Identification of functional modules in protein networks by near-clique detection

Shu Tadaka[1,a)] Takeshi Obayashi[1,b)] Kengo Kinoshita[1,2,3,c)]

Abstract: In analysis of protein-protein interaction (PPI) networks, detection of functional module is one of the most important problems for understanding of cellular function of uncharacterized proteins. Identification of functional modules has been mainly done by searching densely connected subgraph, and some methods have been proposed to identify modules by using different criteria of densely connected subgraph. Here, we propose a new method NCMine to detect functional modules by extracting near-clique subgraphs aiming to get better identification of functional modules. We tested NCMine and other methods by using human PPI network from HPRD. When NCMine is applied to the network, it extracts about 2000 modules and 55% of them have at least one enriched GO term that is shared among members of a module. On the other hand, the percentage of GO-enriched modules extracted by other methods was lower than that of NCMine. This indicates that NCMine is superior to other methods in identification of biologically meaningful modules.

1. Introduction

Protein is a highly polymerized compound which consists of chains of amino acids, and it associated with various phenotypes. However, a biological phenomenon is not conducted by a single protein, but is conducted by a set of proteins or a protein complex. These sets of cooperatively working proteins are called functional modules.

Interactions between proteins are represented by a network in which nodes correspond to proteins, and edges correspond to physical interactions between proteins, and proteins that are included in a functional modules tend to form a densely connected region in a network. Therefore identification of functional modules is done by searching densely connected region in networks. For example, MCODE[1], CFinder[2] and NeMo[3] have been proposed to identify functional modules by searching densely connected region using different criteria of densely connected region. However, some problems exist in the previous methods: (i) some protein complexes are known to change its composition of proteins depending on conditions and that means, in network node clustering problem, one protein might be assigned to multiple clusters, however, a few methods consider such situations. (ii) previous methods don't seem to be enough in aspect of running time when applied to real biological networks.

Here we propose a new method NCMine to detect functional modules by extracting near-clique subgraphs from networks aiming to overcome the problems in the previous methods. Here, the near-clique subgraph is defined as a subgraph obtained by removing a few edges from a clique. A previous research[4] succeeded to find protein complexes by mining of cliques, however, clique is very hard restriction for biological aspect. Therefore, we consider we can obtain structures from PPI networks which correspond to real complex by relaxing the requirement.

2. Proposed method

NCMine calculates PageRank[5] of nodes in the given network and uses them as node weights. Then NCMine selects node with the highest weight (Fig. 1), and the node is treated as initial local cluster. Next, NCMine searches nodes around the local cluster in the order of node weight, and adds them to the local cluster until cliqueness of the cluster is grater than pre-defined threshold. The "cliqueness" is defined as ratio of the number of edges in a cluster to the number of the possible edges when the cluster is a clique. In this way, a local cluster is constructed, and local clusters are built from every node in the network. After construction of the local clusters, they are merged by overlap of nodes, resulting in the final solution. (Fig. 2)

The reasons why PageRank is used as node weight are as follows: (i) PageRank reflects connectivity of nodes in a network, so that calculation of node weight and traversal of nodes in the order of its weight lead to effective local cluster construction. (ii) the PPI network tested in this paper is an undirected network, however, when we apply NCMine to some directed networks such as regulation networks, it is expected that we can construct local clusters more effectively than other weighting scheme.

[1] Graduate School of Information Sciences, Tohoku University, 6-3-09 Aoba, Aramaki-aza, Aoba-ku, Sendai, Miyagi 980-8575, Japan
[2] Tohoku Medical Megabank Organization, Tohoku University, 2-1, Seiryo-cho, Aoba-ku, Sendai, Miyagi 980-8575, Japan
[3] Institute of Development, Aging, and Cancer, Tohoku University, 4-1, Seiryo-cho, Aoba-ku, Sendai, Miyagi 980-8575, Japan
[a)] tadaka@sb.ecei.tohoku.ac.jp
[b)] obayashi@ecei.tohoku.ac.jp
[c)] kengo@ecei.tohoku.ac.jp

