

Fast and Accurate Method for Identifying High-Quality Protein-Interaction Modules by Clique Merging and Its Application to Yeast

Chi Zhang, Song Liu, and Yaoqi Zhou*

Howard Hughes Medical Institute Center for Single Molecule Biophysics, Department of Physiology & Biophysics, State University of New York at Buffalo, 124 Sherman Hall, Buffalo, New York 14214

Received October 31, 2005

Molecular networks in cells are organized into functional modules, where genes in the same module interact densely with each other and participate in the same biological process. Thus, identification of modules from molecular networks is an important step toward a better understanding of how cells function through the molecular networks. Here, we propose a simple, automatic method, called MC², to identify functional modules by enumerating and merging cliques in the protein-interaction data from large-scale experiments. Application of MC² to the *S. cerevisiae* protein-interaction data produces 84 modules, whose sizes range from 4 to 69 genes. The majority of the discovered modules are significantly enriched with a highly specific process term (at least 4 levels below root) and a specific cellular component in Gene Ontology (GO) tree. The average fraction of genes with the most enriched GO term for all modules is 82% for specific biological processes and 78% for specific cellular components. In addition, the predicted modules are enriched with coexpressed proteins. These modules are found to be useful for annotating unknown genes and uncovering novel functions of known genes. MC² is efficient, and takes only about 5 min to identify modules from the current yeast gene interaction network with a typical PC (Intel Xeon 2.5 GHz CPU and 512 MB memory). The CPU time of MC² is affordable (12 h) even when the number of interactions is increased by a factor of 10. MC² and its results are publicly available on <http://theory.med.buffalo.edu/MC2>.

Keywords: module • network • clique • protein–protein interaction

Introduction

Living biological cells function through a network of interacting molecules. It is increasingly evident that interacting molecules such as proteins, DNAs, and others are organized (or clustered) in the groups called functional modules to perform specific tasks.^{1–3} Identification of functional modules from a known interaction network allows us to “divide and conquer” the problem of understanding the function of the interaction network.

An important parameter to measure the “modularity” of an interaction network is connection density Q (fraction of the actual number of connections in the maximal number of possible connections).^{4–6} A highly connected sub-network with a high connection density close to 1 is a candidate for a functional module. Several computational methods have been developed to identify modules by locating highly connected sub-networks (i.e., sub-networks with high Q values) from protein–protein interaction and/or coexpression network.^{3,7–11}

Cliques, fully connected subgraphs with $Q = 1$, are likely either parts of modules or modules themselves. This is supported by the recent work of Spirin and Mirny who showed

the statistically significant enrichment of the number of cliques in a protein interaction network over random networks.³ In our work, we use cliques in protein interaction network as initial “seeds” for modules and merge the partially connected cliques for final module identification. This algorithm (called MC² – Merging Cliques for Module identification) is applied to the well documented interaction data of yeast *S. cerevisiae*. We find that this algorithm based on the interaction information alone produces a number of modules with highly co-functional, co-localized, and coexpressed proteins. The predicted modules are deposited in an on-line MC² server to facilitate further exploration of genes inside or close to the identified modules.

Theory, Method, and Material

Interaction Data. The comprehensive protein-interaction data of yeast are obtained by combining sources from both physical and genetic interactions. The physical interaction data are downloaded from the Database of Interacting Proteins (DIP), which includes 15 147 nonredundant, heterogeneous interactions for 4738 genes.¹² The data of genetic interactions were assembled by Kelley and Ideker, which had 5212 linkages involving 1167 genes.¹³ The combination of these two datasets results 19 614 nonredundant interactions for 4950 genes.

* To whom correspondence should be addressed. Tel: (716) 829-2985. Fax: (716) 829-2344. E-mail: yqzhou@buffalo.edu.

Network Construction. A protein interaction graph⁶ is constructed as an undirected graph that is made of all proteins in the interaction data set collected above. In this graph, each vertex represents a distinct protein, and two vertices are linked by an edge if they have an interaction measured by supporting experiments (either physical or genetic). The density of a graph (or subgraph), Q , is defined as $Q = 2E/V(V - 1)$, where E is the total number of edges and V is the number of vertices. A complete sub-graph, for which each vertex has connections to all of the rest vertices, is called a clique. The density of a clique is one ($Q = 1$).

