

Comparing the biological coherence of network clusters identified by different detection algorithms

DONG Dong^{1,2}, ZHOU Bing² & Jing-Dong J. HAN^{2†}

¹ Graduate School, College of Life Sciences, Beijing Normal University, Beijing 100875, China;

² Chinese Academy of Science Key Laboratory for Molecular Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

Protein-protein interaction networks serve to carry out basic molecular activity in the cell. Detecting the modular structures from the protein-protein interaction network is important for understanding the organization, function and dynamics of a biological system. In order to identify functional neighborhoods based on network topology, many network cluster identification algorithms have been developed. However, each algorithm might dissect a network from a different aspect and may provide different insight on the network partition. In order to objectively evaluate the performance of four commonly used cluster detection algorithms: molecular complex detection (MCODE), NetworkBlast, shortest-distance clustering (SDC) and Girvan-Newman (G-N) algorithm, we compared the biological coherence of the network clusters found by these algorithms through a uniform evaluation framework. Each algorithm was utilized to find network clusters in two different protein-protein interaction networks with various parameters. Comparison of the resulting network clusters indicates that clusters found by MCODE and SDC are of higher biological coherence than those by NetworkBlast and G-N algorithm.

network cluster detection algorithms, biological relevance, function entropy, protein-protein interaction network

Protein-protein interaction (PPI) networks are crucial for many biological functions^[1]. Almost every cellular process relies on transient or permanent physical bindings of proteins. Currently, many experimental and computational methods are available to detect or predict PPI in different organisms^[2–9]. The datasets thus generated have provided us a chance to examine the biological functions at the interactome network level.

The next challenge is to understand the biological functional significance of the PPI networks^[10]. A great amount of evidence has suggested that functional modules are cellular entities of complex biological systems^[11,12]. PPI network can be described as an undirected graph whose nodes represent proteins and whose edges correspond to pair-wise interactions. Network clusters are defined as sub-groups of PPI network whose proteins are closely linked and work together for the same cellular process. The connections within a cluster

are denser than the connections to the rest of the network. These highly connected network clusters often correspond to the basic molecular machineries in the cells. Proteins within each cluster are relatively homogeneous in function, and the clusters are relatively independent of each other^[13]. Identifying such functional neighborhoods from PPI network is essential for understanding the functions, organization, dynamics and evolution of biological systems.

Several methods have been applied to the PPI network in order to detect modular structures. These algorithms differ from each other in their definitions of network clusters and in the clusters detected. Furthermore, when different parameters are specified at the runtime,

Received February 15, 2007; accepted June 19, 2007

doi: 10.1007/s11434-007-0454-z

†Corresponding author (email: jdhan@genetics.ac.cn)

Supported by the National Natural Science Foundation of China (Grant No. 30588001)

the algorithm may also yield drastically different results. The users are often provided with little guidance in selecting these algorithms and the appropriate parameters. Therefore it is necessary to compare these algorithms in terms of their reliabilities and specificity. Brohee and van Helden^[14] recently evaluated four sophisticated network cluster detection algorithms: Markov clustering (MCL), restricted neighborhood search clustering (RNSC), super paramagnetic clustering (SPC), and molecular complex detection (MCODE). The first three algorithms are not commonly used and have no ready-to-use implementations. The evaluations are partially based on the topological separation/tightness of the clusters detected, which does not necessarily correspond to the best biological functional coherence. As a quantitative measurement for an algorithm's performance, they use coverage, accuracy and separation values based on known yeast protein complexes annotated by MIPS^[15]. MIPS annotated complexes only contain the protein complexes in yeast, which limits their algorithm comparison in yeast only and for protein complexes only. To address these limitations, we compared four more commonly used algorithms for biological networks based on the biological function homogeneity or coherence of the genes inside a cluster. If a network cluster is biologically relevant or reveals a certain biological context, the nodes in the cluster should be functionally coherent. In this study, using function entropy as a measurement for functional homogeneity, we evaluated the performance of MCODE^[16], NetworkBlast^[17], shortest-distance clustering (SDC)^[12] and Girvan-Newman (G-N) algorithms^[18].

