Biological Significance of Modular Structures in Protein Networks

Keywords: module, protein networks, systems biology

INTRODUCTION

It is widely assumed that cellular functions are organized in a highly hierarchical and modular mannar (Hartwell, Hopfield et al. 1999). Each module is a discrete object composed of a group of tightly linked components and performs an independent task, seperable from the function of other modules (Hartwell, Hopfield et al. 1999; Ravasz, Somera et al. 2002). With the advent of genome scale data, many efforts have been devoted into identifying modular structures and their biological significance (Barabasi and Oltvai 2004). To thoroughly study the modular structure(Ravasz, Somera et al. 2002; Rives and Galliski 2003; Yook, Oltvai et al. 2004) of large scale networks require effective and automatic method that can separate modules. Clustering could be one of the methods to discover module structure in networks using topological structure (Ravasz, Somera et al. 2002; Giot 2003; Yook, Oltvai et al. 2004). And some studies combined clustering with functional genomics data also gave good result (Stiart, Segal et al. 2003; Tornow and Mewes 2003). By using RNA expression profile data, Han etc. (Han, Bertin et al. 2004) divided the hubs into data hub and party hub and showed modularity organization in yeast protein-protein interaction networks. Fraser (Fraser 2005) studied the evolutionary conservation of data hubs and party hubs and suggested the occurring of modules through exon shuffling.

However, the module separation by clustering usually gives ambiguity result (Barabasi and Oltvai 2004), part of which is because of the network's hierarchical structure. But lack of objective judgment could be another reason. Newman (Newman and Girvan 2004) proposed a method to measure the modularity of the separated modular structure and devised a greedy method to separate module according to edge betweenness (Newman 2004). Based on the modularity definition, many other algorithms were invented to get the global maximized modularity because greedy methods could be easily trapped in local maximization (Duch and Arenas 2005; Guimera and Amaral 2005; Massen and Doye 2005). Heuristic algorithms often give better result, especially for the networks with relatively low hierarchical structure. Using stimulated annealing to maximizing the network's modularity, Guimera and Amaral (Guimera and Amaral 2005) was able to identify the functional organization of metabolic networks. According to the topological properties, functional roles were determined for each node and they showed the evolutionary conservation among different roles of nodes. Nevertheless, their result that intra-module hubs are less conserved than the intermodule nodes, which contradicts with Fraser's (Fraser 2005) result for protein networks that intramodule hubs are more conserved.

In this paper, modules are separated solely based on the topology of protein networks. The biological significance of modular structures is accessed by functional and evolutionary data. The modular structures show highly evolutionary conservation when comparing the orthologous proteins in yeast and fly modules.

RESULTS

Modular structures in protein networks

The modules of protein-protein interaction networks of yeast *Saccharomyces cerevisiare* and fly

Drosophila melanogaster was identified using simulation annealing algorithm. For the protein networks, the largest component contains 3862 nodes (about 94% out of 4216 nodes for the entire network) forming 7208 edges for yeast and 6279 nodes (95% of all) forming 10094 edges. The overall modularity is 0.666 for yeast and 0.685 for fly suggestting a high modular organization of the network (table 1).