MC² Algorithm. Algorithm 457 from Communications of the ACM developed by Bron and Kerbosch¹⁴ is used to enumerate all nonredundant cliques in a protein interaction network. The obtained cliques are ranked according to their sizes. While cliques are fully connected themselves, it is possible for two cliques to share some nodes and/or to have some connections between them. Here we describe an algorithm that merges those partially connected cliques.

We start from the largest clique c_m (parent) and locate all smaller cliques (size ≥ 4) that share nodes with it. We use the following criteria to determine if any of the smaller cliques is to be merged with the largest clique. The first natural criterion for a merging operation to occur is that the interaction density after merging (Q_{merge}) needs to be greater than a cutoff value Q_{cutoff} . This is based on the assumption that biological modules are highly connected. However, the application of this criterion alone favors the merge between small cliques because only a small number of edges between them is needed to reach high Q_{merge} . Therefore, we add a second criterion that requires the number of the shared nodes between two cliques to be greater than a certain threshold. However, an absolute cutoff value for the number of shared nodes favors the merge between large cliques. To balance the need for merging cliques small and large, we use a relative cutoff value. A smaller clique c_j (child, with at least 4 nodes) will be merged to the parent clique c_m if the ratio of the number of their shared nodes (n^j_{shared}) to the maximal number of common nodes between c_m and all other available cliques [$\max(n^i_{\text{shared}}, \forall i)$] is larger than a certain cutoff. That is, $n^j_{\text{shared}}/\max(n^i_{\text{shared}}, \forall i) > f_{\text{cutoff}}$.

All smaller cliques satisfying the above two criteria are merged with the largest clique to form a subgraph called the draft module. These merged smaller cliques are removed from the list of cliques. The remaining cliques are then compared with the established draft module. Any shared nodes between the remaining cliques and the draft module are removed from the cliques. This produces a clean set of new cliques (or clique fragments) that do not share any nodes with the established draft module. Next, we search for possible merges among new cliques by starting from the largest clique in the new list and following the same merging procedure and criteria described above. The merge operation continues until no more merging partner can be found in the remaining cliques.

The above procedure yields a set of draft modules (merged cliques) and some remaining un-merged orphan cliques (cliques with size of 3 and clique fragments). To reduce the number of orphan cliques, we allow an orphan clique to merge with a draft module if the number of interactions within the orphan clique or between the orphan clique and the draft module, k_{in} , is larger than k_{out} , the number of other connections between the orphan clique and its environment (i.e., the number of connections that are neither within the orphan clique nor between the orphan clique and the module).¹⁵ In addition, the density after

merging should be greater than a to-be-determined cutoff value, $Q_{\text{cutoff}}^{\text{orphan}}$. The completion of merging orphan cliques with draft modules provides a list of final predicted modules. More detailed description of the algorithm is provided in the supplement material.

Module Assessment. Whether the modules predicted by clique merging are truly biological modules has to be assessed independently. An identified module is biologically meaningful if it consists of proteins involved in the same cellular process. We use GO::TermFinder developed by Boyle et al.¹⁶ to assess whether the modules resulted from merging cliques are significantly enriched with certain “biological process” GO terms.¹⁷ In GO::TermFinder, the statistical significance is calculated using the P -value based on hyper-geometric distribution under multiple hypothesis corrections.¹⁶ The homogeneity for a given module is defined as the fraction of nodes (or proteins) with the most enriched GO term. The highest possible homogeneity is 1 (i.e., 100% of proteins in the module have the same enriched GO term). Since each module may be annotated by many different GO terms, we present the specific GO term with the most significant P -value. Similarly, we evaluate locational homogeneity based on the GO annotation of cellular components. This is to assess if a predicted module is enriched with proteins located in the same cellular component. The gene ontology file version is 1.419 on Sep. 30, 2005, and annotation file version is 1.1190 on Sep. 30, 2005.

