Available online at www.sciencedirect.com
SCIENCE DIRECT
Biochemical and Biophysical Research Communications 345 (2006) 302–309
BBRC
www.elsevier.com/locate/ybbrc

Integrated analysis of multiple data sources reveals modular structure of biological networks

Hongchao Lu [b,c,1], Baochen Shi [a,c,1], Gaowei Wu [b], Yong Zhang [a,c], Xiaopeng Zhu [a,c], Zhihua Zhang [a,c], Changning Liu [b,c], Yi Zhao [b], Tao Wu [a,c], Jie Wang [a,c], Runsheng Chen [a,b,*]

[a] *Bioinformatics Laboratory and National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, PR China*
[b] *Bioinformatics Research Group, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, PR China*
[c] *Graduate School of the Chinese Academy of Sciences, Beijing 100039, PR China*

Received 2 April 2006
Available online 27 April 2006

Abstract

It has been a challenging task to integrate high-throughput data into investigations of the systematic and dynamic organization of biological networks. Here, we presented a simple hierarchical clustering algorithm that goes a long way to achieve this aim. Our method effectively reveals the modular structure of the yeast protein–protein interaction network and distinguishes protein complexes from functional modules by integrating high-throughput protein–protein interaction data with the added subcellular localization and expression profile data. Furthermore, we take advantage of the detected modules to provide a reliably functional context for the uncharacterized components within modules. On the other hand, the integration of various protein–protein association information makes our method robust to false-positives, especially for derived protein complexes. More importantly, this simple method can be extended naturally to other types of data fusion and provides a framework for the study of more comprehensive properties of the biological network and other forms of complex networks.

Keywords: Integrated analysis of multiple data sources; Hierarchical clustering algorithm; Protein–protein interaction network; Modular structure of biological networks

In the post-genomic era, with the ever-increasing amounts of high-throughput data available [1–7], more attention is being paid to how to comprehensively integrate these data to understand the functions of individual proteins as well as to further investigate the systematic and dynamic organization of biological networks. Some progress has been made in predicting protein–protein interactions and annotating uncharacterized proteins using the Bayesian methods [8,9]. In this study, a simple hierarchical clustering algorithm is employed to uncover the modular structure of the network by integrating various protein–protein association data. By ''network'' is here meant the protein–protein interaction (PPI) network, in which proteins are depicted as vertices and interactions as undirected edges.

Important statistical characteristics of PPI networks have been discussed in earlier studies, such as topological properties [10,11], motifs [12], and the modular architecture in which proteins have more interactions within modules than with the rest of the network [13]. The analysis of modules is useful for reduction of network complexity and extraction of biological information from the network, and it has also made some progress in studying the

* Corresponding author. Fax: +8610 64877837.
E-mail address: crs@sun5.ibp.ac.cn (R. Chen).
[1] These authors contributed equally to this work.

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.04.088

evolution of networks [14]. The module can be understood as a separated substructure of a network, members in the same module having strong functional associations with each other [13]. The reliably functional predictions can be provided for uncharacterized protein components in terms of the functional coherency within the derived modules [15–17].

In general, two types of modules are distinguished based on different biological properties [16]. Protein complexes are defined as groups of proteins that bind each other at the same time and place, forming a stably multimolecular machine, such as a transcription factor complex. In a more generic way, the functional module often contains protein complexes, but conceptually it consists of proteins that participate in a common cellular process (such as a biochemical pathway, etc.). With different dynamic requirement for protein–protein dosage relationship, the protein complex and functional module generally have different expression profiles. However, without temporal and spatial information, it is difficult to distinguish protein complexes from dynamic functional modules from the topological structure of the PPI network alone [16]. Integrating the PPI network with the added spatio-temporal information seems one of the solutions to change the situation.

Unsupervised clustering methods have been recently developed and successfully used in the analysis of the expression patterns [6] and the modular structure of the budding yeast PPI network [15–20]. In this paper, we present a simple hierarchical clustering method that integrates multiple data sources. The method is developed from our earlier approach in which the adjacency matrix of the PPI network is employed as the similarity matrix for the clustering [20].

In theory, interacting proteins tend to be localized in the same subcellular compartments [21] and are more likely to have similar expression profiles [22]. We also observed these characteristics in the yeast PPI networks, especially significant in protein pairs of protein complexes (see Additional Fig. 1). By integrating the protein–protein interaction data with subcellular localization data and expression profile data, our clustering method effectively reveals the information-rich modular structure of the network and distinguishes protein complexes from functional modules (Fig. 1). We provide a functional context for the uncharacterized components within functional modules with high confidence based on the functional coherency of proteins within functional modules. Moreover, our clustering result is also robust to false-positives in experimental data by the integration of the heterogeneous data, especially for derived protein complexes. What is more valuable is that this simple method can be extended naturally to other types of biological data fusion to study more comprehensive characteristics of the biological network, as well as in analysis of other forms of complex networks, such as technological (e.g., internet, world-wide web), ecological, and social networks.