Table 1. Summery of the characters of the largest component of yeast networks

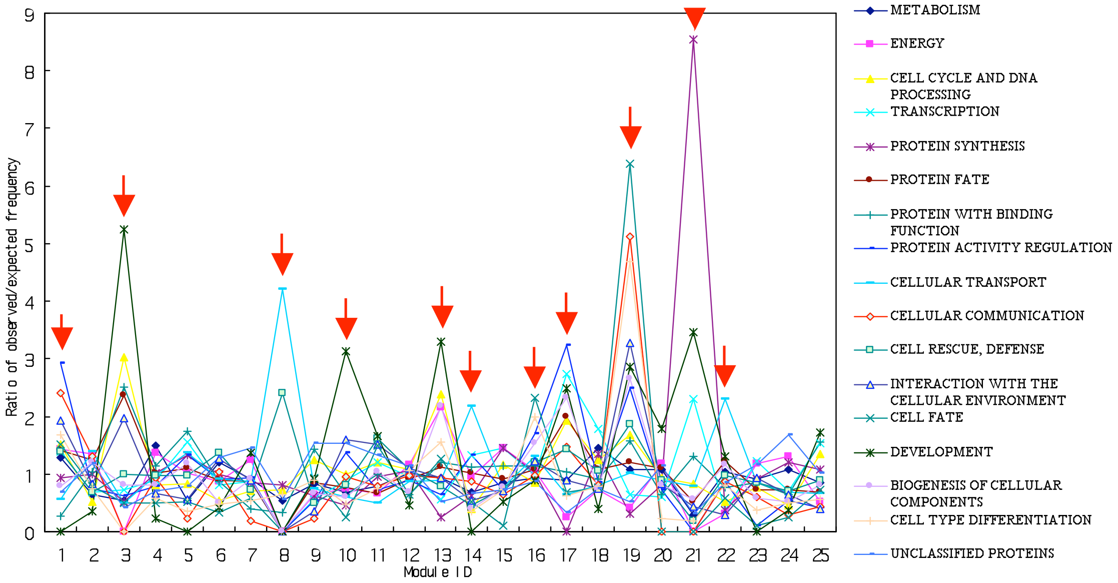
Organism	Average degree	Average shortest path-length	The largest Shortest Path-length	Modularity	Number of modules	Average number of nodes in a module	Number of nodes in the largest module	Number of nodes in the smallest module
Yeast	3.73	4.84	13	0.666	25	154.5	377	16
Fly	3.22	6.48	19	0.685	27	232.6	536	55

Functional correspondence of modules

To correlate the functional properties and topological modules, functional classification established by MIPS was used in which each protein is assigned a function category according to the enzyme function. I filtered the data that only functional categories containing more than 10 proteins were used, and totally there are 17 functional categories including unclassified proteins.

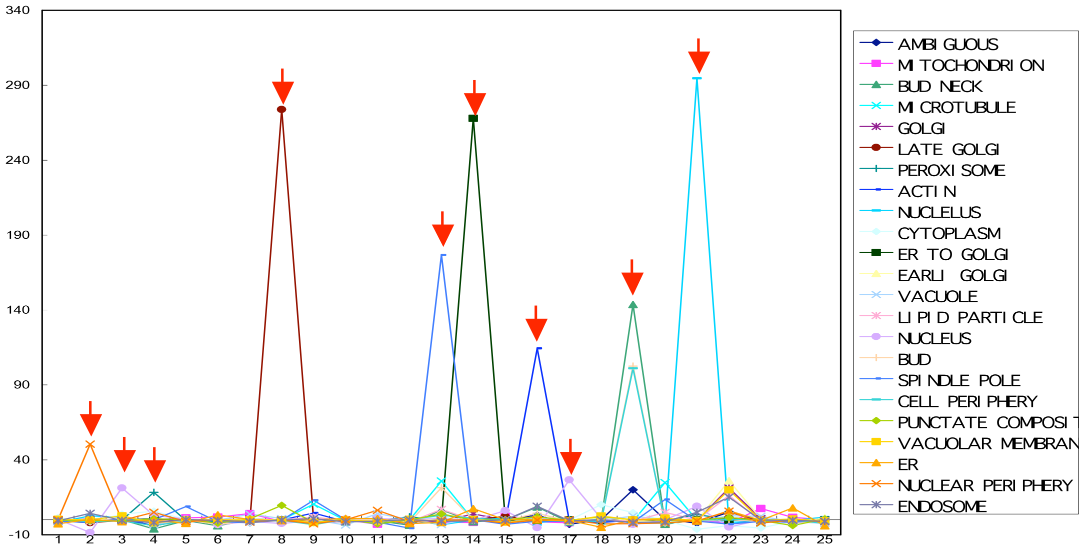
For each module, the sum of genes in each functional category was calculated. To get the expected number, random module separation ran for 1000 times, and the average was used as the mathematical expectation. I used χ^2 test for each module, and 19 of 25 modules show significant biased distribution after Boffereni correction, which suggests biological meaning of topological module separation. χ^2 test was also used to each functional category within each module. In each module, there are some functions biased distributed suggesting the functional correspondence (figure 1). Functional profile is used to visualize the enriched distribution of each functional category among modules. According to the profile, module #1, #3, #8, #10, #13, #14, #16, #17, #18, #19, #21, and #22 show significant biased distribution of enriched functional categories (details in supplementary table 1).