Cutoff Values. There are three adjustable parameters (f_{cutoff} , Q_{cutoff} , and $Q_{\text{cutoff}}^{\text{orphan}}$) in MC² algorithm. From their definitions, all three parameters are between 0 and 1. Moreover, $Q_{\text{cutoff}}^{\text{orphan}}$ shall be smaller than Q_{cutoff} in order to further incorporate some orphan cliques into modules. To reduce the number of adjustable parameters for MC² in this study, we set f_{cutoff} equal to Q_{cutoff} , and hence use only one parameter r_{cutoff} to represent both of them. That is, $r_{\text{cutoff}} = f_{\text{cutoff}} = Q_{\text{cutoff}}$ and $r_{\text{cutoff}} \in (0, 1)$. Thus, there are only two parameters (r_{cutoff} and $Q_{\text{cutoff}}^{\text{orphan}}$) for our merging algorithm. The cutoff value of $Q_{\text{cutoff}}^{\text{orphan}}$ controls the number of orphan cliques merged into draft modules. Because major merges occur at the formation of draft modules, the variation of this parameter is not expected to make a significant change in the overall results. Here, we set $Q_{\text{cutoff}}^{\text{orphan}} = 0.6r_{\text{cutoff}}$. As a result, there is only one key adjustable parameter in our algorithm, r_{cutoff} , that controls the formation of draft modules. A large r_{cutoff} value tends to prohibit clique merging by raising the required minimum for the overlap between two cliques as well as for the interaction density of the merged cliques. Because the initial number of cliques is fixed, increasing r_{cutoff} will increase the number of modules (Supporting Information Figure 1), decrease the number of proteins per module (Supporting Information Figure 2), and likely enhance functional homogeneity per module. Figure 1 shows functional homogeneity per module as a function of r_{cutoff} . Indeed, homogeneity increases as r_{cutoff} increases. However, the increase is not monotonic; there is a local maximum of the average homogeneity over all modules at $r_{\text{cutoff}} \approx 0.5$. The global maximum will be at $r_{\text{cutoff}} = 1$ when there are no merges between cliques. However, this is not the solution that we are seeking. This solution would produce a significantly lower number of genes involved in modules (Supporting Information Supplement Figure 3). Thus, throughout this paper, we set $r_{\text{cutoff}} = 0.5$. Note that the average homogeneity is close to a constant value between $r_{\text{cutoff}} = 0.4$ and $r_{\text{cutoff}} = 0.7$. This suggests that the overall result is not very sensitive to the cutoff value for this range of r_{cutoff} . To confirm that the production of final modules

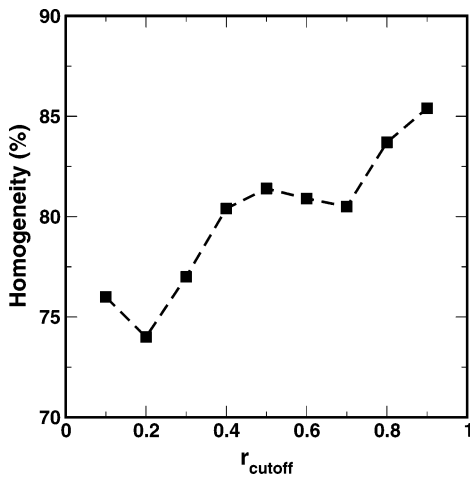


Figure 1. Average percentage of the number of proteins that share an enriched GO process term in one discovered module as a function of the cutoff value r_{cutoff} with $Q_{\text{cutoff}}^{\text{orphan}}/r_{\text{cutoff}} = 0.6$.

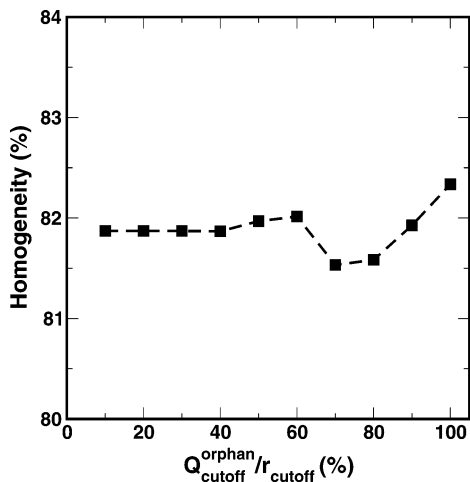


Figure 2. Average percentage of the number of proteins that share an enriched GO process term in one discovered module as a function of $Q_{\text{cutoff}}^{\text{orphan}}/r_{\text{cutoff}}$ with $r_{\text{cutoff}} = 0.5$.

