Published online September 8, 2004

4804–4811 *Nucleic Acids Research, 2004, Vol. 32, No. 16*
doi:10.1093/nar/gkh814

The interactome as a tree—an attempt to visualize the protein–protein interaction network in yeast

Hongchao Lu[2], Xiaopeng Zhu[1], Haifeng Liu[2], Geir Skogerbø[2], Jingfen Zhang[2], Yong Zhang[1], Lun Cai[2], Yi Zhao[2], Shiwei Sun[2], Jingyi Xu[2], Dongbo Bu[2] and Runsheng Chen[1,2,*]

[1]Bioinformatics Laboratory, Institute of Biophysics and [2]Bioinformatics Research Group, Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, Peoples Republic of China

Received February 25, 2004; Revised May 28, 2004; Accepted August 20, 2004

ABSTRACT

The refinement and high-throughput of protein interaction detection methods offer us a protein–protein interaction network in yeast. The challenge coming along with the network is to find better ways to make it accessible for biological investigation. Visualization would be helpful for extraction of meaningful biological information from the network. However, traditional ways of visualizing the network are unsuitable because of the large number of proteins. Here, we provide a simple but information-rich approach for visualization which integrates topological and biological information. In our method, the topological information such as quasi-cliques or spoke-like modules of the network is extracted into a clustering tree, where biological information spanning from protein functional annotation to expression profile correlations can be annotated onto the representation of it. We have developed a software named PINC based on our approach. Compared with previous clustering methods, our clustering method ADJW performs well both in retaining a meaningful image of the protein interaction network as well as in enriching the image with biological information, therefore is more suitable in visualization of the network.

INTRODUCTION

It is now thought that the complexity of organisms rise not from the number of their macromolecules but rather from the relationships between them (1). Protein interactions are one of the major sources of this complexity. Since several high-throughput protein interaction detection approaches already have been developed (2–7), the information about protein interactions in yeast, which is one of the best-characterized model organisms, has grown considerably. A number of large-scale protein interaction data sets (2–5) have recently been published. These large-scale data sets provide a yeast protein–protein interaction network, in which proteins are depicted as vertices and interactions as edges.

The challenge coming along with the network is to find better ways to make it accessible for biological investigation. A visualization method would be helpful for extraction of useful biological information from the network. The present algorithms represent a vertex as a point, an edge as a straight line, and focus on how to draw these graphs more nicely and neatly either on two-dimensional (2D) picture (8) or in three-dimensional (3D) space (8,9) by adjusting positions of the points and lines. However, as there are thousands of proteins and tens of thousands of interactions in the yeast protein–protein interaction network, much information in the network remains hidden with this kind of visualization methods. For instance, in high linkage density areas such as cliques, the lines will overlap in a 2D picture, and is difficult to obtain information from dense parts even in a 3D representation.

Since the limitations of traditional visualization methods are unavoidable, we here present an alternative visualization method, which aims at combining the topological and biological information in a better way. The topological information is extracted from the network and displayed in a clustering tree. Based on the clustering tree, a graphical representation was created. The representation takes the form of a graphical adjacency matrix where proteins are listed according to the order of the clustering tree, and in which a pixel depicts an interaction between two proteins. Biological information is then, added into the graphical representation of the network. Different colors could be used to represent different biological information, spanning from protein functional annotation to expression profile correlations. We also provide a software (PINC, protein interaction network clustering) which can cluster and visualize protein–protein interation networks based on our method.

Topological clustering methods have proven to be a good solution for metabolic networks (10) and complicated networks in other areas (11,12). Recently, two research groups separately applied two clustering methods on the yeast protein interaction network (13,14). Several studies (15–19) have shown that there exists meaningful topological information

*To whom correspondence should be addressed. Tel: +86 10 6256 5533 5716; Fax: +86 10 6256 7724; Email: crs@sun5.ibp.ac.cn
Correspondence may also be addressed to Dongbo Bu. Tel: +86 10 6256 5533 5716; Fax: +86 10 6256 7724; Email: bdb@ncic.ac.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

in a protein–protein interaction network, commonly in the form of quasi-clique (20) or spoke-like patterns (21). Both patterns can be clustered in separate branches of topological clustering trees, thus revealing information about sub-topological modules of the network. However, different clustering methods reveal different parts of the network. Here, we provide two methods. One is a new topological clustering method which is called ADJW clustering, in which a modified adjacency matrix of the network was employed as the similarity matrix for the clustering. The other is a topological clustering method based on Hall (12). In this method the proteins were first projected into Euclidian space and then clustered according to their positions using a hierarchical clustering method. We applied both to the yeast protein–protein interaction network. Compared to two previously published methods (13,14), the ADJW clustering method is more suitable for visualizing several aspects of the network. We further analyzed the distribution of protein complexes and unclassified proteins in the ADJW clustering tree. The results show that the ADJW clustering method is an efficient tool for clusters of protein complexes. Further analyses show the Hall's method provides different and complementary results.