MCODE^[16] identifies densely interconnected sub-graphs of a molecular interaction network. NetworkBlast^[17] is based on a probabilistic model. It compares the fitness of a sub-network to the observed structure versus its expected likelihood with random interactions. SDC^[12] forms a pair-wise shortest distance matrix among all the nodes in the network, and it assumes that each node in a graph has a unique profile of shortest distances to any other nodes. Network clusters can then be formed by hierarchical clustering based on the similarity of shortest distance profiles between a pair of nodes. The nodes inside a cluster have similar shortest distance profiles. The G-N algorithm^[18] is based on the assumption that the edges connecting communities have higher edge 'betweenness', which is defined the number of shortest paths between pairs of vertices that run

through an edge, and the underlying community structure of the graph can be revealed by sequentially cutting at the edge of the highest betweenness. To find the optimal cutting level, Girvan and Newman proposed a quantitative definition of network cluster delimitation metric, the Q value^[19], or the network modularity value. We followed Girvan and Newman and used the highest Q value to delimit the clusters. Although the highest Q value corresponds to the best cut topologically, it does not necessarily lead to the highest biological coherence of the clusters.

We calculated the node coverage and functional homogeneity of the network clusters using different input parameters of these algorithms.

1 Materials and methods

1.1 Datasets used in this study

PPI datasets generated by large-scale experiments and computational predictions generally contain false positives^[20] that may complicate the elucidation of biological process or cellular function^[21]. We therefore only used the *S. cerevisiae* CORE PPI dataset (DIPCORE)^[22] in the Database of Interacting Proteins (DIP) and the human PPI datasets curated by Human Protein Reference Database (HPRD)^[23]. DIPCORE data was generated by filtering large scale PPI data to improve their reliability^[24], and HPRD was manually curated from literatures by expert biologists. They are both considered as reliable PPI data. After removing self-connecting links, there are 2583 yeast proteins and 6285 interactions in DIPCORE, 7568 human proteins and 25072 interactions in HPRD dataset.

The functional category annotation was downloaded from MIPS^[15] (<ftp://ftpmips.gsf.de/yeast/>) and the original Gene Ontology (GO)^[25] was downloaded from NCBI (<ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/>). GO has three organizing principles: molecular function, biological process, and cellular components. In our analysis, we categorized genes by their biological process.

1.2 Automatically derive clusters from all-pair-shortest-distance matrix

The original shortest-distance clustering algorithm entails manually delimiting clusters from binary shortest distance matrix after hierarchical clustering. To improve the consistency of the results, we used a method to derive clusters from the matrix automatically. We first

scanned through the hierarchical clustering tree from the root of the tree, for pairs of sub-clusters under a certain cluster. For each gene in a sub-cluster of sub-cluster pair, we calculated the average shortest-distance of the gene toward all the genes in the other sub-cluster of sub-cluster pair. Then we ranked the genes based on their average shortest distance. And we calculated the spearman rank correlation coefficient of the two gene ranks in the two sub-clusters. We used the Spearman rank correlation coefficient as a criterion to determine at which level we should stop dividing the cluster into sub-clusters. That is, if the correlation coefficient is no longer larger than a certain cutoff ranging from 0.6 to 0.95, we accept the parent cluster as a network cluster.

1.3 Biological validity of the clusters

If a cluster is biologically relevant, the genes belonging to the same cluster should have similar biological functions^[10]. We evaluated the quality of the clusters by the degree of functional homogeneity among genes within a network cluster^[26,27] and used the function entropy to compute the degree of functional homogeneity. Function entropy is calculated to represent the frequency distribution of gene functions of a group of genes. As function annotations, we used MIPS functional category^[15] for yeast genes and GO for human genes. Function entropy of a cluster is calculated as the sum of the appearance frequencies of all function annotations in the cluster and multiplies the logarithms of those frequencies. It can be defined by the following formula: $-\sum F_i \log F_i$, where F_i is the appearance frequencies of the function annotation i , and $F_i = T_i / \sum_i^n T_i$ where T_i is the number of times that the function annotation appears in the clusters and n is the number of distinct function annotation present in the cluster. If the genes in the same cluster have consistent functions, the value of function entropy will be low, and it will be zero when the genes have only one function.

2 Results

Comparison of biological relevance of network clustering algorithms

We applied four network clustering algorithms to the DIPCORE and HPRD PPI network with different parameters. The clustering results were filtered and only the clusters whose sizes ranged from 5 to 300 were re-

tained.