Figure 1. Profile of functional categories in each module



Protein complex is another biological functional module. A protein complex is composed of several proteins that are closely connected to each other and perform a certain function as a whole "module". I analyzed the possibility that a pair of interacted proteins also appear in the same complex. The basic logic is that if the network topological modules could correspond to the protein complex, we should observe a higher possibility to find a pair of interacted proteins within the same module than between different modules. For within module connections, there are 1046 interacted pairs of protein and 641 of them are in the same complex. While for between module connections, only 58 out of 333. The difference is very significant (2 by 2 contingency test, $p < 0.0001$) which strongly suggest the correspondence between protein complex and topological modules.

A discrete module that performs a certain function is much likely located in the same cellular location. Based on this idea, global protein localization data in budding yeast is used to check the biased distribution of each module. The same as functional category, each location of proteins in modules is counted and simulation was used to calculate the expected distribution. χ^2 test was used to test for statistical significance. From the result (figure 2),



Evolutionary conservation of topological structure

Nonsynonymous changes in sequences will result in protein sequences change, and could be used to measure the distance or evolutionary conservation of proteins. And a low evolutionary rate usually indicates strong functional constraint. Nonsynonymous substitutions per site (dN) was calculated for all the genes with ortholog in *S. bayanus* a closely related species. The Spearman's rank correlation was used to evaluate the evolutionary conservation of topological structure (table 1).

The result shows only dN and within-module degree is very significantly and negatively correlated which indicate that within-module hubs are more conserved. And this result is robust because there is no significant correlation between between-module degree and dN which would not confound the correlation between within-module degree.

Table 1. Correlations between evolutionary conservation and topological structure

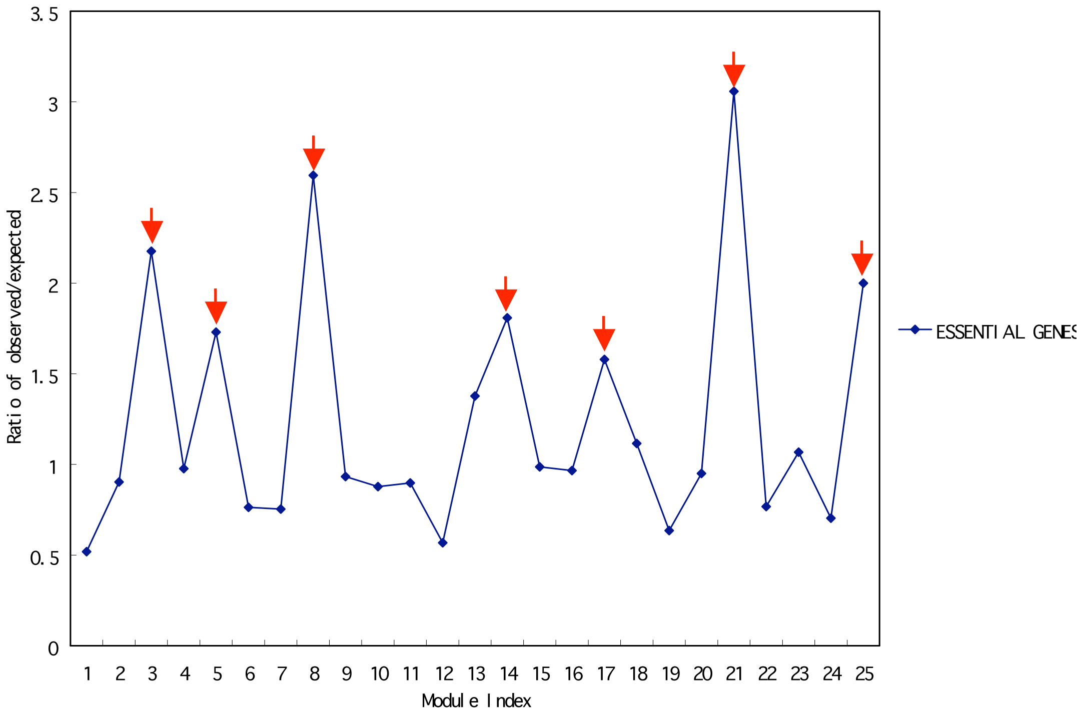
	within-module degree		between-module degree		participation coefficient		betweenness centrality	
	ρ	p	ρ	p	ρ	p	ρ	p
gene loss	-0.03	0.0892	-0.02	0.2685	0	0.8115	-0.02	0.3342
dN	-0.04	0.0128	-0.02	0.2321	-0.01	0.5869	-0.02	0.1930