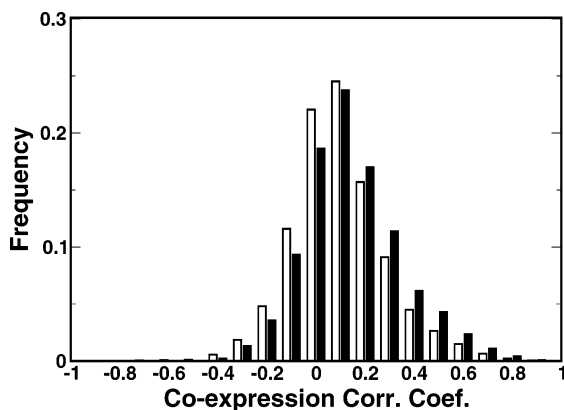


Figure 3. Distribution of the pairwise correlation coefficients of gene expression profiles for interaction proteins in the whole interaction network (open bar) and in the modules obtained (filled bar).

is less sensitive to the value of $Q_{\text{cutoff}}^{\text{orphan}}$, we show in Figure 2 that the average percentage of the number of proteins share an enriched GO term in a discovered module as a function of $Q_{\text{cutoff}}^{\text{orphan}}/r_{\text{cutoff}}$. As expected, the average homogeneity makes

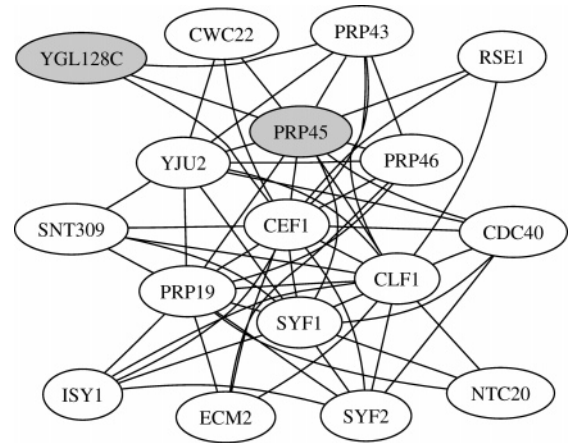


Figure 4. Interaction network of Module #15 for nuclear mRNA splicing (GO level 8). All genes except *YGL128C* and *PRP45* (shaded circles) are known to involve in this process based on the current GO annotation. *PRP45* is linked to this process in a recent work.²¹ *YGL128C* is an unknown gene.

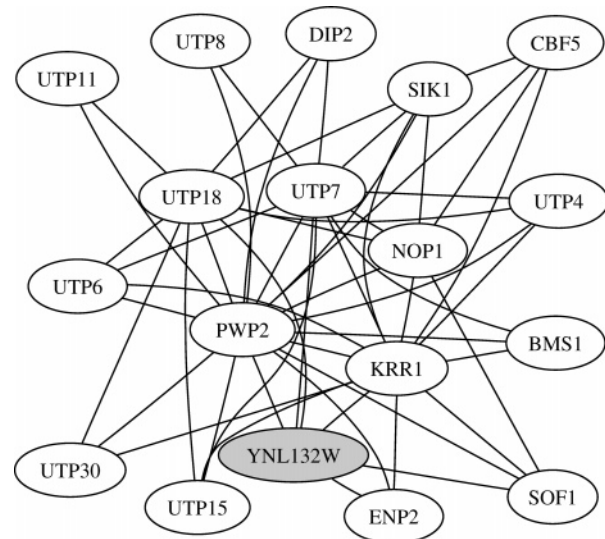


Figure 5. Interaction network of Module #13 for rRNA processing (GO level 7). All except the unknown gene *YNL132W* (shaded circle) are known to involve in this process.

little change (0.5%) within the whole range of $Q_{\text{cutoff}}^{\text{orphan}}/r_{\text{cutoff}}$ (from 0 to 1). The largest average homogeneity occurs at $Q_{\text{cutoff}}^{\text{orphan}}/r_{\text{cutoff}} = 1$ where no orphan clique is included in the formation of modules. As $Q_{\text{cutoff}}^{\text{orphan}}/r_{\text{cutoff}}$ decreases, the average homogeneity decreases until $Q_{\text{cutoff}}^{\text{orphan}}/r_{\text{cutoff}} = 0.6$. Figure 2 is somewhat similar to Figure 1 except that the former has a much smaller change in homogeneity.

To further check the dependence of the MC^2 algorithm on r_{cutoff} , we also study how the average locational homogeneity and the significance of enrichment of coexpressed proteins changes with different r_{cutoff} input. Results shown in Supplement Figure 4 and Figure 5 indicate that r_{cutoff} with a value of 0.5 offers an optimal performance for both cases.

