Identifying hierarchical community structures in biological networks

Jianhua Ruan[1] and Weixiong Zhang[1,2]
Department of Computer Science and Engineering[1] and Department of Genetics[2]
Washington University in St. Louis, St. Louis, MO 63130, USA
jruan@cse.wustl.edu, zhang@cse.wustl.edu

October 31, 2006

Biological networks, similar to many real-world networks, often have the so-called community structures, i.e., groups of bio-molecules that are highly associated among themselves, while having relatively fewer and/or weak associations with the rest of the network. A community in a gene network often corresponds to genes involved in the same biological process. For example, genes within the same community of a protein-protein interaction (PPI) network often belong to some known protein complexes, while genes within the same community of a metabolic network are often located on the same metabolic pathway. Identifying and analyzing communities, therefore, help us focus on the coarse-grained, high-level organizational principles of biological networks rather than the functions of individual bio-molecules.

Recently, several algorithms have been developed to identify network communities based on the optimization of an objective function called the modularity (Q). Empirical studies have shown that modularity optimization is a very effective way to search for natural communities, without prior knowledge about the number of communities and sizes of the communities. However, the optimization of Q is NP-hard, and existing algorithms are often trapped by local optima. A simulated annealing approach has been previously proposed, with impractical running time for large networks. The best algorithm so far in terms of both efficiency and effectiveness is from Newman (Proc Natl Acad Sci USA, 103: 8577-8582, 2006), who first proposed the Q function.

In this poster, we present an efficient heuristic algorithm, called *Qcut*, which can find higher Q values than the Newman's method on a large number of simulated and real networks. We also show that, for many simulated networks, when the communities are not very small compared to the networks, higher Q values indeed correspond to better community structures. Using standard measures, Qcut has achieved much higher accuracy than the Newman's method in recovering the known communities (e.g., 100% vs. 60%, and 90% vs. 40%), especially when the community structures are weak, i.e., when there are a large number of inter-community edges.

On the other hand, when the community sizes are small, or when the networks have hierarchical community structures, we show that optimizing Q is not always a good strategy, since it often merges small communities and ignores low-level sub-communities. The Q function uses a null model to estimate the expected number of edges between two sub-networks, and partition them into two communities only if the actual number of edges between them is smaller than expected. However, for two small communities, even a single edge between them may seem unexpected by the null model. As a result, they cannot be separated without reducing the Q value. To deal with this issue, we propose another algorithm, *HQcut*, which recursively applies *Qcut* to each already identified community to search for sub-communities. To ensure that the sub-community structures are genuine, we use Monte-Carlo test to estimate the statistical significance of the partitions. Applied to a large number of simulated networks that contain both large and small communities, *HQcut* is able to recover all the embedded communities with very high accuracies, while the Newman's method often fails completely.

We have applied the *Qcut* algorithm to gene co-expression networks, and evaluated the communities using Gene ontology (GO) annotations, ChIP-chip data, as well as PPI data. Compared to several popular clustering algorithms, including k-means, self-organizing maps, and spectral clustering, the *Qcut* algorithm can identify communities of genes that are statistically more enriched with common functional terms, are more likely to be co-regulated, and more frequently interact among themselves physically. The results have been submitted as a paper to the RECOMB Satellite Conferences on Systems Biology.

We have also applied the *Qcut* algorithm to a yeast protein-protein interaction network, and identified $\sim$100 protein communities. Using GO analysis, we find that most communities correspond to highly specific functional modules. The community sizes range from 2 to $\sim$400, and approximately follow a power-law distribution. Compared to the hand-curated known protein complexes in the MIPS database, the smaller communities often match well with individual protein complexes, while the larger ones usually contain several functionally related complexes in their entirety. With the *HQcut* algorithm, we are able to further partition the larger communities into smaller sub-communities that have better correspondence in the MIPS protein complexes database.

Finally, we have analyzed the yeast PPI communities identified by *HQcut* to study the relationships between community roles and gene essentiality. Following Guimera & Nunes Amaral (Nature, 433:895-900, 2005), we compute a participation index for each gene to measure how diverse its interaction partners are distributed among different communities. Several studies have shown that the so-called hub genes, those with a large number of interactions, are more likely to be essential to the cell's viability. Surprisingly, we find that the hub genes with intermediate participation indices, i.e., those whose interaction partners are distributed among a small number of communities, have the highest possibility to be essential, even though they are not the most connected in the network. On the other hand, the hub genes whose interaction partners are mostly restricted within a single community or distributed among many communities are less likely to be essential, but are still more frequently essential than the non-hub genes. We are currently exploring the possible biological and alternative explanations for this phenomenon.