Another way to measure the evolutionary conservation is the phylogenetic conservation across species. If the gene subjects high functional constraints, it is very unlikely to be lost during the evolution. So, by counting the gene loss events on a phylogenetic tree also reflects the evolutionary conservation for the gene. I used all *S. cerevisiea* gene to blast against the other 9 species and identified a number of gene loss events on each brach. And in the following work I will analyze the distribution and correlation with the network topological properties.

Functional differences among modules

Because of the functional correspondence to module, the importance of each module may be varied depending on the functional importance. To access the differences among modules, I examined the distribution of essential genes which would cause death if knocked out. If one module has more essential genes, it tends to be more important in terms of function. While if there is no functional differences among modules, essential genes should be randomly distributed in each module. χ^2 test was used to test the distribution bias (figure 3).

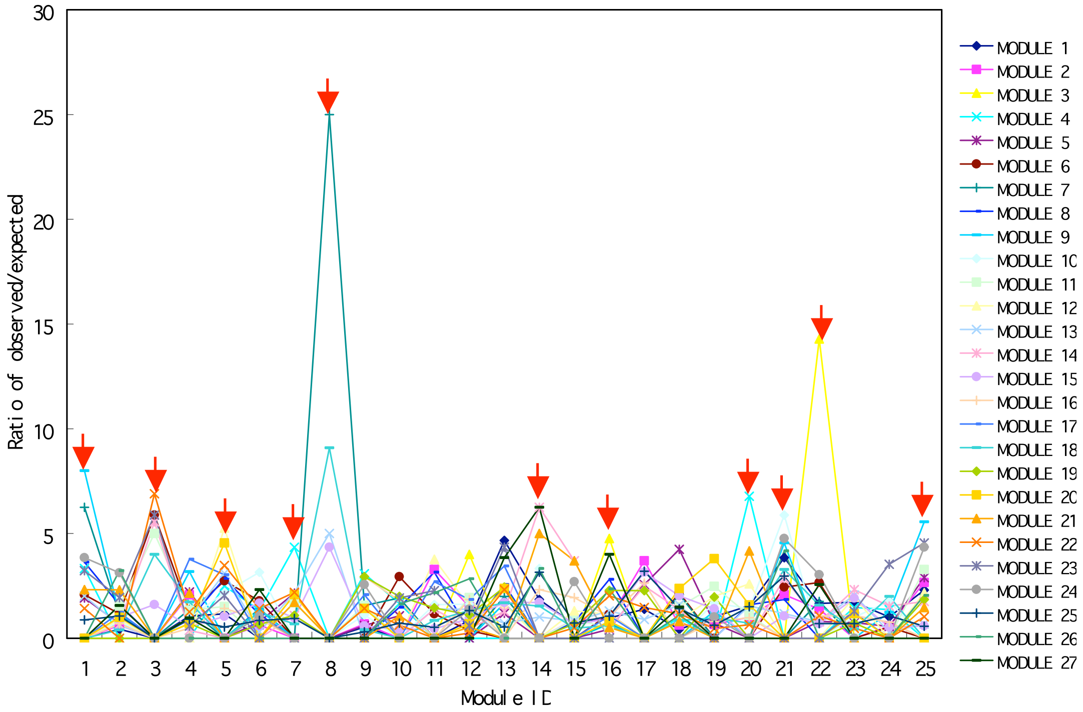
Figure 3 Profile of essential genes among moduels



The χ^2 test for the entire distribution is very significant ($p < 0.00001$) which suggest a great biased distribution of essential genes among modules. And after Bofferenii correction, there are still 3 modules (module ID 12, 21, 25) which are statistically deviated from the expectation. Two of them are observed a much higher number of essential genes than expectation, and one of them has much lower number of essential genes.