Results

Application of the MC^2 algorithm to the yeast interaction data set yields 84 modules with sizes ranging from 4 to 69 proteins and an average size of 10.6 proteins. Of the 84

Table 1. List of Modules Obtained (size ≥ 8) by MC² in the Yeast Protein–Protein Interaction Network and Their Most Enriched GO Terms for Biological Processes

| no. | size ^a | C ^b | H(%) ^c | GO ID | level ^d | description ^e | P-value ^f |
|-----|-------------------|----------------|-------------------|---------|--------------------|--|-----------------------|
| 1 | 69 | 287 | 38 | 0007017 | 7 | microtubule-based process | 6.1×10^{-29} |
| 2 | 34 | 68 | 91 | 0007028 | 5 | cytoplasm organization & biogenesis | 2.4×10^{-39} |
| 3 | 31 | 50 | 77 | 0006508 | 7 | proteolysis | 7.9×10^{-32} |
| 4 | 30 | 24 | 83 | 0016071 | 6 | mRNA metabolism | 6.3×10^{-33} |
| 5 | 22 | 21 | 91 | 0006366 | 8 | transcription from Pol II promoter | 3.2×10^{-24} |
| 6 | 21 | 17 | 71 | 0006606 | 7 | protein-nucleus import | 3.1×10^{-27} |
| 7 | 21 | 18 | 71 | 0016043 | 4 | cell organization & biogenesis | 6.3×10^{-5} |
| 8 | 20 | 34 | 85 | 0000398 | 8 | nuclear mRNA splicing, via spliceosome | 2.1×10^{-27} |
| 9 | 20 | 43 | 95 | 0006259 | 5 | DNA metabolism | 5.6×10^{-19} |
| 10 | 19 | 24 | 68 | 0030036 | 8 | actin cytoskeleton organization & biogenesis | 6.1×10^{-18} |
| 11 | 19 | 13 | 79 | 0006888 | 6 | ER to Golgi transport | 5.2×10^{-25} |
| 12 | 18 | 18 | 78 | 0031123 | 7 | RNA 3'-end processing | 4.2×10^{-32} |
| 13 | 18 | 17 | 94 | 0006364 | 7 | rRNA processing | 8.5×10^{-25} |
| 14 | 17 | 13 | 94 | 0006412 | 6 | protein biosynthesis | 1.0×10^{-12} |
| 15 | 17 | 16 | 88 | 0000398 | 8 | nuclear mRNA splicing, via spliceosome | 1.5×10^{-24} |
| 16 | 17 | 15 | 65 | 0000114 | 8 | G1-specific transcription in mitotic cell cycle | 1.6×10^{-26} |
| 17 | 15 | 8 | 47 | 0000001 | 7 | mitochondrion inheritance | 5.0×10^{-12} |
| 18 | 15 | 11 | 67 | 0006413 | 8 | translational initiation | 7.6×10^{-17} |
| 19 | 15 | 10 | 80 | 0006402 | 7 | mRNA catabolism | 1.6×10^{-21} |
| 20 | 15 | 9 | 100 | 0006383 | 8 | transcription from Pol III promoter | 2.2×10^{-34} |
| 21 | 13 | 8 | 54 | 0006357 | 9 | regulation of transcription from Pol II promoter | 1.0×10^{-7} |
| 22 | 12 | 6 | 100 | 0007035 | 7 | vacuolar acidification | 1.0×10^{-29} |
| 23 | 12 | 5 | 100 | 0007031 | 6 | peroxisome organization & biogenesis | 5.0×10^{-27} |
| 24 | 12 | 7 | 92 | 0006753 | 8 | nucleoside phosphate metabolism | 5.9×10^{-27} |
| 25 | 12 | 8 | 58 | 0000289 | 9 | poly(A) tail shortening | 6.3×10^{-19} |
| 26 | 11 | 3 | 100 | 0008054 | 6 | cyclin catabolism | 3.5×10^{-32} |
| 27 | 11 | 5 | 64 | 0006368 | 9 | RNA elongation from Pol II promoter | 4.9×10^{-14} |
| 28 | 10 | 7 | 80 | 0016192 | 5 | vesicle-mediated transport | 2.7×10^{-8} |
| 29 | 10 | 5 | 90 | 0006487 | 8 | protein amino acid N-linked glycosylation | 5.7×10^{-18} |
| 30 | 10 | 6 | 60 | 0007010 | 6 | cytoskeleton organization and biogenesis | 1.7×10^{-4} |
| 31 | 10 | 8 | 80 | 0006413 | 8 | translational initiation | 1.4×10^{-14} |
| 32 | 9 | 4 | 100 | 0006888 | 6 | ER to Golgi transport | 4.3×10^{-17} |
| 33 | 9 | 6 | 89 | 0006511 | 10 | ubiquitin-dependent protein catabolism | 6.4×10^{-12} |
| 34 | 9 | 5 | 44 | 0006357 | 9 | regulation of transcription from Pol II promoter | 6.0×10^{-3} |
| 35 | 9 | 4 | 67 | 0006333 | 8 | chromatin assembly or disassembly | 8.8×10^{-8} |
| 36 | 9 | 3 | 44 | 0000077 | 7 | DNA damage checkpoint | 8.0×10^{-8} |
| 37 | 8 | 5 | 63 | 0045835 | 9 | negative regulation of meiosis | 6.1×10^{-14} |
| 38 | 8 | 3 | 75 | 0007046 | 7 | ribosome biogenesis | 2.7×10^{-6} |
| 39 | 8 | 2 | 100 | 0016571 | 8 | histone methylation | 2.0×10^{-20} |
| 40 | 8 | 3 | 88 | 0006350 | 6 | transcription | 9.4×10^{-6} |
| 41 | 8 | 1 | 88 | 0006267 | 8 | pre-replicative complex formation & maintenance | 1.1×10^{-17} |
| 42 | 8 | 2 | 100 | 0000910 | 5 | cytokinesis | 1.3×10^{-13} |
| 43 | 8 | 3 | 75 | 0000082 | 7 | G1/S transition of mitotic cell cycle | 7.1×10^{-10} |