Methods

Data source. We assembled a PPI network containing 4537 budding yeast (*Saccharomyces cerevisiae*) proteins and 13,344 physical interactions obtained with the two-hybrid assay [2], HMS-PCI [3], and TAP methods [4]. The global subcellular localization data in which proteins are classified into 22 distinct subcellular localization categories [5] and integrated expression profile data containing 643 condition/time points [7] were included to provide spatio-temporal information in the analysis of the modular architecture of yeast PPI network.

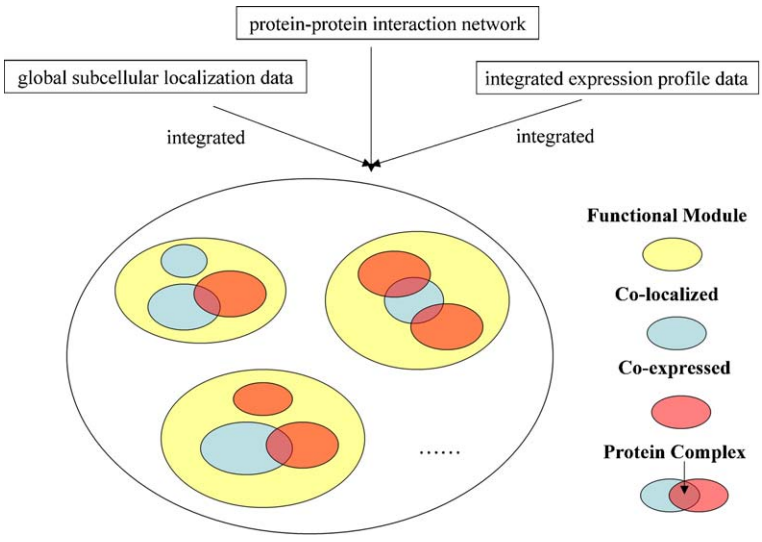

Fig. 1. Integrated analysis reveals modular structure of the PPI network with the spatial and temporal information included. The figure is a general outline that visualizes the spatio-temporal information within the modules, and clearly reveals the relationship between protein complexes and functional modules by integrating the protein–protein interaction data with subcellular localization data and expression profile data. That is, a functional module consists of proteins that participate in a common biological process at the same time or place; in contrast, protein complexes are the intersections of co-localized and co-expressed protein groups that are usually included in the functional modules.

Hierarchical clustering algorithm. The PPI network can be represented as a bi-directed graph $G(V,E)$, in which proteins are depicted as vertices and the interactions between them as edges. Let the adjacency matrix denote the interaction information, where $A=(a_{ij})$, $a_{ij}=1$ when there is an edge between vertices i and j, and $a_{ij}=0$ otherwise. Then, the adjacency matrix A of the PPI network is defined as:

$$A=(a_{ij}), a_{ij}=\begin{cases}1 & \langle i,j\rangle \text{ have an interaction}\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

Other protein–protein association information can be depicted in a similar way. For the spatial consistency information of the subcellular localization, the localization matrix L, it follows that:

$$L=(L_{ij}), L_{ij}=\begin{cases}1 & \langle i,j\rangle \text{ is in the same subcellular compartment}\\ 0 & \text{otherwise}\end{cases} \tag{2}$$

For the gene expression profile information, let $E=(e_{ij})$ be the matrix, where the Pearson correlation coefficient e_{ij} describes the temporal consistency of the expression levels between proteins. The Pearson correlation coefficient takes the form:

$$E=(e_{ij}), e_{ij}=C_{\text{Pearson}}(i,j)=\sum_k \frac{(X_k^{(i)}-\overline{X}^{(i)})(X_k^{(j)}-\overline{X}^{(j)})}{\sqrt{\sum_p (X_p^{(i)}-\overline{X}^{(i)})^2\cdot\sum_q (X_q^{(j)}-\overline{X}^{(j)})^2}} \tag{3}$$

where $X_k^{(i)}$ is the primary data for protein i under condition k and $\overline{X^{(i)}}$ is set to the mean of observations on the protein i.