MATERIALS AND METHODS

Data source

This approach is applied to the yeast *Saccharomyces cerevisiae* protein–protein interaction network. The protein–protein interactions data detected by experiments, such as the yeast two-hybrid assay (5), HMS-PCI and TAP methods (2), were collected from the MIPS (http://mips.gsf.de/), PreBIND (http://bind.ca/index2.phtml?site=prebind), BIND (http://bind.ca/) and GRID (http://biodata.mshri.on.ca/grid/servlet/Index). In a preprocessing step, self interactions and redundant interactions were filtered out. For interactions detected by the HMS-PCI and TAP methods, the spoke model data (21,22) that assign interactions only between the bait and the associated proteins were used. This yielded an interaction data set containing 13 344 physical interactions among a total of 4537 yeast proteins (see Supplementary Material).

Methods

The topological information of the network was represented in a clustering tree produced by a Hierarchical Clustering algorithm. Biological information was annotated with color into the adjacency matrix base on the order of the clustering tree. Functional *P*-value and *P*-value of complex were employed as criteria to compare our topological clustering with the methods of Brun *et al*. (14) and Rives and Galitski (13).

Hierarchical Clustering Algorithm. A protein–protein interaction network is represented as a bi-directed graph $G(V,E)$, i.e. each protein is noted as a vertex and each interaction between proteins as an edge between vertices. Let A be the adjacency matrix, where $A = (a_{ij})$, $a_{ij} = 1$ when there is an edge between vertices i and j, and $a_{ij} = 0$ otherwise.

ADJW. The adjacency matrix A is employed as the similarity matrix. The average linkage hierarchical clustering is applied to this matrix. For two groups M and N, their average linkage is

$$D_{MN} = \frac{\sum_{m\in M}\sum_{n\in N} a_{mn}}{|M||N|}. \quad \mathbf{1}$$

Here D_{MN} represents the density of the edges between these two groups. The average linkage hierarchical clustering is a greedy algorithm based on D_{MN}. Two groups I and J which have the max value of D_{IJ} are clustered into one group in each step. By iterating these steps, a hierarchical clustering tree is generated.

In the beginning of the clustering, the method treats all of the edges as the same, and the proteins with edge are clustered together. As we know, the more common neighbors the two vertices in an edge have, the better the initial clustering of their edge. In order to decide which edge should be clustered first we made a modification based on the adjacency matrix A. Thus, we defined a similarity matrix as

$$S = A + w\cdot A^2, \quad \mathbf{2}$$

where w is a very small number, here, with an assigned value of 10^{-8}. The modification $w\cdot A^2$ ensures that the interacting protein pair which shares more neighbors should be clustered first.

The tree based on A is called the ADJ Tree while the tree based on S is called the ADJW Tree.

Hall Clustering (12). The clustering tree was obtained through a two-step process. First, the proteins in the network were projected into Euclidian space based on an optimization. Second, a hierarchical cluster method was applied to the proteins according to their positions in the Euclidian space.

Step 1: Projecting proteins into an r-dimensional Euclidean space. The vertices were projected into an r-dimensional Euclidian space according to the principle that two vertices should be as near as possible if there is an edge between them (12). For simplification, the problem to find one dimension to match the above requirement equals the problem of finding the vector $X = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$ by minimizing the following formula:

$$\min Q = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 a_{ij} = X^{\mathrm{T}}LX, \quad |X| = 1. \quad \mathbf{3}$$

Here, let $L = D - A$ be the *Laplacian* matrix, and D be the *diagonal* matrix $D_{ii} = \sum_k a_{ik}$, $D_{ij} = 0$ $(i \neq j)$.

It has been proved that Q will reach a minimum when X is the eigenvector with the minimal eigenvalue of the *Laplacian* matrix L (12). Hence, finding the first dimension can easily be solved by setting it as the eigenvector of the minimal eigenvalue.

Notice that all eigenvectors are mutually orthogonal because L is a symmetric matrix. The eigenvectors are employed to produce the r-dimensional space we aim at. That L is a positive semi-definite (12) matrix should facilitate the computation of eigenvectors. The rank of L will be $n - 1$ if G is a connected graph, which means only one eigenvalue of L equals 0. This trivial solution is not useful because it would mean that all proteins would be projected into one point. So we

can start from the second minimal eigenvector (Fiedler Vector) to get r eigenvectors for r dimensions. Here, $r = 350$ was chosen.

Step 2: Ward clustering method. The Ward hierarchical clustering method (23) was applied to cluster the projected vertices into a clustering tree. The Euclidean distances were used as a metric to measure the topological distance of vertices in the network. Since the two groups with the smallest sum form a new group for each step, the groups closer in the distance space will be chosen earlier during the clustering process. As a result, a clustering tree is produced step by step.