^a The size of each module. ^b The number of merging operations applied to the module (including merging orphan cliques). ^c The fraction of genes in the module has the same most-significantly enriched GO function. ^d The level of the corresponding GO term below the “root” in the shortest path. ^e The description of the corresponding GO term. ^f The *P*-value of the most-significantly enriched GO process term.

modules, eighty (95%) are significantly (P -value $< 10^{-3}$) enriched with a highly specific GO process term: a term that is at least 4 levels below the “root” in the shortest path. [A protein annotated to a GO node also automatically belongs to all its ancestral nodes.¹⁷ As in ref 18, we defined a GO term as sufficiently specific if it is 4 levels below the “root” of GO tree, that is, the length of the shortest path from this term to the root of this tree is at least 4.] Additionally, three of the remaining four modules are also enriched with a specific GO term but with a less significant *P*-value ($< 10^{-2}$). The only module with P -value $> 10^{-2}$ (statistically insignificant) is a small module whose size is 4. The average level of all enriched GO terms for all 84 modules is 7.1. The average homogeneity rate is 82%. Table 1 lists those modules with sizes of 8 proteins or more. The table shows that the majority of large modules (87% for sizes of 10 or greater, 79% for sizes of 8 or greater) have *P*-values that are less than 10^{-10} . This highlights the high quality of the discovered modules.

Table 1 also shows the number of merging operations applied to each module. The majority of modules are from the merge of many cliques. In general, the larger the module, the

more reliable it is (based on the *P*-value), and the greater the number of merging operations is. This indicates that the proposed algorithm successfully merges the cliques with the same biological function without using functional information during merging.

It should be noted that only the GO term with the most significant *P*-value is shown in Table 1. Other GO terms may have a less significant *P*-value but a much higher homogeneity. For example, the *P*-value and homogeneity of module #17 for GO term GO:0000001 are 5.0×10^{-12} and 47%, respectively. The corresponding values for GO:0051179 term are 1.2×10^{-4} and 73%, respectively.

One can also assess if the predicted modules are enriched with proteins in the same cellular component annotated in GO. Of the 84 predicted modules, 71 (85%) are significantly (P -value $< 10^{-3}$) enriched with a specific cellular component. The average locational homogeneity is 78%. More detailed results are shown in Table 2.

The discovered modules, if co-functional, should more likely be coexpressed. To test this, we obtained the comprehensive yeast expression profile *Rosetta compendium*,¹⁹ which includes