Average-linkage hierarchical clustering has been successfully used in the analysis of expression patterns [6] and the PPI network [20]. In this paper, we have developed an improved average-linkage hierarchical clustering method to integrate multiple data sources. The proximity between two protein groups M and N is defined as P_{MN}:

$$P_{MN}=D_{MN}(1+B_{MN}) \tag{4}$$

$$D_{MN}=\frac{1}{|M||N|}\sum_{m\in M,n\in N} a_{mn} \tag{5}$$

$$B_{MN}=\frac{1}{|M||N|}\sum_{m\in M,n\in N} b_{mn} \tag{6}$$

D_{MN} represents the linkage density between two protein groups in the PPI network with protein numbers $|M|$ and $|N|$, respectively. In this context, B_{MN} describes the similarity of other association information between protein groups, where the matrix b_{mn} is either L or E representing the added spatial and temporal information, respectively.

We integrated the PPI network with the added spatial information, temporal information or both. Thus, for the clustering, we defined the proximity in three different ways, as follows:

$$\text{ADJL}: P_{MN}=D_{MN}(1+L_{MN}) \tag{7}$$

$$\text{ADJE}: P_{MN}=D_{MN}(1+E_{MN}) \tag{8}$$

$$\text{ADJB}: P_{MN}=D_{MN}(1+L_{MN})(1+E_{MN}) \tag{9}$$

This agglomerative method was employed to generate a hierarchical clustering tree by the greedy algorithm. After adding the spatio-temporal information, these hierarchical clustering methods ensure that co-localized and co-expressed protein groups should be clustered first. The scale of the proximity value in the ADJL or ADJE methods ranges from zero to two, and from zero to four in ADJB method.

The F-measure of a complex. The reduced F-measure [23] was introduced to measure the correspondence between a branch in the clustering tree and MIPS (Munich Information Center for Protein Sequences) complex annotation [24]. Given a particular complex C_r of size n_r and a particular cluster S_i of size n_i, suppose n_{ri} proteins in the cluster S_i belong to C_r, then the F-measure between this complex and cluster is defined as:

$$F(C_r,S_i)=\frac{2n_{ri}}{n_r+n_i} \tag{10}$$

where an F-measure close to 1 means that the complex and the cluster have a similar set of proteins. The F-measure of complex C_r is the maximum F-measure value attained by any branch in the hierarchical clustering tree. That is,

$$F\text{-measure}(C_r)=\max_{S\in T} F(C_r,S_i) \tag{11}$$

The P-value of a functional module. For each cluster we used the hypergeometric distribution *P-value* to model the probability of by chance observing at least k proteins in a cluster size n belonging to a category containing C proteins in a total genome size of G proteins, such that the *P-value* is given by

$$P=1-\sum_{i=0}^{k-1}\frac{\binom{C}{i}\binom{G-C}{n-i}}{\binom{G}{n}} \tag{12}$$

The above test measures whether a cluster is more enriched with proteins from a particular category than that would be expected by chance. It is taken to be a significant biologically functional module if the *P*-value is less than 0.001 based on the Bonferroni correction for multiple independent hypotheses ($0.01/N$, where N is the number of categories in the MIPS annotations, herein $N=12$) [21,24,25]. We manually assigned each derived module the functional category with the lowest *P-value*, and used this to evaluate different clustering methods and annotate uncharacterized proteins within modules.

Linkage density of a cluster in the network. The linkage density (LD) of a cluster in a particular network can be defined as:

$$LD=\frac{2\cdot\sum_{i,j} m_{ij}}{N\cdot(N-1)} \tag{13}$$

where m_{ij} is the protein–protein association information matrix, i and j denote the vertexes in the cluster, and N is the size of the cluster. For the PPI network, m_{ij} is just the adjacency matrix A, and the linkage density shows the tendency for a cluster to form a quasi-clique. Similarly, in the co-location and the co-expression networks, m_{ij} is defined as matrix L and matrix E, respectively; the linkage densities of these indicating the fractions of co-located or co-expressed components of the clusters.

Results

We applied our method to the yeast protein–protein interaction data set, combining the global subcellular localization data [5] and the integrated expression profile data [7] as the spatial and temporal information of the PPI network (see Methods). The clustering result was displayed using TreeView [6] and our PINC [20] with functional annotations [21,24] and spatio-temporal association information (see Additional Files).

The ADJB tree performs better than four previously reported methods

In order to illustrate the advantages of the methods, we also applied four previously reported methods including those of methods (Rives and Galitski [17], Samanta and Liang [15], Brun et al. [18], and our earlier ADJW method [20]) on the PPI network, and compared them to our ADJL, ADJE, and ADJB methods according to (a) biological information enrichment and (b) correspondence between clustered branches and experimental protein complexes.

Biological information enrichment is an important criterion for evaluation of a clustering method. Applying the MIPS functional annotations [21,24], the *P-value* was introduced to measure these enrichments in a clustering

tree. We calculated the *P-value* of each branch in the seven trees. Compared to other trees, biological information is more enriched through clustering by ADJB and ADJL than through clustering by other methods (Fig. 2A).