We evaluated the quality of clusters based on the network (node) coverage and functional homogeneity of the clusters^[20]. The network *coverage* is the number of nodes retrieved by the clusters divided by the total number of nodes in the network. The *average function entropy* is the mean value of function entropy of the clusters obtained by an algorithm using a particular set of input parameters, and it represents the functional homogeneity of the genes in each cluster (see section 1).

Each algorithm has one or more parameters, and it may affect the clustering results. For each algorithm, we sampled different parameters to measure their effects. The complete results are listed in Supplementary Table 1 (at <http://www.SpringerLink.com>). The cluster results indeed differ very much for different input parameters of each algorithm. We therefore took advantage of this fact and obtained performance curves between accuracy and coverage based on shifting network coverage generated through different input parameter values. The values of different input parameters have different relationships with network coverage and average function entropy for each algorithm. For the MCODE algorithm, both the coverage and average function entropy increase coordinately with increasing 'vertex weight' values, which is an input parameter to determine the cutoff value for the vertex weight and reflects the density of the resulting network. However, the parameters almost have no simple direct relationship with network coverage and average function entropy in the other three algorithms (Supplementary Table 1 is available at <http://www.SpringerLink.com>). The values of parameters are labeled next to the data points in Figure 1 to provide some reference to the users as to what parameter values one may use for each algorithm.

There are a few input parameters for MCODE. It is reported that vertex weight parameter has the strongest effect on the network cluster results^[16]. The value of parameter is between 0 and 1.0. When the parameter value is larger than 0.5, the clusters tend to merge into a large component (data not shown). So we only selected 25 different values for the vertex weight parameters ranging from 0.01 to 0.5 with haircut and no fluff. Different values of this parameter give rise to different network coverage, and the average function entropy increases with the increasing coverage with both the DIPCORE and HPRD PPI networks (Figure 1(a) and (b);

Supplementary Table 1 is available at <http://www.SpringerLink.com>). Among the four algorithms, MCODE has both the lowest network coverage and the lowest average function entropy, suggesting that al-

though the algorithm does not retrieve as many nodes as the other algorithms, the nodes inside the clusters are functionally more homogeneous (Figure 2). Therefore this algorithm is very stringent at identifying function-

