Structure-based Multilevel Approaches for Large-scale Proteomic Networks.

S. Oliveira and S.-C. Seok
Department of Computer Science
The University of Iowa
Iowa City, IA 52242 U.S.A.

October 24, 2006

Most cellular processes are believed to be carried out by groups of highly interacting proteins called functional modules, protein complexes, or molecular complexes. Recent large-scale high-throughput experiments, and integration of published data, have generated large protein-protein interaction (PPI) networks. Even the simplest organism, yeast, has more than 17 thousand proteins. Protein complexes can be detected by identifying highly connected sets of proteins in PPI networks. Computational identification of functional modules or protein complexes can provide an inexpensive guideline for biological experiments.

There are a number of challenges in treating protein-protein interaction data. One is that many high-throughput experiments have high error rates, which results in a many false positives for interactions between proteins. However, the biggest obstacle to identifying functional modules in PPI networks is the uniform weighting of edges.This uniform weighting limits the information on how close pairs of proteins are. Moreover, large-scale networks demand excessive computational time as the size increases.

We have developed various multilevel algorithms to improve existing graph clustering algorithms for protein-protein interaction (PPI) networks [5, 6]. Our matching based algorithm combines a greedy algorithm for weighted graphs and a minimal matching idea in [5]. Matching based algorithms try to merge at most two nodes joined by an edge or node related information. However, these matching based algorithms on unweighted networks are hampered by the lack of closeness information between nodes. Many of these algorithms fail to merge correct pairs of proteins because they do not take advantage of any structured analysis. In [6] we developed Triangular Clique (TC) based algorithms which merge highly connected triples of nodes. Our TC-based multilevel algorithm was inspired by Spirin et al's use of cliques to identify highly connected clusters [7]. These TC based algorithm showed more proteins correctly merged into supernodes than any other matching based algorithm. This is because all three proteins in a TC have very high chance of being a part of the same functional module. A weakness of the TC based algorithms is that there are very many TCs found in even a moderately-sized clique. For example, there are 560 TCs in a clique of 16 nodes.

Our next approach generalizes these TC based algorithms to cliques of various sizes. A recent approach for identifying protein complexes used maximal cliques to create subgraphs with high densities [8]. Cliques are called maximal when they are not completely contained in another clique. In general, enumerating all maximal cliques takes more time than finding all TCs which take only about $O(n \ln n)$ in PPI networks. Fortunately, scale-free networks are quite sparse, so all maximal cliques are enumerated quickly. Our experimental results show the quality of maximal cliques have high probability to be found in the same protein complexes. They also show that the bigger the maximal clique, the higher the chance the nodes are included in the same protein complex. This strongly motivates us to consider maximal cliques in conjunction with our multilevel algorithms. We call these new approaches structure-based multilevel algorithms.

One of the distinct features of PPI networks is that they have a power-law property [2]. These networks have many low-degree nodes and a relatively small number of high-degree nodes. Very high degree nodes correspond to proteins that interact with most other proteins, and these interactions can obscure the connections between other proteins. The *2*-core network is built by removing nodes belonging to the non-cyclic part of the graph and nodes of degree one or less. The removed nodes can later be added back to the group of clusters found. We show the effectiveness of the *2*-core network approaches by presenting the clustering quality for these removed nodes. We apply the new algorithms to a recently reported protein-protein interaction network in the yeast *Saccharomyces cerevisiae* [4]. The number of levels to produce best results is a common issue in multilevel algorithms. TC based algorithms showed that one level of grouping (or coarsening) is enough to produce as good clustering results as other matching based algorithms. Similarly, the new maximal clique based multilevel algorithms achieves this with one or two levels of coarsening.

Further applications of our clustering algorithm may include structure mining of complex networks which have a power-law property. Examples of such networks are widespread and varied. They include genetic networks, the World Wide Web, citation networks, biological networks and social networks. Current clustering algorithms are based on graph partitioning, so no vertex is allowed in more than one group. However, some proteins are involved in more than one cellular process. Similarly, some documents can be classified under two or more topics. Recently, various researchers have been working on clustering algorithms considering overlapping [1, 3]. Research on exploiting multilevel algorithms with overlap will be carried out to address this problem.

References

[1] A. Banerjee and et al. Model-based overlapping clustering. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 532–537, 2005.

[2] S. Bornholdt and H.G. Schuster, editors. *Handbook of Graphs and Networks.* Wiley VCH, 2003.

[3] S. Mallela I. Dhillon and D. Modha. Information-theoretic co-clustering. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 89–98, 2003.

[4] N. Krogan and et al. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–643, 2006.

[5] S. Oliveira and S.C. Seok. A multilevel approach for identifying functional modules in protein-protein interaction networks. *Proceedings of IWBRA 2006, Lecture Notes in Computer Science*, 3992:726–773, 2006.

[6] S. Oliveira and S.C. Seok. Triangular clique based multilevel approaches to identify functional modules. the 7th International Meeting on High Performance Computing for Computational Science (VECPAR 2006), 2006.

[7] V. Spirin and L.A. Mirny. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, 100(21):12123–12128, October 2003.

[8] C. Zhang and et al. Fast and accurate method for identifying high-quality protein-interaction modules by clique merging and its application to yeast. *J. Proteome Res.*, 5(4):801 –807, 2006.