Second, the *F*-measure was used to measure the correspondence between the clustering branch and the experimental protein complex. We calculated the coincidences between 260 known complexes from the MIPS database [24] and the branches in the above seven trees. The *F*-measure of a particular complex is the maximum *F*-measure value attained to any branch in the hierarchical clustering tree. Measured by this criterion, the ADJB tree performs better than any of the other six clustering trees (Fig. 2B).

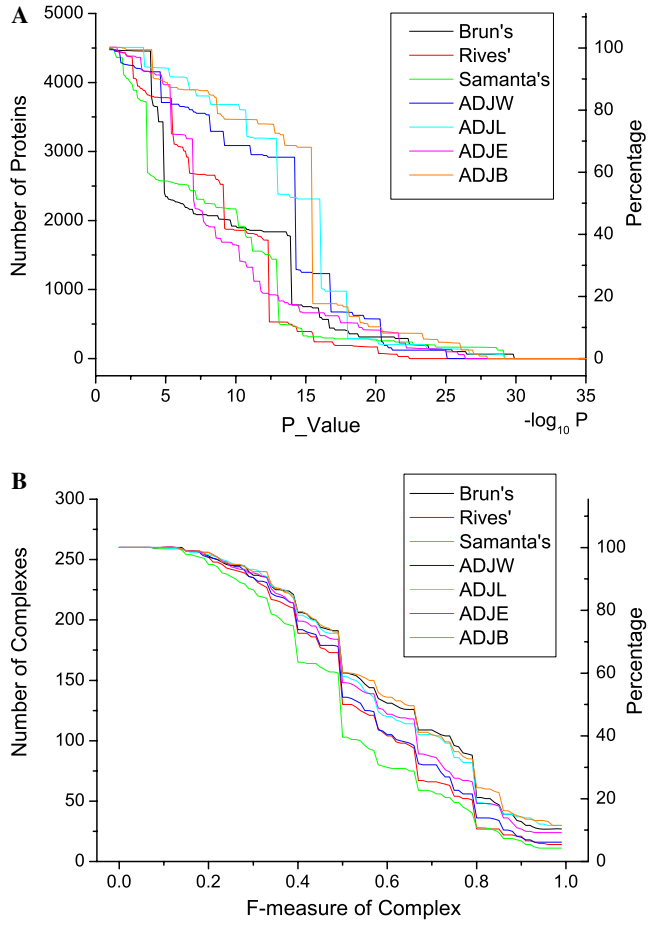

Fig. 2. Comparisons of the clustering methods. (A) Comparison with respect to the enrichment of biological information in the hierarchical clustering trees. Applying MIPS functional annotations [21,24], the *P-value* was introduced to measure the biological information enrichment in a clustering tree. We calculated the *P-value* of each branch in the trees. On the *Y*-axis we plot the number of proteins in the branches, which achieve the *P-value* in the corresponding *X*-axis value or less. (B) Comparison with respect to discovered complexes. The *F*-measure values were calculated for a list of 260 protein complexes from MIPS [24]. The *F*-measure of a particular complex is the maximum *F*-measure value attained at any branch in the hierarchical clustering tree. The *Y*-axis presents the number of protein complexes, which achieve the *F*-measure in the corresponding *X*-axis value or more (see Methods).

The ADJB method better highlights protein complexes from the PPI network

Generally speaking, it is difficult to distinguish protein complexes from functional modules in the hierarchical clustering tree. However, as integrating both spatial and temporal information, the ADJB method better highlights protein complexes from the PPI network than methods without this amount of information, and ensures that the co-localized and co-expressed protein groups are clustered first in the agglomerative clustering process. In the ADJB tree, we take the proximity P_{MN} as the parameter to delimit protein complexes. Using this method the spatial and temporal consistency of the proteins of clusters degrades gradually as the P_{MN} value falls during agglomerative clustering. We experimented with a range of values for P_{MN} whilst observing the *F*-measure values of the MIPS experimental complexes [24] in the ADJB tree. Fig. 3 shows that, for experimentally verified complexes, the highest number of best-matched clusters appears when P_{MN} approaches 1. This means that the protein complex information is more enriched when taking 1 as the optimal value. Thus, we used 1 as the lower bound for delimitation of complex-like clusters in the ADJB tree. (A complete list of the above co-localized and co-expressed clusters is found in Additional Table 1, which also includes some related information, such as linkage density information of clusters and MIPS annotations [24]). We hope this information may be helpful to experimentalists for identifying possible unknown or poorly studied complexes.

