Shasha. Netmatch: a Cytoscape plugin for searching biological networks. Bioinformatics, 23(7): 910-912, 2007.

[22] Z. Li, S. Zhang, Y. Wang, X. Zhang, L. Chen. Alignment of molecular networks by integer quadratic programming. Bioinformatics, 23: 1631-1639, 2007.

[23] Y. Tohsato, H.Matsuda, A.Hashimoto. A Multiple Alignment Algorithm for Metabolic Pathway Analysis Using Enzyme Hierarchy. in Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, 376-383, 2000.

[24] K. Pawlowski, L. Jaroszewski, L. Rychlewski, A. Godzik. Sensitive sequence comparison as protein function predictor. Pac Symp Biocomput, 42-53, 2000.

[25] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman. Basic Local Alignment Search Tool. Journal of Molecular Biology, 215: 403-410, 1990.

[26] Maureen Heymans, Ambuj K. Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Bioinformatics, 19: i138-i146, 2003.

[27] Maureen Heymans, Ambuj K.Singh. Building phylogenetic trees from the similarity analysis of metabolic pathways. Department of Computer Science, University of California, Santa Barbara. Tech. Rep. 2002-33, 12 2002.

[28] S.Maslov, K.Sneppen. Specificity and stability in topology of protein networks. Science, 296: 910-913, 2002.

[29] Ralf Mrowka, Andreas Patzak, Hanspeter Herzel. Is there a bias in proteome research? Genome Research, 11(12): 1971-1973, 2001.

[30] Charlotte M. Deane, Lukasz Salwinski, Ioannis Xenarios, David Eisenberg. Protein interactions: two methods for assessment of the reliability of high throughput observations. Molecular Cell Proteomics, 1(5): 349-356, 2002.

[31] Sourav Bandyopadhyay, Roded Sharan, Trey Ideker. Systematic identification of functional orthologs based on protein network comparison. Genome Research, 16: 428-435, 2006.

[32] Huahai He, Ambuj K.Singh. Closure-Tree: An Index Structure for Graph Queries. in ICDE'06: Proceedings of the 22nd International Conference on Data Engineering, IEEE Computer Society: 38, 2006.

[33] Jiang Xie. Numerical Researches on Protein-Protein Interactions Network. Doctoral thesis of Shanghai University, 2008.

[34] NCBI National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov.

[35] G. Mishra, M. Suresh, K. Kumaran, N. Kannabiran. Human Protein Reference Database - 2006 Update. Nucleic Acids Research, 34: D411–D414, 2006.

[36] Andrew Chatr-aryamontri, Arnaud Ceol, Luisa Montecchi Palazzi, Giuliano Nardelli, Maria Victoria Schneider, Luisa Castagnoli, Gianni Cesareni. MINT: the Molecular INTeraction database. Nucleic Acids Research, 35: D572-574, 2007.

[37] Jiang Xie, Xiaobin Zhang, Wu Zhang. PSE-Bio: a grid enabled problem solving environment for bioinformatics. in Proceedings of the Third IEEE International Conference on e-science and Grid Computing, 529-535, 2007.

[38] I.Xenarios, L.Salwinski, X. J. Duan, P. Higney, S. M. Kim, D. Eisenberg. DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. Nucleic Acids Research, 30: 303-305, 2002.

[39] MIPS database, http://mips.gsf.de/.

[40] BioGrid database, http://www.thebiogrid.org/index.php.

[41] R.J. Wilson, J.L. Goodman, V.B. Strelets, the FlyBase Consortium. FlyBase: integration and improvements to query tools. Nucleic Acids Research, 36: D588-D593, 2008.

School of Computer Engineering and Science, Institute of Systems Biology, Shanghai University, Shanghai, 200072 China
E-mail: jiangx@shu.edu.cn and wzhang@shu.edu.cn

Academy of Mathematics and Systems Science, CAS, Beijing,100190 China
E-mail: zsh@amss.ac.cn

School of Life Sciences, Institute of Systems Biology, Shanghai University, Shanghai, 200444 China
E-mail: tqwen@staff.shu.edu.cn