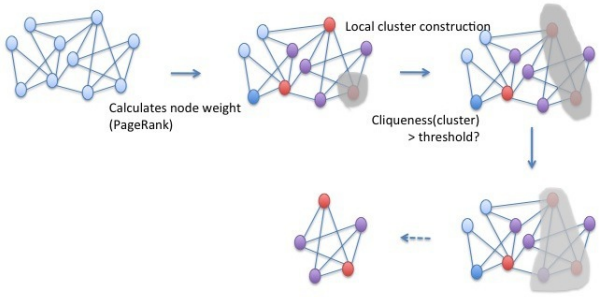

Fig. 1 Overview of NCMine: construction of local clusters

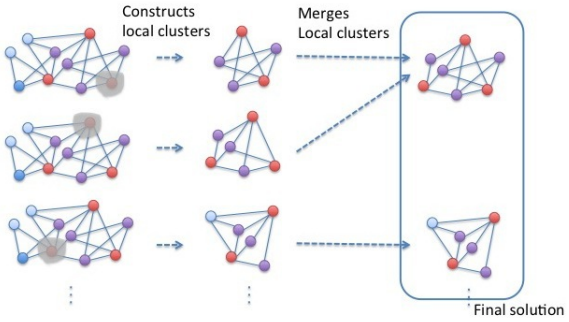

Fig. 2 Overview of NCMine: merge of local clusters

3. Comparison with other methods

We compared NCMine and other methods by using two datasets: (i) artificially generated networks with near-clique structures embedded, (ii) human PPI network obtained from HPRD[6].

3.1 Comparison on artificially generated networks

We generated random networks, embedded some near-clique structures in the random networks, and tested that NCMine can successfully extracts embedded clusters. We also investigated characteristics of clusters extracted by other methods. We ran clustering methods with various parameters, compared clusters obtained as output and clusters we actually embedded, and calculated recall and precision from the results. The figure 3 indicates that NCMine shows better recall-precision trade off than the other methods.

We also checked the dependency of running time on network size. The figure 4 shows NCMine is competitive to other methods.

3.2 Comparison on human PPI network obtained from HPRD

We applied NCMine and other methods to human PPI network obtained from HPRD. The result is shown in table 1. The "biologically meaningful clusters" in the table means clusters which have at least one enriched Gene Ontology (GO) term that is shared among members of a cluster, and seems to be biologically interpretable by GO. GO enrichment test is performed using Fisher's exact test. A term whose p-value is less than 0.05 and shared by more than 60% of the cluster member is treated as enriched GO term. The result shows that NCMine is superior to other methods in identification of biologically meaningful modules when applied to the network. The running time of NCMine was also reasonable and competitive to the other methods.

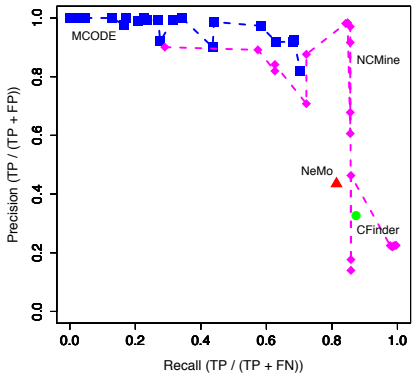

Fig. 3 Recall and precision when applied to artificially generated networks

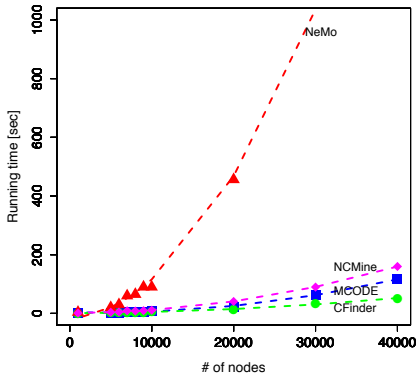

Fig. 4 Dependency of running time on network size when applied to artificially generated networks

Table 1 Human PPI clustering results

Method	# of found clusters (A)	# of biologically meaningful clusters (B)	Ratio (B/A)	Running time
NCMine	2385	1341	0.55	15s
MCODE	205	93	0.45	7s
CFinder	764	315	0.41	28s
NeMo	1510	784	0.52	180s

References

[1] Bader GW et al.: An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, Vol.4 (2003)
[2] Adamcsek B. et al.: CFinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics*, Vol.8, pp.1021-1023 (2006)
[3] Rivera CG et al.: NeMo: Network Module identification in Cytoscape, *BMC Bioinformatics*, Vol.11 (2010)
[4] Khner, S. et al. Proteome organization in a genome-reduced bacterium. *Science*, Vol.326, pp.1235—1240 (2009)
[5] Lawence P. et al.: The PageRank Citation Ranling: Bringing Order to the Web, *Stanford Digital Libraries Technologies Project*, (1998)
[6] Keshava Prasad, T. et al.: Human Protein Reference Database–2009 update, *Nucleic acids research*, Vol.37 (2009)