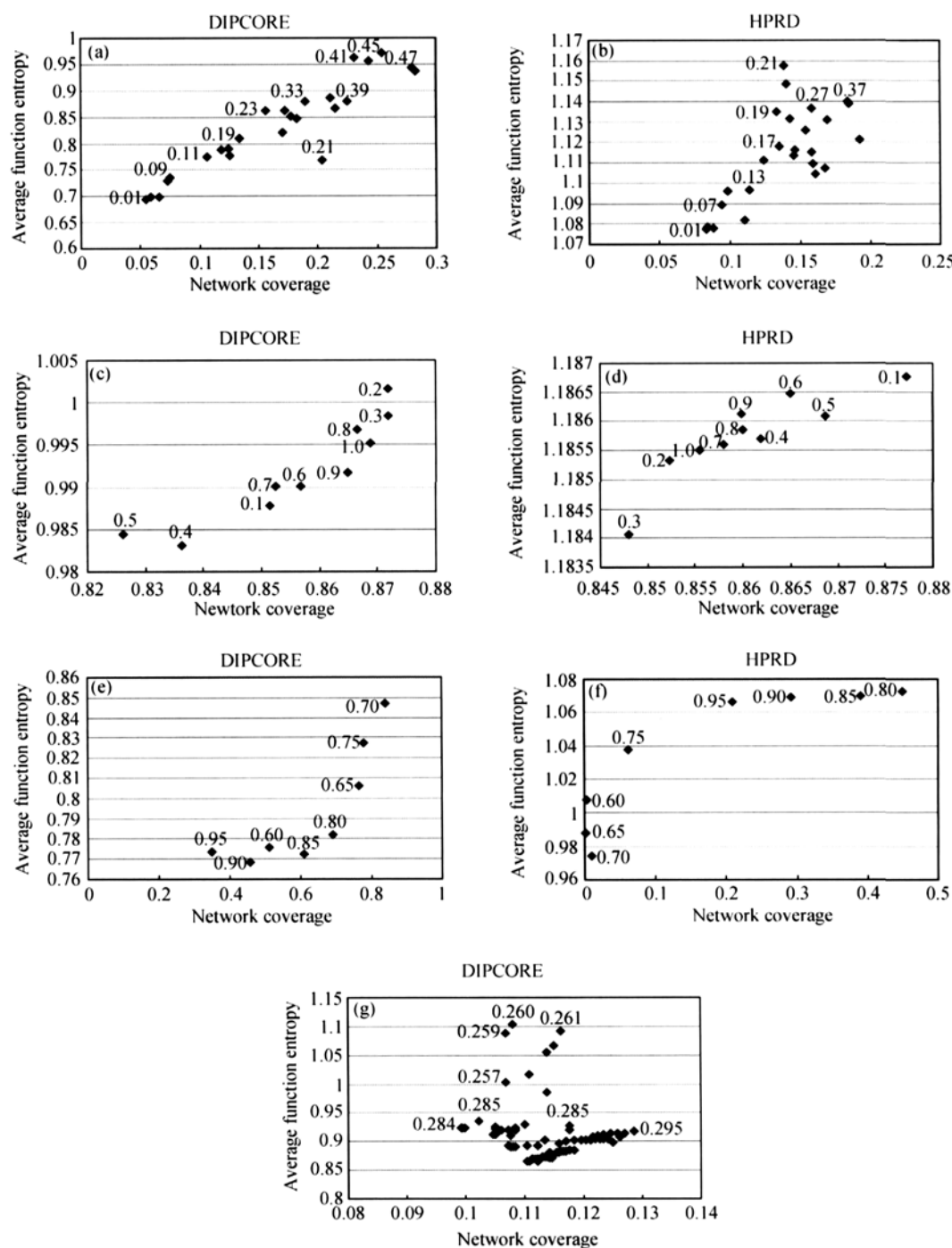


Figure 1 Impact of the different input parameters on network cluster finding results. (a) and (b), MCODE; (c) and (d), NetworkBlast; (e) and (f), SDC; (g), G-N algorithm. Each data point on the graph represents a cluster detection result under the specific input parameter value as indicated next to the data point. These parameters are the vertex weight parameter, significance threshold, pair-wise cluster Spearman correlation coefficient and the Q value for MCODE, NetworkBlast, SDC and G-N algorithms, respectively. The left and right panels display the clusters finding results in the DIPCORE and HPRD networks, respectively.

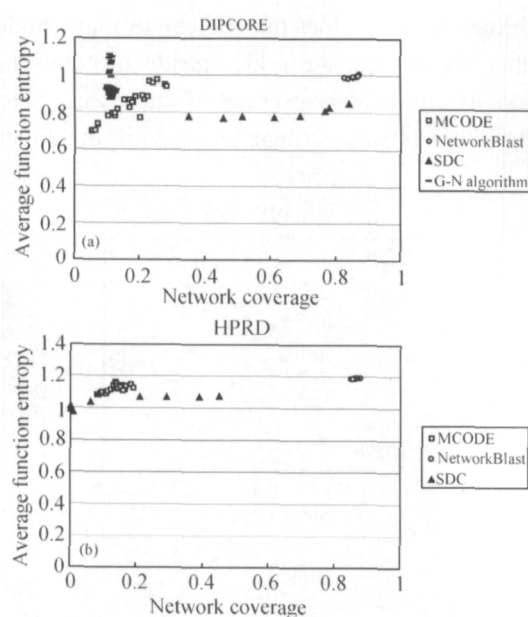


Figure 2 Comparison of the network clustering results. Lower average function entropy represents greater functional homogeneity, while the higher network coverage indicates higher detection sensitivity for an algorithm.

ally coherent clusters.

For the SDC algorithm, clusters can be manually dissected from the hierarchical clustering result. Here, we used our automatic method to delimit the cluster results with different thresholds ranging from Spearman rank correlation coefficient of 0.6 to 0.95 (see section 1 for details), because the network breaks into a few very large clusters when the threshold is below 0.6. Different thresholds for dissecting the clusters give rise to a large span of network coverage (Figure 1(e) and (f)). Concerted changes can be seen between the network coverage and the average function entropy, but the slope of increase is not very steep (Figure 2). Despite the scattered distribution of the network coverage, the average function entropy does not change dramatically. The clusters derived by this algorithm tend to have lower average function entropy, and higher functional homogeneity than those of the other three algorithms.