The tree based on this method is called the H Tree.

Visualization and annotation. Using our software PINC, we drew the adjacency matrix of the interaction network with row and column protein headings ordered according to the clustering tree. A filled-in row/column entry indicates an interaction between the two proteins heading the row and column. The color of the entry was used to symbolize different biological information such as information concerning individual protein (function annotation, complex annotation and degree of evolutionary conservation) or protein interactions (expression profile correlations, interaction confidence and regulatory relationship). Different patterns in the picture imply different topological structures in the network, i.e. a block means that the involved proteins form a clique while a line means a spoke module.

Comparison and validation

P-value of a branch. As a branch may involve different functional categories, *P*-values (24,25) were employed to assign each branch a main function, which is a criterion of coincidence of topological cluster and biological function.

Hypergeometric cumulative distribution was applied to model the probability of observing, by chance, at least k proteins in a branch size n belonging to a category containing C proteins from a total genome size of G proteins, such that the *P*-value is given by

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-C}{n-i}}{\binom{G}{n}}. \quad \mathbf{4}$$

The above test measures whether a branch is more enriched with proteins from a particular category than that would be expected by chance. If the *P*-value of a category is near 0, the proteins of the category in a branch will have low probability be chosen by chance. The functional category with the lowest *P*-value in a branch was assigned as its main function, and used to evaluate our clustering method.

The *P*-value of a complex

Each branch was assigned a *P*-value (see Equation 4) for a complex containing C proteins. The *P*-value of the complex $P(C_j)$ is the minimum *P*-value attained at any branch B_i in the hierarchical clustering tree T. That is,

$$P(C_j) = \min_{B_i \in T} \{P(B_i, C_j)\}. \quad \mathbf{5}$$

RESULTS

We applied our clustering method ADJW to a yeast *S.cerevisiae* protein–protein interaction data set containing 4537 yeast proteins and 13 344 physical interactions (see Materials and Methods). The ADJW clustering tree was displayed using TreeView (26) (http://rana.lbl.gov/EisenSoftware.htm) with functional annotation (22,27). The outline of the tree is shown in Figure 1. The graphical representation of the interaction network according to the ADJW Tree is outlined in Figure 2a (for details see Supplementary Figure 1).

This representation revealed many hidden modules in the network such as quasi-cliques and spoke-like modules. These modules, which are not easily revealed through conventional visualization, are believed to be biological meaningful and ready to be further analyzed by adding biological information. Among them, we presented two examples in Figure 2b, which was a representation of a branch of the clustering tree. A densely interconnected module (square block), which was clustered together, was a quasi-clique pattern (20) (Figure 2c). On the other hand, proteins that were clustered together in spoke-like patterns, were represented by slender blocks at right angles (21) (Figure 2d). Using our software PINC, the MIPS functional annotations are added into the representation (Figure 3). It indicated that the proteins in Figure 2c belong to *cellular fate/organization function category* and most proteins in Figure 2d belong to an unclassified category.

Using PINC, we analyzed quasi-cliques revealed in the ADJW Tree by adding biological annotation. Most of them are protein complexes. Details about the distribution of complexes in the ADJW Tree are shown in Supplementary Table 1 and Supplementary Figure 1. This representation also revealed a number of unclassified proteins clusters which are potentially new complexes or executing special biological functions (20). A selection of clusters from Tree comprising mostly unclassified proteins are shown in Supplementary Table 2.

In order to illustrate the advantages of the ADJW clustering method, we compared this method with the previously reported methods of Brun *et al*. (14) and Rives and Galitski (13), which have been used on the yeast protein–protein interaction network. We also applied these two methods on our data set along with the ADJ method and the Hall clustering method. The resulting trees were called the B Tree, the R Tree, the ADJ Tree and the H Tree, respectively. Together with the ADJW Tree, these trees were compared by several criteria.

A good test of how well the interaction information is retained in the clustering tree would be to compute the distribution of the shortest path in the tree between interacting proteins. We computed these distributions in the five trees, the result showing in Figure 4a. In comparison, the ADJW Tree is as good as the R Tree and better than the B Tree and the H Tree (Figure 4a). The topology of the network is well preserved in the ADJW clustering tree.