Hierarchical structural conservation across species

Although the interactions between proteins are less conserved across distant species (Gandhi, Zhong et al. 2006), they may show some conservation at modular structure level in which orthologous genes across species share the same module. To test this hypothesis, I used the same method to separate modules for *D. drosophila* protein-protein interaction network. And the homologous genes are identified using reciprocal blast best hits. If the homologous genes are grouped into the same module, the hierarchical modular structure will be conserved. The result shows this (figure 4). The same as function and location, conservation profile is used to show the relationship. And the χ^2 test was used to show the statistical significance.



DISCUSSION

The hierarchical structure of the protein networks could correspond to the functional category, which is consistent with previous studies(Tornow and Mewes 2003; Yook, Oltvai et al. 2004). However, my result does not suggest a direct one-to-one or one-to-multiple correspondence between functional categories and modules. Noise in the data maybe one reason; another reason could be the module structure may not reflect the enzyme based functional classification. The direct evidence that pairs of proteins within the same module tend to be in the same protein complex structure indicate that protein complex is better correspondence to the module structure (Spirin and Mirny 2003).

Jeong etc. (Jeong, Mason et al. 2001) studied the lethality and centrality in protein networks, and their result suggest a strong correlation between degree centrality (the same as node degree) and gene essentiality. My data analysis also proved it in which the number of connections of nodes is highly and negatively correlated with evolutionary conservation dN ($p < 0.00001$). Moreover, because no correlation between between-module degree and dN but significant negative correlation between within-module degree and dN suggest the overall correlation comes from within-module degree. This result suggests that the core of the module is more conserved and more important than the periphery nodes which is also consistent with Fraser's (Fraser 2005) study but different from Guimera and Amaral's (Guimera and Amaral 2005) study.

Because highly pleiotropic genes tend to have multiple functions and they could be the nodes link to many other modules, they are usually thought to be more conserved (Fraser 2005). Participation coefficient measures how the between-module links distributed (Guimera and Amaral 2005; Guimera and Amaral 2005). This result suggest that either participation coefficient is not corresponded to pleiotropy or there is no correlation between pleiotropy and evolutionary conservation.

The result of no significant correlation between dN and betweenness centrality is somehow surprising. Because Hahn and Kern's (Hahn and Kern 2004) study suggests a significant correlation between betweenness centrality and gene essentiality in three eukaryotic protein networks, although the correlation is very week.

Finally, although the differences among modules are statistically significant, only a few of them show a strong biased distribution. And querying these modules to functional category distribution gives no significant biased distribution.

In fact, protein-protein interaction data set is highly noisy (Barabasi and Oltvai 2004). This is partly come from the random errors in large scale experiments. But they may also come from the method of yeast two hybrids to detect interacting protein pairs. A number of studies (Aloy and Russell 2002; Han, Dupuy et al. 2005) suggested artifacts in protein interaction networks and sampling may also result in biased data set. This result in the genome scale analysis becomes very hard because noise reduces the signals significantly.

MATERALS AND METHODS

Protein-protein interaction networks data set for *Saccharomyces cerevisiea* was downloaded from MIPS (http://mips.gsf.de/). Only the nodes in the largest component of the network were used to separate module by simulated annealing algorithm.

Functional annotation for *S. cerevisiea* genes were also downloaded from MIPS and protein complex data came from IntAct database (http://www.ebi.ac.uk/intact). I compared the distribution of gene in each module and the functional category for each gene to evaluate the functional correspondence of module structure. Protein complex data was another way to measure the biological meaning of network module. I also calculated the betweenness centrality for each node using Pejak.