NetworkBlast has only one input parameter and it provides a threshold for significance level compared to random network configurations. The value of parameter ranges from 0 to 1.0. The clusters identified by NetworkBlast seem to be very stable toward different parameters (Figure 1(c) and (d); Supplementary Table 1 is available at <http://www.SpringerLink.com>), so we only

selected 10 different parameters. Among all the four algorithms, the results of NetworkBlast have the highest coverage and highest average function entropy (Figure 2). Therefore the clusters identified are not very stringent in biological functional homogeneity.

The G-N algorithm is very time-consuming when the input network is of a large size. It is practically impossible to finish for networks as large as the HPRD network. We therefore only performed analysis on the DIPCORE PPI network. Different Q values were chosen to arrive at different network coverage. Surprisingly, no distinct relationship between network coverage and average function entropy can be observed for network clusters found by this algorithm (Figure 1(g)). Among the four algorithms, the G-N algorithm gives the lowest network coverage and the highest average function entropy (Figure 2(a)). Therefore it is doubtful whether the assumption of the network partition principle that the highest 'betweenness' edge serves to connect different functional neighborhoods is biologically relevant. Based on structural properties, biological networks can be categorized into the influence networks and flow networks, which have very distinctive topological features^[28]. PPI networks are influence networks, where the interactions are 'influence-based'. The flow networks, such as metabolic networks, on the other hand, are networks with a specific variable, such as mass, conserved at each node. A modified G-N algorithm, which is based on node betweenness rather than edge betweenness, has been shown to perform well in decomposing metabolic networks^[29]. Apparently, when it is applied to PPI networks, the results are not very satisfactory, indicating that the algorithm might not be suitable for dissecting clusters in influence networks. Also given the unreasonably high computational demand, for example it takes 24 h to search a network of 2583 nodes and 6285 edges on a 3.2 GHz dual-processor Linux server, this algorithm is not appropriate for large PPI network.

3 Discussion

In this study, we compared the biological significance of four network cluster detection algorithms. Our evaluation is based on the functional homogeneity and the network coverage of the clusters to reveal the quality of the results. The results show that average function entropy is a good estimator for biological function homogeneity although it is slightly dependent on the number

of proteins in a cluster. Some algorithms may trim nodes from the final clusters, producing many small clusters and affecting the functional entropy values. The incompleteness of output clusters may also decrease the sensitivity, as measured by the network node coverage.

Our analysis revealed that different input parameters have different impacts on the final cluster results and the functional homogeneity of clusters. Although in general and as expected, an increase in sensitivity (network coverage) of an algorithm is accompanied by a decrease in accuracy of the clusters (a decrease in function homogeneity as indicated by an increase in function category entropy). G-N algorithm seems to be an exception. This may have something to do with the particular input parameter we chose. As we did not survey all the possible ranges of the parameter, we cannot be sure that this exception is only a coincidence for the chosen parameters. Furthermore, the difference in the biological significance evaluated by our method is based on the average performance of the network clusters using the known functional annotations. This does not rule out the possibility that some of the clusters detected by the worst performing algorithm are more useful than the clusters detected by the best performing algorithm under a specific biological context. For example, the current functional annotations might be biased for stable complexes, whereas transient complexes are hard to annotate as similar functions or they may represent more diverse functions.

Among the four commonly used approaches, MCODE focuses on the extraction of densely connected sub-graphs from PPI network, Shortest-distance clustering (SDC) uses the conventional clustering analysis based on shortest interaction distances to decompose the network into functional units. MCODE and SDC can yield highly connected clusters containing the proteins that form a single multi-molecular machine or part of it. Brohee et al.^[14] also evaluated the MCODE algorithm. Their result shows that MCODE finds a smaller number of clusters which is in agreement with its lower coverage we observed. However, they mainly compared an algorithm's performance based on the cluster separation and its agreement to the known protein complexes. Our results show that clusters identified by MCODE have

higher functional coherence than the other three algorithms. NetworkBlast algorithm identifies unexpectedly tightly connected network clusters which may correspond not only to protein complexes, but also to biological pathways or processes. It always yields larger clusters than the other three algorithms and a concurrent decrease in sensitivity; G-N algorithm or edge 'betweenness' dissects a network into modules by sequentially cutting at the edge that supports the most network traffic flow. Our results indicate that modules found by the G-N algorithm have the lowest coverage and the highest average function entropy. To delimit the clusters, it splits the clustering dendrogram from top to bottom at the optimal split-point where the quality of the communities, estimated by Q value, is the highest. Due to the splitting strategy on the clustering dendrogram, the G-N algorithm neglects the uniformity of clusters and leads to a large variation in cluster sizes. As a result, network coverage also varies a lot among different splits. Although we have used the Q value developed by the same group to find the optimal module separation boundary, currently we still cannot rule out the possibility that the algorithm may perform better under different cluster delimitation strategies.