Besides the network topology reservation, biological information enrichment is another important criterion for evaluation of a clustering method. Applying MIPS functional annotation (27), the *P*-value was introduced to measure these enrichments in a clustering tree. We calculated the *P*-value of each branches in the five trees. In total, 264 branches covering 541 proteins in the ADJW Tree had *P*-values below 0.001 (25,28). Compared to other trees (see Table 1), the ADJW

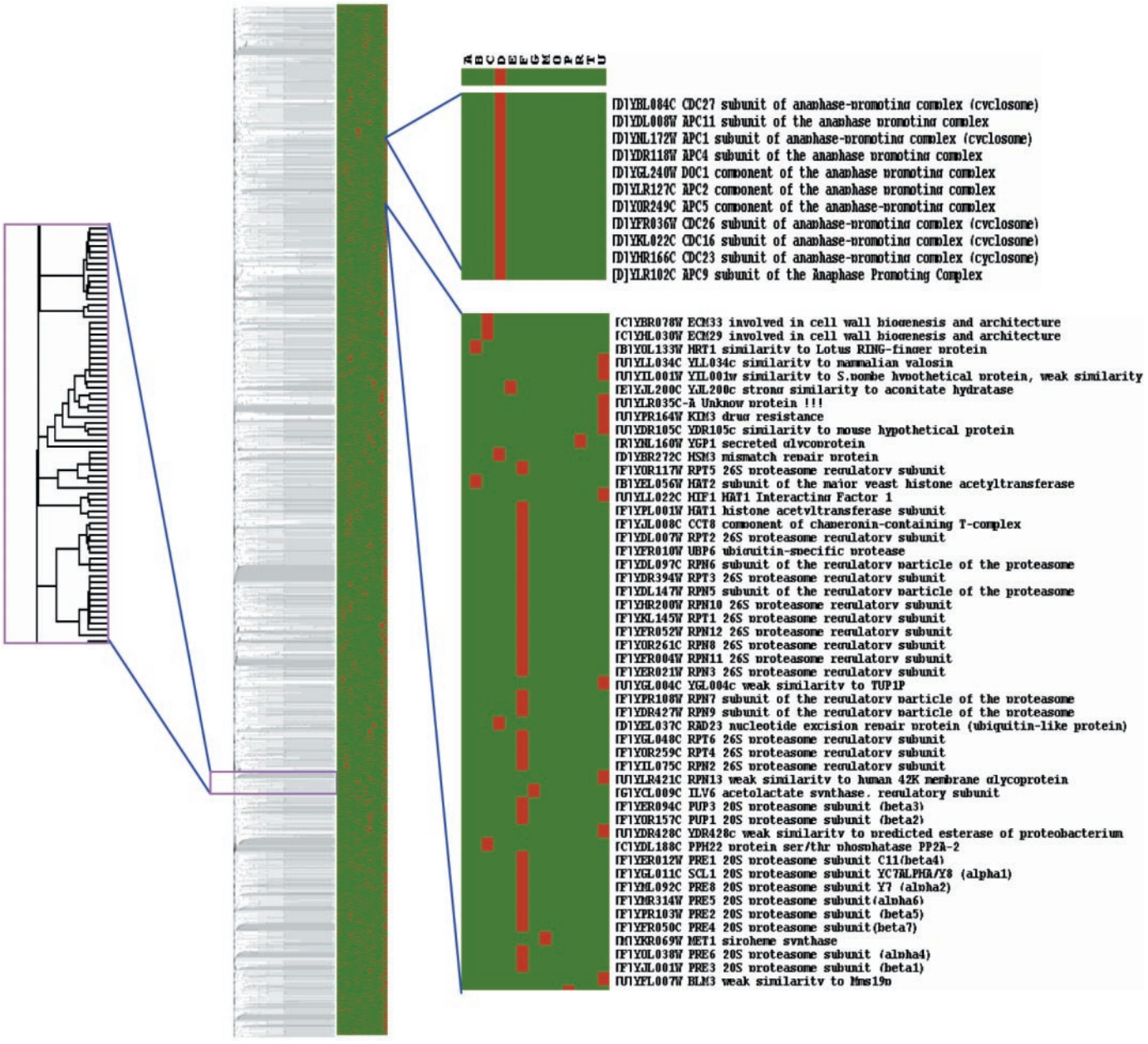

Figure 1. The ADJW clustering tree with protein functional annotation added (MIPS) which is drawn by TreeView (26). The length of the branches indicates the average linkage density of the local group. Two branches with protein annotations are highlighted on the right. The proteins were divided into 12 functional categories and one category for unclassified proteins (U). The color pattern shows the functional category of each protein. Proteins in a single branch tend to share common roles. The two branches highlighted mostly consist of proteins with Genome Maintenance (D) and Protein Fate (F) functions respectively. Other functional categories are as follows: E, energy production; G, amino acid metabolism; M, other metabolism; P, translation; T, transcription; B, transcriptional control; O, cellular organization; A, transport and sensing; R, stress and defense; C, cellular fate/organization.

Tree is almost as good as the B Tree and much better than the R Tree, H Tree and the ADJ Tree.

The *P*-value (see Methods) was used to measure the coincidence between the protein complexes and branches in a tree. We calculated the coincidence between 307 complexes from MIPS (27) (or *complex categories*) and the branches in five trees. The result is shown in Figure 4b. Among 307 complexes, 222 complexes had a *P*-value lower than 10^{-5} in the ADJW Tree. Compared to the other trees (see also Table 1), the ADJW Tree is better than the other four trees measured by this criterion. As the results show, biological information is much enriched through clustering by ADJW.