Table 2. List of Modules Obtained (size ≥ 8) by MC² in the Yeast Protein–Protein Interaction Network and Their Most Enriched GO Terms for Cellular Components

| no. | size ^a | H (%) ^b | GO ID | level ^c | description ^d | P-value ^e |
|-----|-------------------|--------------------|---------|--------------------|--|-----------------------|
| 1 | 69 | 30 | 0015630 | 4 | microtubule cytoskeleton | 4.8×10^{-21} |
| 2 | 34 | 85 | 0005730 | 4 | nucleolus | 2.6×10^{-33} |
| 3 | 31 | 81 | 0000502 | 3 | proteasome complex (sensu Eukaryota) | 2.9×10^{-50} |
| 4 | 30 | 60 | 0046540 | 5 | U4/U6 x U5 tri-snRNP complex | 4.1×10^{-35} |
| 5 | 22 | 82 | 0000119 | 6 | mediator complex | 4.1×10^{-46} |
| 6 | 21 | 67 | 0005643 | 5 | nuclear pore | 1.6×10^{-24} |
| 7 | 21 | 19 | 0005794 | 4 | Golgi apparatus | 3.1×10^{-2} |
| 8 | 20 | 70 | 0005685 | 5 | snRNP U1 | 1.1×10^{-33} |
| 9 | 20 | 95 | 0005634 | 3 | nucleus | 4.7×10^{-9} |
| 10 | 19 | 63 | 0030479 | 6 | actin cortical patch | 8.6×10^{-22} |
| 11 | 19 | 63 | 0030120 | 5 | vesicle coat | 2.8×10^{-23} |
| 12 | 18 | 94 | 0005849 | 5 | mRNA cleavage factor complex | 2.8×10^{-45} |
| 13 | 18 | 78 | 0005732 | 4 | small nucleolar ribonucleoprotein complex | 2.3×10^{-25} |
| 14 | 17 | 94 | 0000314 | 6 | organellar small ribosomal subunit | 3.1×10^{-36} |
| 15 | 17 | 71 | 0005669 | 6 | transcription factor TFIIID complex | 5.9×10^{-30} |
| 16 | 17 | 82 | 0005681 | 4 | spliceosome complex | 7.4×10^{-24} |
| 17 | 15 | 47 | 0005885 | 5 | Arp2/3 protein complex | 3.1×10^{-18} |
| 18 | 15 | 40 | 0005852 | 5 | eukaryotic translation initiation factor 3 complex | 9.1×10^{-15} |
| 19 | 15 | 80 | 0000178 | 3 | exosome (RNase complex) | 1.4×10^{-32} |
| 20 | 15 | 100 | 0005666 | 4 | DNA-directed RNA polymerase III complex | 2.3×10^{-42} |
| 21 | 13 | 46 | 0008023 | 5 | transcription elongation factor complex | 9.4×10^{-12} |
| 22 | 12 | 75 | 0016471 | 5 | hydrogen-translocating V-type ATPase complex | 7.0×10^{-22} |
| 23 | 12 | 92 | 0005777 | 5 | peroxisome | 1.4×10^{-21} |
| 24 | 12 | 92 | 0005753 | 4 | proton-transporting ATP synthase complex (sensu Eukaryota) | 3.6×10^{-27} |
| 25 | 12 | 83 | 0030014 | 6 | CCR4–NOT complex | 3.0×10^{-27} |
| 26 | 11 | 100 | 0005680 | 5 | anaphase-promoting complex | 3.0×10^{-30} |
| 27 | 11 | 64 | 0008023 | 5 | transcription elongation factor complex | 6.2×10^{-15} |
| 28 | 10 | 50 | 0005794 | 4 | Golgi apparatus | 3.8×10^{-5} |
| 29 | 10 | 50 | 0005851 | 4 | eukaryotic translation initiation factor 2B complex | 4.9×10^{-14} |
| 30 | 10 | 80 | 0008250 | 5 | oligosaccharyl transferase complex | 1.3×10^{-22} |
| 31 | 10 | 40 | 0005832 | 5 | chaperonin-containing T-complex | 3.1×10^{-8} |
| 32 | 9 | 100 | 0030008 | 6 | TRAPP complex | 2.2×10^{-27} |
| 33 | 9 | 89 | 0005839 | 4 | proteasome core complex (sensu Eukaryota) | 2.3×10^{-20} |
| 34 | 9 | 33 | 0016585 | 5 | chromatin remodeling complex | 2.7×10^{-3} |
| 35 | 9 | 78 | 0000790 | 5 | nuclear chromatin | 1.1×10^{-13} |
| 36 | 9 | 89 | 0005634 | 3 | nucleus | 1.6×10^{-3} |
| 37 | 8 | 38 | 0000118 | 6 | histone deacetylase complex | 2.0×10^{-5} |
| 38 | 8 | 88 | 0005634 | 3 | nucleus | 2.6×10^{-2} |
| 39 | 8 | 88 | 0035097 | 5 | histone methyltransferase complex | 9.0×10^{-21} |
| 40 | 8 | 88 | 0000118 | 6 | histone deacetylase complex | 1.7×10^{-16} |
| 41 | 8 | 75 | 0000808 | 4 | origin recognition complex | 4.7×10^{-18} |
| 42 | 8 | 75 | 0005940 | 5 | septin ring | 4.5×10^{-14} |
| 43 | 8 | 88 | 0005737 | 3 | cytoplasm | 7.3×10^{-1} |

^a The size of each module. ^b The fraction of genes in the module has the same most-significantly enriched GO term. ^c The level of the corresponding GO term below the “root” in the shortest path. ^d The description of the corresponding GO term. ^e The P-value of the most-significantly enriched GO term for cellular components.