In order to make sure that the evaluation results are not specific to a particular dataset, we performed the same analysis on both yeast and human PPI network (due to computational constraints, G-N algorithm was only performed on the much smaller yeast PPI network). The coverage and average function entropy of each algorithm are very similar across the two different datasets (Figure 2). This indicates that our evaluation methods are not biased for any particular dataset.

In summary, this study presents an objective evaluation of the network cluster detection algorithms from a perspective of biological function coherence. The evaluation results provide biological researchers useful guidance for selecting network clustering algorithms and input parameters to find network clusters with important biological significance.

The authors thank Zhang Qingpeng and Xia Kai for carefully reading the manuscript and the four anonymous reviewers for invaluable suggestions.

- 1 Hartwell L H, Hopfield J J, Leibler S, et al. From molecular to modular cell biology. *Nature*, 1999, 402: 47–52
- 2 Gavin A C, Bosche M, Krause R, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*,

2002, 415: 141–147

- 3 Giot L, Bader J S, Brouwer C, et al. A protein interaction map of *Drosophila melanogaster*. *Science*, 2003, 302: 1727–1736
- 4 Ho Y, Gruhler A, Heilbut A, et al. Systematic identification of protein

- complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 2002, 415: 180–183
- 5 Ito T, Chiba T, Ozawa R, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA*, 2001, 98: 4569–4574
- 6 Li S, Armstrong C M, Bertin N, et al. A map of the interactome network of the metazoan *C. elegans*. *Science*, 2004, 303: 540–543
- 7 Rual J F, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 2005, 437: 1173–1178
- 8 Stelzl U, Worm U, Lalowski M, et al. A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 2005, 122: 957–968
- 9 Uetz P, Giot L, Cagney G, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 2000, 403: 623–627
- 10 Barabasi A L, Oltvai Z N. Network biology: Understanding the cell's functional organization. *Nat Rev Genet*, 2004, 5: 101–113
- 11 Ravasz E, Barabasi A L. Hierarchical organization in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 2003, 67: 026112
- 12 Rives A W, Galitski T. Modular organization of cellular networks. *Proc Natl Acad Sci USA*, 2003, 100: 1128–1133
- 13 Spirin V, Mirny L. A Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA*, 2003, 100: 12123–12128
- 14 Brohee S, van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 2006, 7: 488
- 15 Mewes H W, Frishman D, Guldener U, et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 2002, 30: 31–34
- 16 Bader G D, Hogue C W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 2003, 4: 2
- 17 Sharan R, Suthram S, Kelley R M, et al. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA*, 2005, 102: 1974–1979
- 18 Girvan M, Newman M E. Community structure in social and biological networks. *Proc Natl Acad Sci USA*, 2002, 99: 7821–7826
- 19 Newman M E, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 2004, 69: 026113
- 20 von Mering C, Krause R, Snel B, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 2002, 417: 399–403
- 21 Han J D, Dupuy D, Bertin N, et al. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol*, 2005, 23: 839–844
- 22 Xenarios I, Salwinski L, Duan X J, et al. DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 2002, 30: 303–305
- 23 Peri S, Navarro J D, Amanchy R, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 2003, 13: 2363–2371
- 24 Deane C M, Salwinski L, Xenarios I, et al. Protein interactions: Two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 2002, 1: 349–356
- 25 Ashburner M, Ball C A, Blake J A, et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 2000, 25: 25–29
- 26 Han J D, Bertin N, Hao T, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 2004, 430: 88–93
- 27 Snel B, Bork P, Huynen M A. The identification of functional modules from the genomic association of genes. *Proc Natl Acad Sci USA*, 2002, 99: 5890–5895
- 28 Mahadevan R, Palsson B O. Properties of metabolic networks: Structure versus function. *Biophys J*, 2005, 88: L07–09
- 29 Holme P, Huss M, Jeong H. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 2003, 19: 532–538