Another important criterion for visualization is the distance between interacting proteins in the visualized adjacency matrix ordered according to the clustering tree. An interaction will be near the diagonal of the matrix if the distance between the proteins is small. If there are many interacting proteins which have a small distance in the matrix, a major part of the

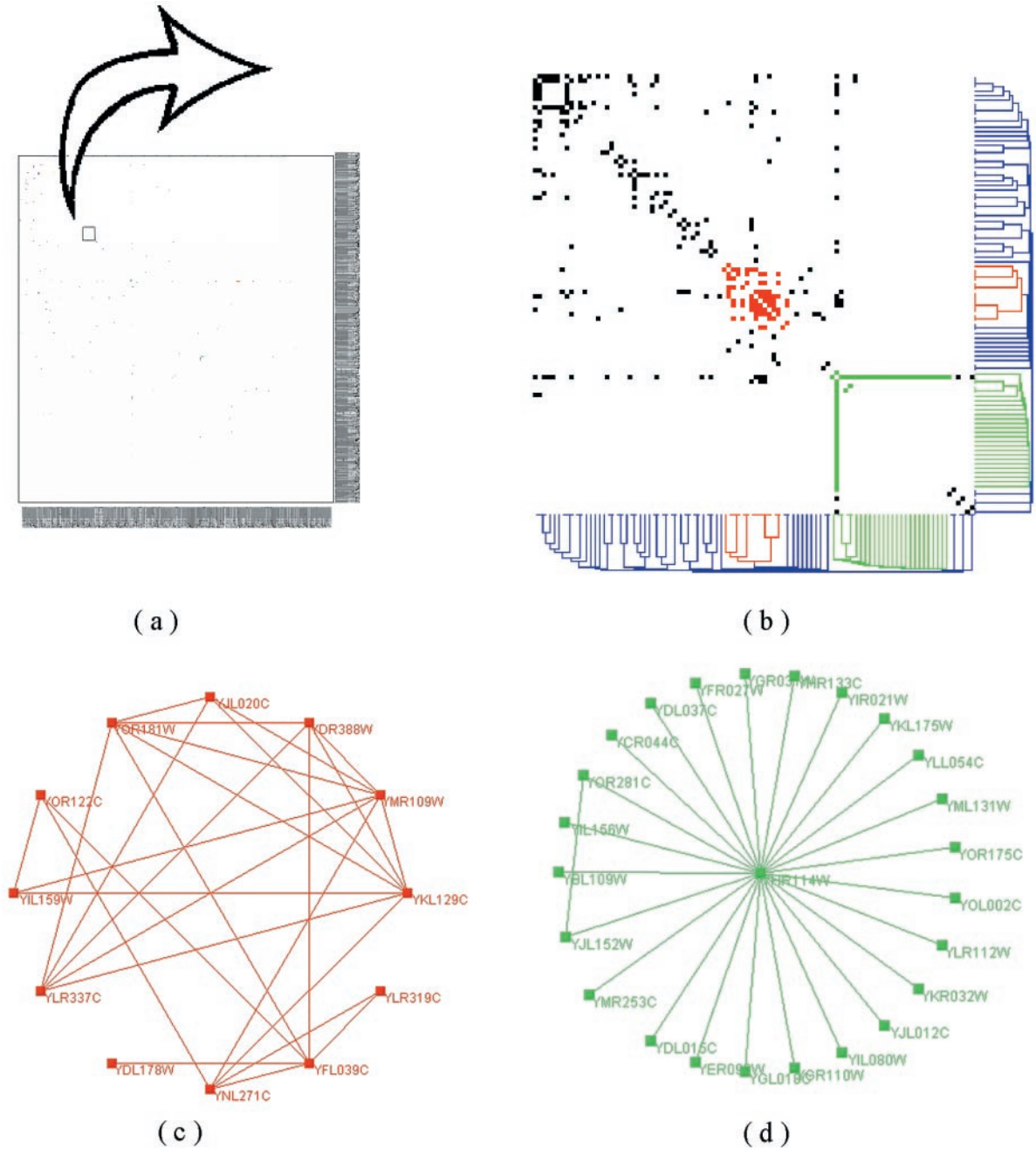

Figure 2. Details of a single branch. The outline of the graphic representation of the clustering tree (**a**), and a branch of it (**b**) are shown. The branch of the tree consists of a quasi-clique (see proteins in red) and a spoke-like fashion (see proteins in green). The traditional visualization of the quasi-clique and the spoke-like fashion is depicted in (**c**) and (**d**), respectively.

information will be found in this relatively narrowly diagonal area. We recorded separately the interactions in the selected area in the adjacent matrix of the five trees (Figure 4d), showing that the ADJW method performed well also on the basis of this criterion (Figure 4c).