To test the biological function coherence of proteins in the derived functional modules, we computed the fractions of functional categories in each derived functional module associated with MIPS annotations [21,24]. We found that in 71% of functional modules, more than 50% of annotated constituent proteins belong to the same functional category as we had assigned manually to the module with the lowest

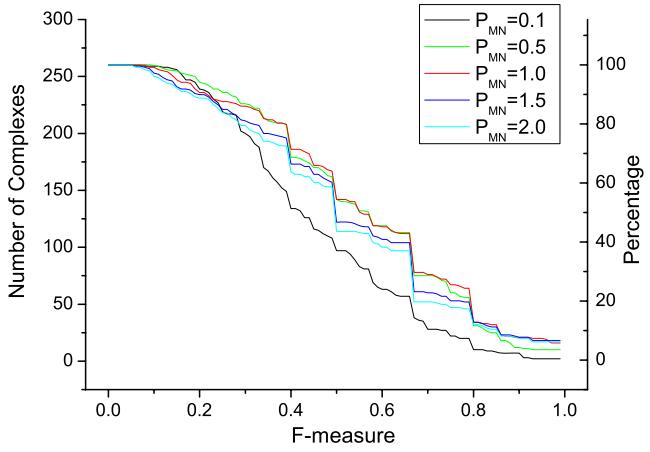

Fig. 3. Choice of P_{MN} value to delimit protein complexes in the ADJB tree. We tested with a range of values for P_{MN} and observed the *F*-measure values of the MIPS experimental complexes [24] in the ADJB tree. The figure shows that for experimental complexes, the highest number of well-matched clusters appears when P_{MN} approaches 1.

P-value. This indicates that most of proteins in the modules are involved in the same biological process. Using the functional coherency to analyze the derived modules, we predicted functions for 108 proteins annotated in the latest update of MIPS database [24] (June 20, 2005 release), obtaining 55% with very good agreement with the MIPS annotation (Additional Table 2).

The ADJB method is more robust to false-positives in all perturbed networks

The realization of computational strategies usually depends on the limitations of experimental data, and in high-throughput interaction screens is commonly a high rate of false positives [21]. To investigate the extent to which false positives affected the clustering tree, we randomly removed 10–90% of the links, added 10–100% links, and rewired 10–100% of the links without changing the power-law nature in the PPI network. We evaluated the robustness of our ADJB method by comparing the average *F*-measure values of 260 known protein complexes [24] in the original and the perturbed networks. We likewise compared the numbers of computational modules in the original and the perturbed networks. As Fig. 4A shows, noise in the form of random addition of interactions has less deteriorating effect than other forms of perturbations. Even after adding links up to two times the original number, our ADJB method still does well in retaining co-localized and co-expressed protein groups, thus, clustering of a PPI network added temporal and spatial data effectively resisted the effects of false positives. The numbers of computational modules went down in the noisy networks, but even after increasing the number of interactions by 100% we were still able to detect more than 50% of the functional modules found in the original network (Fig. 4B).

