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

Received by the editors October 12, 2009 and, in revised form, April 28, 2010.
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