In summary, the ADJW clustering method performed well both in retaining the interaction information and in enriching of the network with biological information. This clustering method thus has an advantage over other methods in visualizing the protein–protein interaction network.

However, the different clustering results revealed different information of the network. For example, the H Tree revealed an unusual, hidden module (Supplementary Figure 3), consisting of two quasi-cliques, in which the proteins were mostly unclassified according to the MIPS annotation. Classified proteins in this module were mostly related to RNA processing, and the unclassified proteins in one of the quasi-cliques have also been suggested to be involved in pre-rRNA processing (20,29). This module could not be easily seen in the other clustering trees.

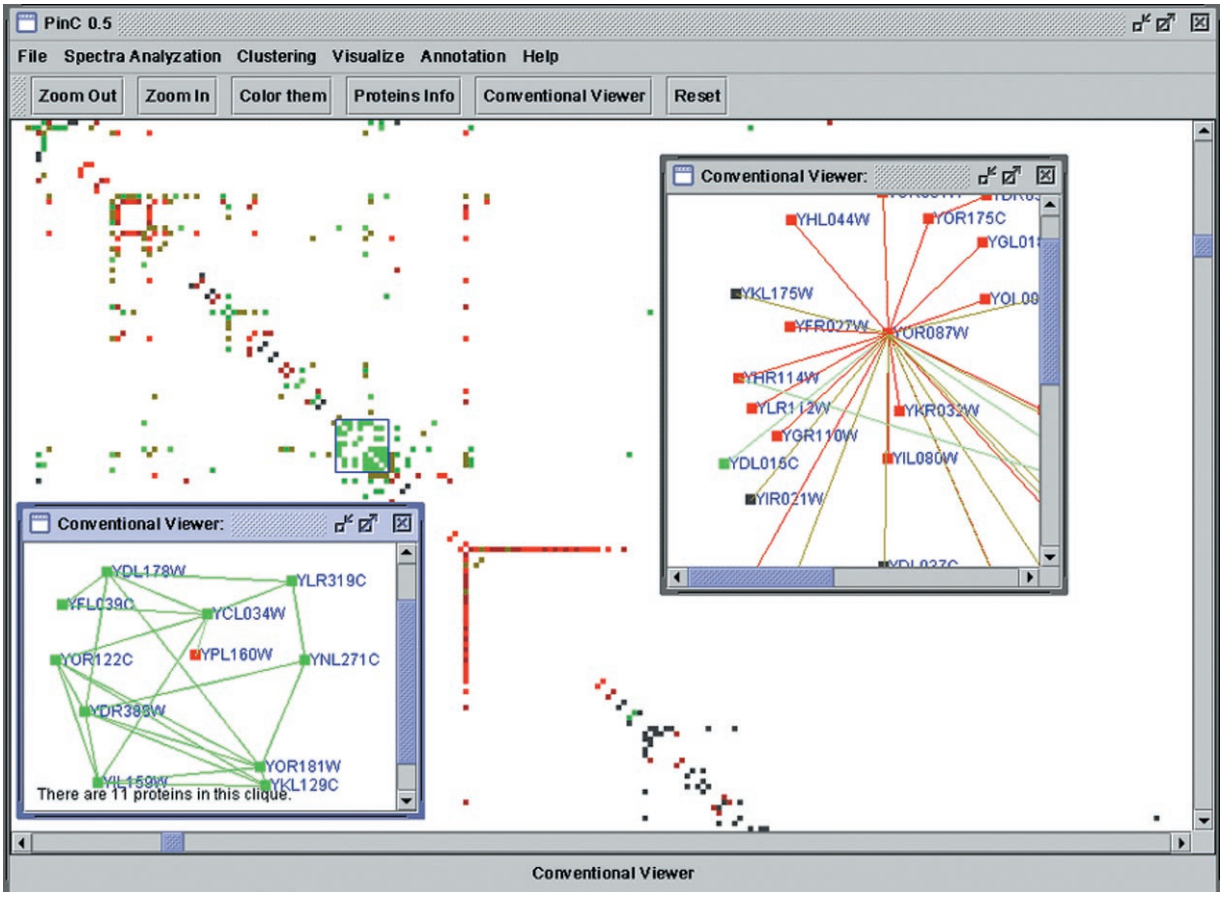

Figure 3. Visual impact of the yeast protein network using our software PINC. Unclassified proteins are in red, the cellular fate/organization proteins are in green and the others are in black. The color of the interaction between two proteins is computed by PINC using linear interpolation between the color codes representing the functional category of each of the linked proteins.

Table 1. The *P*-value of branches and of protein complexes

	R	B	H	ADJ	ADJW
$P < 0.001$					
Branches	173	274	155	163	264
Proteins	362	568	331	387	541
$P < 10^{-5}$					
Protein complexes	204	214	182	212	222

The first and second row show, for all five trees, the number of branches with a *P*-value less than 0.001, and the number of proteins covered by the respective branches. The third row shows the number of protein complexes with a *P*-value less than 10^{-5}.