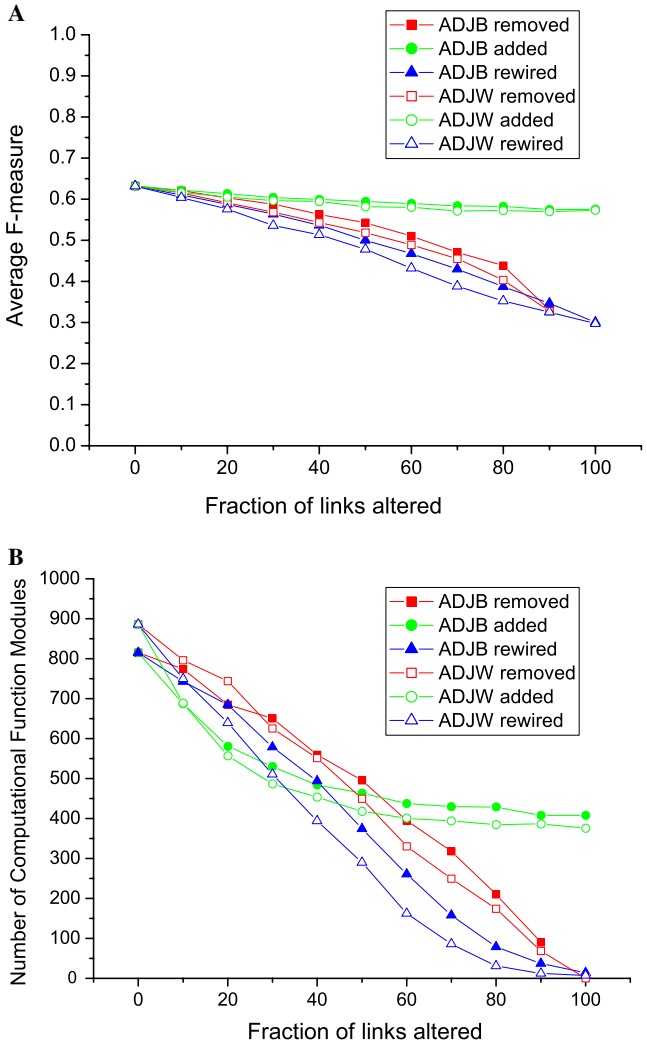

Fig. 4. Robustness test of the ADJB and ADJW methods. To investigate the extent to which false positives affected the clustering tree, we randomly removed 10–90% of the links, added 10–100% links, and rewired 10–100% of the links without changing the power-law nature in the PPI network. Each perturbation was repeated 10 times. (A) The average *F*-measure values of 260 known protein complexes [24] are compared to evaluate the extent to which the ADJB method retains co-localized and co-expressed clusters in the original with perturbed networks. (B) We observed the numbers of computational modules in the original and perturbed networks. Compared to our earlier ADJW method [20], the ADJB method is more robust to false-positives in all perturbed networks.

More importantly, compared to our earlier ADJW method [20], the ADJB method is more robust to false-positives in all perturbed networks (Fig. 4). The increased robustness of the ADJB method to false-positive noise arises from the integration of heterogeneous protein–protein association information to the PPI network.

On the other hand, nearly no functional modules are found when 100% of links are rewired, as this perturbed network can be considered as a random network of the same power-law distribution as the original network. Compared to the number detected in the origin network, this further substantiates that the functional modules derived in the ADJB tree are statistically significant and biologically meaningful.

Discussion

The functional module is a set of proteins that take part in a common process, but may carry out different functions within that process. A detailed inspection of the derived functional modules revealed many cases in which our method isolated known biological processes from the PPI network. Fig. 5A represents a module consisting of 21 yeast proteins and 48 physical interactions, in which all components but one uncharacterized protein (YPR084W) are localized in the mitochondrion. Since all annotated components are mitochondrial ribosomal proteins, we can infer a functional context for the uncharacterized one within the module. Of six previously unknown proteins that were annotated in the latest update of MIPS (June 20, 2005 release) [24], all except YDR036C are indeed also mitochondrial ribosomal proteins. Nuclear genes encode most of the mitochondrial ribosomal proteins, which are transported into the mitochondrion from the cytoplasm. YDR036C is a protein