The evolutionary rate dN for *S. cerevisiea* genes were calculated against orthologous genes of *S. bayanus*. The loss of genes on other braches of yeast was also used to evaluate the phylogenetic conservation. Specificly, I use BLAST against the whole genome sequences of other 9 species, *S. paradoxus, S. mikatea, S. bayanus, C. glabrata, K. waltii, K. lactis, D. hansenii, Y. lipolytice, N. crassa and S. pombe* using 0.1 as criteria to detect gene loss. Parsomony method was used to calculate gene loss events on each brach with *S. pombe* as the outgroup. I compared the correlation among within module-degree, between-module degree, participation coefficient, betweenness centrality, dN and gene loss events to get evolutionary conservation for the roles of nodes.

Finally, the protein-protein interaction data for *Drosophila melanogaster* (downloaded from flybase (http://flybase.net)) was used to detect the evolutionary conservation of hierarchical modular structure between yeast and fly.

REFERENCE

Aloy, P. and R. B. Russell (2002). "Potential artefacts in protein-interaction networks." FEBS Letters **530**: 253-254.

Barabasi, A. L. and Z. N. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nature Rev. Biol. **5**: 101-113.

Callebaut, W. (2005). "The Ubiquity of Modularity." Modularity (ed by Werner Callebaut & Diego Rasskin-Gutman).

Duch, J. and A. Arenas (2005). "Community detection in complex networks using extremal optimization." Phys. Rev. E **72**: no. 027104.

Fraser, H. B. (2005). "Modularity and evolutionary constraint on proteins." Nature Genetics **37**: 351-352.

Gandhi, T. K., J. Zhong, et al. (2006). "Analysis of the human protein interactome and comparason with yeast, worm and fly interaction datasets." nature Genetics **38**: 285-293.

Giot, L. *e. a.* (2003). "A protein interaction map of *Drosophila melanogaster*." Science **302**: 1727-1736.

Guimera, R. and L. A. Amaral (2005). "Cartography of complex networks: modules and universal roles." J. Stat. Mechanics: Theory and Experiment **05**: 1742-5468.

Guimera, R. and L. A. Amaral (2005). "Functional cartography of complex metabolic networks." Nature **433**: 895-900.

Hahn, M. W. and A. D. Kern (2004). "Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks." Mol. Biol. Evol. **22**(4): 803-806.

Han, J. D., D. Dupuy, et al. (2005). "Effect of sampling on topology predictions of protein-protein interaction networks." Nature Biotechnology **23**: 839-844.

Han, J. J., N. Bertin, et al. (2004). "Evidence for dynamically organization modularity in the yeast protein-protein interaction network." Nature **430**: 88-93.

Hartwell, L. H., J. J. Hopfield, et al. (1999). "From molecular to modular cell biology." Nature **402**: C47-C52.

Jeong, H., S. P. Mason, et al. (2001). "Lethality and centrality in protein networks." Nature **411**: 41-42.

Massen, C. P. and J. P. Doye (2005). "Identifying communities within energy landscapes." Phys. Rev. E **71**(no. 046101).

Newman, M. E. (2004). "Fast algorithm for detecting community structure in networks." Phys. Rev. E **69**: no. 066113.

Newman, M. E. and M. Girvan (2004). "Finding and evaluating community structure in networks." Phys. Rev. E **69**: no. 026113.

Ravasz, E., A. L. Somera, et al. (2002). "Hierachical organization of modularity in metabolic networks." Science **297**: 1551.

Rives, A. W. and T. Galliski (2003). "Modular organization of celluar networks." Proc. Natl. Acad. Sci. USA **110**: 1128-1133.

Spirin, V. and L. A. Mirny (2003). "Protein complexes and functional modules in molecular networks." Proc. Natl. Acad. Sci. USA **100**(21): 12123-12128.

Stiart, J. M., E. Segal, et al. (2003). "A gene-coexpression network for global discovery of conserved genetic modules." Science **302**: 249-255.

Tornow, S. and H. W. Mewes (2003). "Functional modules by relating protein interaction networks and gene expression." Nucleic Acids Res. **31**: 6283-6289.

Yook, S. H., Z. N. Oltvai, et al. (2004). "Functional and topological characterization of protein interaction networks." Proteomics **4**: 928-942.