DISCUSSION

The refinement and high-throughput of the protein interaction detection methods offer us a chance to study the protein–protein interaction network. However, because of the large number of proteins, the network is too complicated to be easily visualized. Here, we attempt to visualize this network in a simple but information-rich way. We drew the adjacency matrix of the interaction network according to the clustering tree, and used different colors to represent different biological information. Based on this approach, we have developed the software PINC. Using PINC, our approach was subsequently applied to visualize any two or more categories of proteins in yeast protein–protein interaction network, e.g. 'prokaryotic' 'eukaryotic' proteins (i.e. proteins with or without a prokaryote ortholog, see Supplementary Figure 1). Other biological information such as the expression profile (see Supplementary Figure 2), interaction confidence and regulatory relationships can also be integrated into this approach. Given the abundant information linked to proteins and protein relationships, a versatile visualization approach such as ours should be highly useful. Compared with conventional visualization methods, our method has an advantage in that it reveals more hidden modules in a large network.

However, since some information about the network is lost through our visualization method, we also integrated a conventional visualization method into our software PINC. The PINC software has a friendly graphical interface and can be downloaded from http://www.bioinfo.org.cn/clustering.

The ADJW clustering is a newly developed method, whereas application of Hall's method for clustering of genomic networks represents an older strategy applied to a new field. The former method has the advantage of being simple and easy in implementation; however, when further developed, Hall's clustering may still have a potential in visualization of biological networks (see Figure 4c). These methods and all the other clustering methods which are mentioned in this