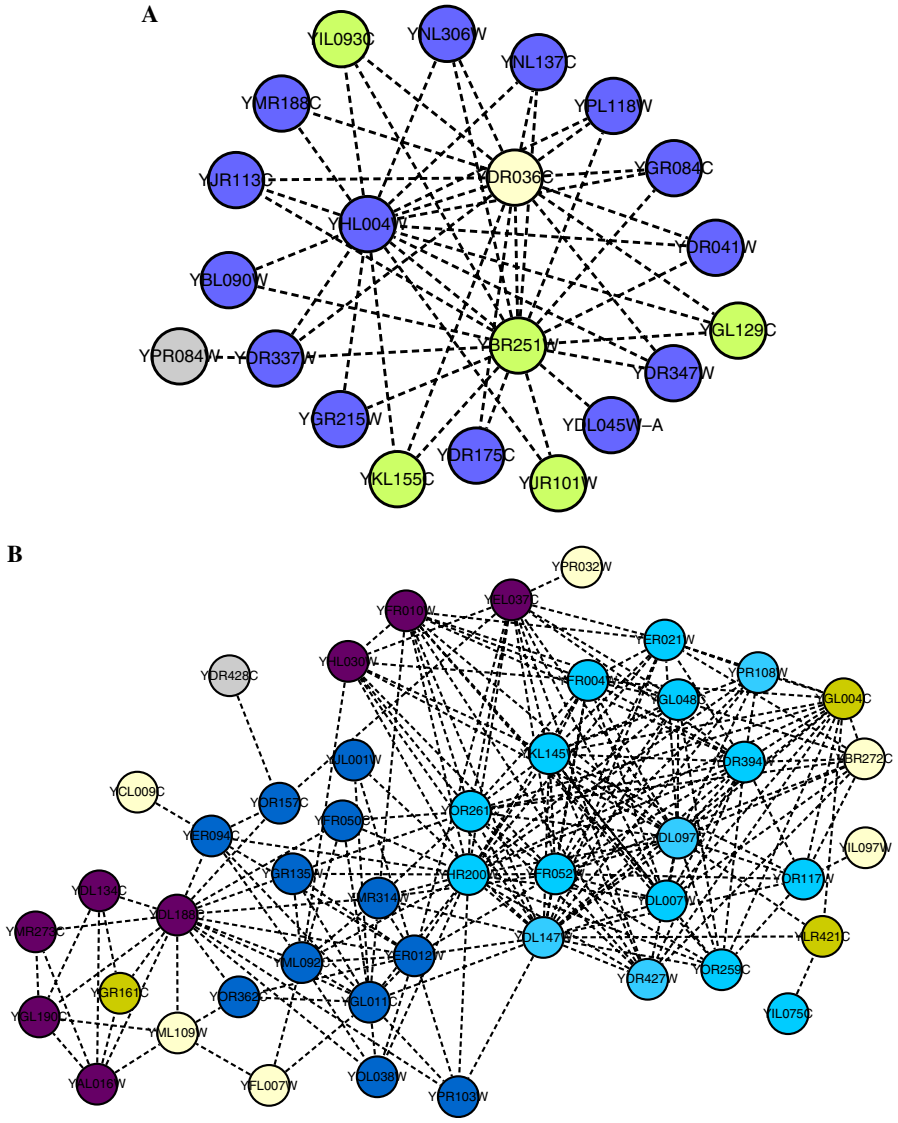

Fig. 5. Examples of discovered functional modules. (A) A module containing 21 yeast proteins, all except the uncharacterized protein YPR084W (gray node) localized in the mitochondrion. All previously functionally known protein (blue vertices) are mitochondrial ribosomal proteins, as are the five most recently annotated proteins [24] (green vertices). The YDR036C (yellow white node) participates in transport across the mitochondrial membranes and acts as an assistant factor in the module. (B) Bigger module consisting of 46 yeast proteins encompasses almost the entire 26S proteasome complex. The proteasomal proteins (different shades of blue) and an additional eight proteins (purple vertices) represent 90% of known components of the complete 26S proteasome. These proteins are all involved in a common process of protein degradation and modification. The 26S proteasome is made up of the 20S proteasome (navy blue vertices) and the 19/22S regulator (bright blue vertices), and all their components appear in the same subcellular compartments [24]. Fig. 5 was constructed using Cytoscape [27]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this paper.)

that is involved in the endocytosis process, and therefore has interactions with most of the mitochondrial ribosomal proteins. From this perspective, YDR036C acts as an interrelated assistant factor that participates in the module and can serve as an example of an unsuccessful prediction by the ADJB method.

Another module depicted in Fig. 5B is a relatively bigger one that consists of 213 physical links involving 46 yeast proteins, which contains almost the entire 26S proteasome complex. The proteasome and additional eight proteins, in total 90% of known components of the module, are all involved in protein degradation and modification. The 26S proteasome is made up of the 20S proteasome and the 19/22S regulator, and all their components appear in the same subcellular compartments (the endoplasmic reticulum and the cell nucleus). Besides identical subcellular localization, the proteins of the proteasome are more co-expressed than the module average [26]. As Fig. 1 outlined, the functional module consists of proteins that participate in a common biological process at the same time or place; in contrast, protein complexes are the intersections of co-localized and co-expressed protein groups that are usually included in the functional modules.

In conclusion, our clustering method integrating localization data and expression profile information effectively reveals the modular architecture of the PPI network, distinguishes protein complexes from functional modules. The method also provides a reliably functional context useful for prediction of function for uncharacterized protein components of modules, and the integration of heterogeneous data makes the method robust to false-positives. More importantly, such a simple method can easily be extended to include more types of biological data, e.g., genetic interactions, protein–DNA and protein–RNA data. Integration of physical interactions networks with orthology data may also provide insights into the origin of cellular modularity. Moreover, we believe the real power of the method will materialize in multiple data source analysis of other forms of complex networks, such as technological, ecological, and social networks.

Acknowledgments

We thank Doctor Geir Skogerbø (Visiting Scientist, Chinese Academy of Sciences, Beijing) for providing helpful discussion and critical review of our manuscript. This work was supported by the National High Technology Development Program of China under Grant No. 2002AA231031, National Key Basic Research and Development Program 973 under Grant Nos. 2002CB713805 and 2003CB715907, the Chinese Academy of Sciences Grant No. KSCX2-2-27, National Sciences Foundation of China Grant Nos. 39890070, 30500104, 30570393 and 60496320, opening task of Shanghai Key Laboratory of Intelligent Information Processing Fudan University No. IIPL-04-001 and Beijing Science and Technology Commission Grant No. H010210010113.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2006.04.088.

References

[1] P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, J.M. Rothberg, A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*, Nature 403 (2000) 623–627.

[2] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, Y. Sakaki, A comprehensive two-hybrid analysis to explore the yeast protein interactome, Proc. Natl. Acad. Sci. USA 98 (2001) 4569–4574.

[3] Y. Ho, A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A.R. Willems, H. Sassi, P.A. Nielsen, K.J. Rasmussen, J.R. Andersen, L.E. Johansen, L.H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B.D. Sorensen, J. Matthiesen, R.C. Hendrickson, F. Gleeson, T. Pawson, M.F. Moran, D. Durocher, M. Mann, C.W. Hogue, D. Figeys, M. Tyers, Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry, Nature 415 (2002) 180–183.

[4] A.C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon, C.M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.A. Heurtier, R.R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, G. Superti-Furga, Functional organization of the yeast proteome by systematic analysis of protein complexes, Nature 415 (2002) 141–147.

[5] W.K. Huh, J.V. Falvo, L.C. Gerke, A.S. Carroll, R.W. Howson, J.S. Weissman, E.K. O'Shea, Global analysis of protein localization in budding yeast, Nature 425 (2003) 686–691.

[6] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA 95 (1998) 14863–14868.

[7] J.M. Stuart, E. Segal, D. Koller, S.K. Kim, A gene-coexpression network for global discovery of conserved genetic modules, Science 302 (2003) 249–255.

[8] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, M. Gerstein, A Bayesian networks approach for predicting protein–protein interactions from genomic data, Science 302 (2003) 449–453.

[9] Y. Chen, D. Xu, Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*, Nucleic Acids Res. 32 (2004) 6414–6424.

[10] H. Jeong, S.P. Mason, A.L. Barabasi, Z.N. Oltvai, Lethality and centrality in protein networks, Nature 411 (2001) 41–42.

[11] S. Maslov, K. Sneppen, Specificity and stability in topology of protein networks, Science 296 (2002) 910–913.

[12] S. Wuchty, Z.N. Oltvai, A.L. Barabasi, Evolutionary conservation of motif constituents in the yeast protein interaction network, Nat. Genet. 35 (2003) 176–179.

[13] L.H. Hartwell, J.J. Hopfield, S. Leibler, A.W. Murray, From molecular to modular cell biology, Nature 402 (1999) C47–C52.

[14] H. Qin, H.H. Lu, W.B. Wu, W.H. Li, Evolution of the yeast protein interaction network, Proc. Natl. Acad. Sci. USA 100 (2003) 12820–12824.

[15] M.P. Samanta, S. Liang, Predicting protein functions from redundancies in large-scale protein interaction networks, Proc. Natl. Acad. Sci. USA 100 (2003) 12579–12583.

[16] V. Spirin, L.A. Mirny, Protein complexes and functional modules in molecular networks, Proc. Natl. Acad. Sci. USA 100 (2003) 12123–12128.

[17] A.W. Rives, T. Galitski, Modular organization of cellular networks, Proc. Natl. Acad. Sci. USA 100 (2003) 1128–1133.

[18] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, B. Jacq, Functional classification of proteins for the prediction of cellular function from a protein–protein interaction network, Genome Biol. 5 (2003) R6.

[19] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, R. Chen, Topological structure analysis of the protein–protein interaction network in budding yeast, Nucleic Acids Res. 31 (2003) 2443–2450.

[20] H. Lu, X. Zhu, H. Liu, G. Skogerbo, J. Zhang, Y. Zhang, L. Cai, Y. Zhao, S. Sun, J. Xu, D. Bu, R. Chen, The interactome as a tree—an attempt to visualize the protein-protein interaction network in yeast, Nucleic Acids Res. 32 (2004) 4804–4811.

[21] C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, P. Bork, Comparative assessment of large-scale data sets of protein–protein interactions, Nature 417 (2002) 399–403.

[22] H. Ge, Z. Liu, G.M. Church, M. Vidal, Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*, Nat. Genet. 29 (2001) 482–486.

[23] Van Rijsbergen, Information Retrieval, second ed., Butterworths, London 1979. Available from: <http://www.dcs.gla.ac.uk/Keith/Preface.html>.
[24] H.W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, B. Weil, MIPS: a database for genomes and protein sequences, Nucleic Acids Res. 30 (2002) 31–34.
[25] L.F. Wu, T.R. Hughes, A.P. Davierwala, M.D. Robinson, R. Stoughton, S.J. Altschuler, Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters, Nat. Genet. 31 (2002) 255–265.
[26] R. Jansen, D. Greenbaum, M. Gerstein, Relating whole-genome expression data with protein–protein interactions, Genome Res. 12 (2002) 37–46.
[27] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.