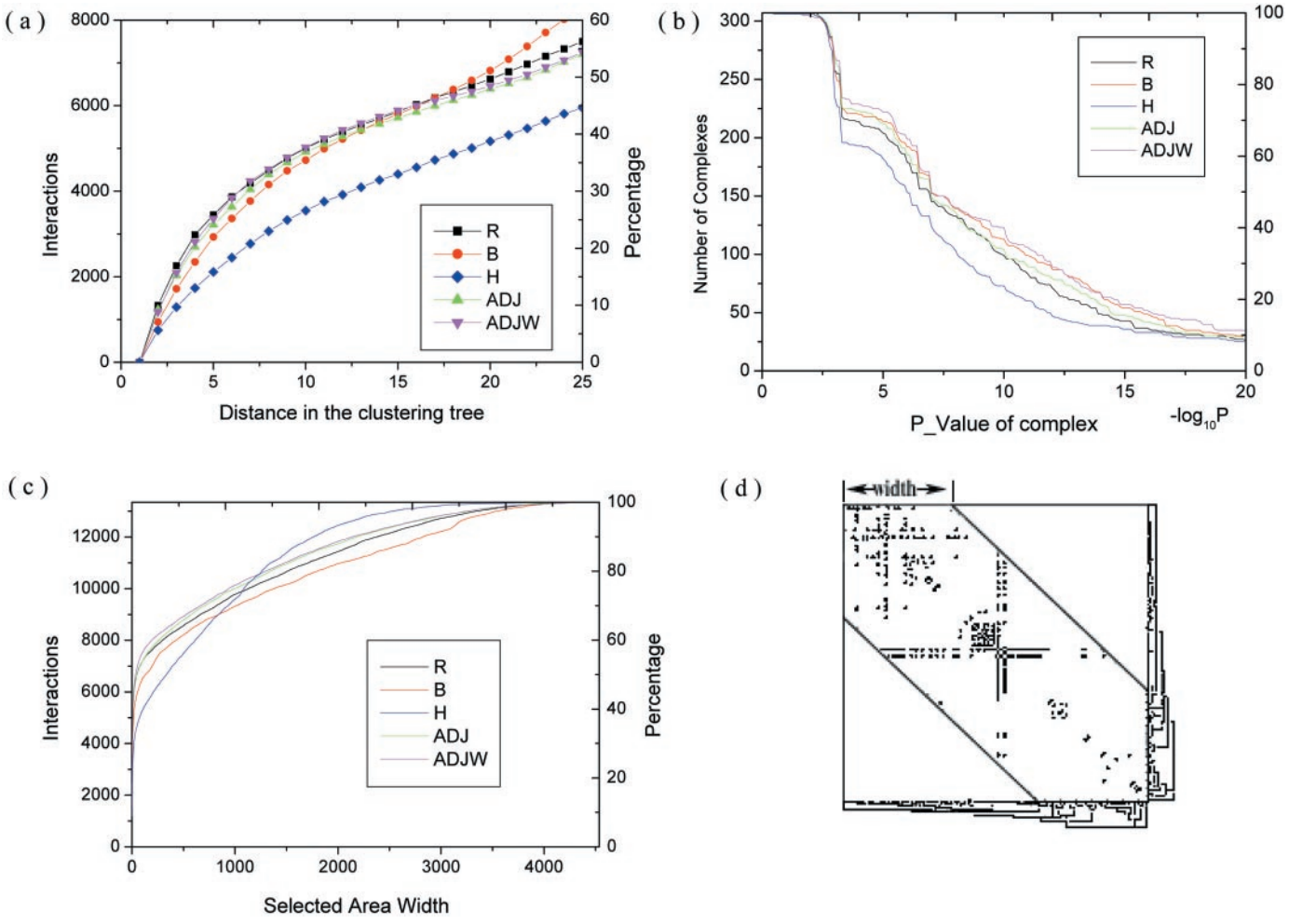

Figure 4. Comparison and validation: (**a**) Distribution of interacting proteins according to the shortest path between them in the tree. A point in the line shows the number of interactions (*y*-axis) that has a path less than a certain distance (*x*-axis) between the two interacting proteins (**b**) Number of complex with a *P*-value lower than a certain level (*x*-axis). The *P*-values were calculated for a list of 307 protein complexes from MIPS (see Methods). (**c**) Distribution of protein interactions according to the selected area in the adjacency matrix; (**d**) definition of selected area width used in (c).

paper were also integrated in PINC. We hope that biologists will find useful biological information based on our software when studying the distribution of the proteins they focus on.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors acknowledge helpful advice from Dr Yan Fu. This work was supported by the Chinese Academy of Sciences Grant No. KSCX2-2-27, National Sciences Foundation of China Grant No. 39890070, 60496320, the National High Technology Development Program of China under Grant No. 2002AA231031, National Key Basic Research & Development Program 973 under Grant No.2002CB713805, 2003CB715900, and Beijing Science and Technology Commission Grant No. H010210010113.

REFERENCES

1. Claverie,J.M. (2001) Gene number. What if there are only 30 000 human genes? *Science*, **291**, 1255–1257.
2. Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
3. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
4. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
5. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
6. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
7. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.

8. Batagelj,V. and Mrvar,A. (2001) Pajek-analysis and visualization of large networks. *LNCS*, **2265**, 477–478.
9. Han,K. and Ju,B.H. (2003) A fast layout algorithm for protein interaction networks. *Bioinformatics*, **19**, 1882–1888.
10. Ravasz,E., Somera,A.L., Mongru,D.A., Oltvai,Z.N. and Barabasi,A.L. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.
11. Girvan,M. and Newman,M.E. (2002) Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA*, **99**, 7821–7826.
12. Hall,K.M. (1970) An *r*-dimensional quadratic placement algorithm. *Management Science*, **17**, 219–229.
13. Rives,A.W. and Galitski,T. (2003) Modular organization of cellular networks. *Proc. Natl Acad. Sci. USA*, **100**, 1128–1133.
14. Brun,C., Chevenet,F., Martin,D., Wojcik,J., Guenoche,A. and Jacq,B. (2003) Functional classification of proteins for the prediction of cellular function from a protein–protein interaction network. *Genome Biol.*, **5**, R6.
15. Maslov,S. and Sneppen,K. (2002) Specificity and stability in topology of protein networks. *Science*, **296**, 910–913.
16. Saito,R., Suzuki,H. and Hayashizaki,Y. (2002) Interaction generality, a measurement to assess the reliability of a protein–protein interaction. *Nucleic Acids Res.*, **30**, 1163–1168.
17. Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.
18. Jeong,H., Mason,S.P., Barabasi,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
19. Goldberg,D.S. and Roth,F.P. (2003) Assessing experimentally derived interactions in a small world. *Proc. Natl Acad. Sci. USA*, **100**, 4372–4376.
20. Bu,D., Zhao,Y., Cai,L., Xue,H., Zhu,X., Lu,H., Zhang,J., Sun,S., Ling,L., Zhang,N. *et al.* (2003) Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic Acids Res.*, **31**, 2443–2450.
21. Bader,G.D. and Hogue,C.W. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.*, **20**, 991–997.
22. Von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.
23. Ward,J.H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236–244.
24. Wu,L.F., Hughes,T.R., Davierwala,A.P., Robinson,M.D., Stoughton,R. and Altschuler,S.J. (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nature Genet.*, **31**, 255–265.
25. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.
26. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
27. Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N., Stumpflen,V. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32** (Database issue), D41–44.
28. Larsen,B. and Aone,C. (1999) *Fast and Effective Text Mining Using Linear-time Document Clustering*.
29. Dragon,F., Gallagher,J.E., Compagnone-Post,P.A., Mitchell,B.M., Porwancher,K.A., Wehner,K.A., Wormsley,S., Settlage,R.E., Shabanowitz,J., Osheim,Y. *et al.* (2002) A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. *Nature*, **417**, 967–970.