300 experiment points. Figure 3 compares the distribution of correlation coefficients between expression profiles obtained from the whole protein–interaction network and that obtained from the modules. It is clear that modules are enriched with coexpressed proteins and depleted with proteins that are not coexpressed [*T*-score and *P*-value are 15.24 and 1.57×10^{-52} , respectively, based on the null hypothesis of no difference between identified modules and the whole network graph].²⁰

Because most discovered modules are highly homogeneous and densely connected, it is tempting to predict the function of *unknown* genes enclosed in modules. For example, module #15 (as listed in Table 1) contains 17 proteins (Figure 4). The 15 out of 17 proteins, except for *PRP45* and *YGL128C*, are involved in the process of nuclear mRNA splicing via spliceosome. *PRP45*, the ortholog of human transcriptional coactivator SKIP, was recently found to be a spliceosome associated pre-mRNA splicing factor.²¹ Thus, it is possible that *YGL128C*, a functionally unknown gene, like all other 16 members in module #15, involves in the process of nuclear mRNA splicing. In another example, module #13 (Figure 5) has 18 genes, all but one (*YNL132W*) are involved in rRNA process-

ing (GO level 7). This suggests that module #13 is a module of rRNA processing pathway and the unknown gene *YNL132W* is likely involved in rRNA processing.

The discovered modules may also be useful to uncover the new role of *known* genes. For example, module #37 is an eight-gene module (Figure 6). According to the current GO annotation, all genes except for *CPR1* belong to the process of establishing and/or maintaining chromatin architecture (GO: 0006325, level 7). Moreover, five of the eight proteins (shown in bold circles) are involved in a more specific process of negative regulation of meiosis (GO:0045835, level 9). *CPR1* was reported as a peptidyl-prolyl cis–trans isomerase (i.e., cyclophilin) involved in protein metabolism.²² The location of *CPR1* in this module suggests its possible role in a chromatin-architecture related process (or more specifically, regulation of meiosis). Indeed, a latest biochemical study and genetic assay did suggest the role of *CPR1* in meiosis controls.²³

Discussion

In this work, we have developed an algorithm, called MC², to discover the modules from a network of interacting proteins

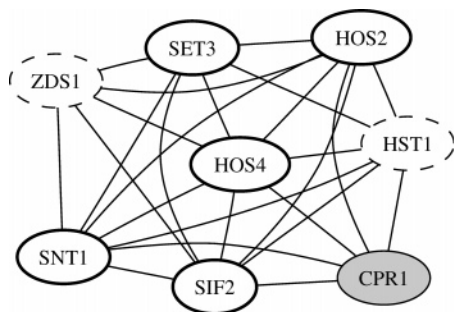


Figure 6. Interaction network of Module #37. Five inner genes (in solid circles) are involved in a more specific process of negative regulation of meiosis (GO:0045835 level 9); All genes except *CPR1* are annotated by GO:0006325 (level 7).

by merging partially overlapped cliques. Application of MC^2 to the yeast protein interaction data produces a list of high-quality modules that are highly homogeneous in function (82% in average) and in cellular locations (78% in average) and are enriched with coexpressed proteins. Detailed inspection of genes contained in modules indicates their usefulness for annotating *unknown* genes and uncovering novel roles played by *known* genes.

Modules identified here may be labeled as the candidate of core components of some biological modules. This is because a real biological module does not have to build on a clique with $Q = 1$. It can also build on a subgraph with Q that is close but not equal to 1. This explains that the total number of genes in all discovered modules here is only 18% of all genes contained in the whole interacting network. Nevertheless, MC^2 allows us to make a number of high-quality predictions that are less susceptible to high false positive rates in large-scale interacting data.^{24,25}

To test the dependence of our results on possible false positives and false negatives in the database of protein–protein interactions, we randomly remove 10% links. We find that this reduction of interactions only makes a small change in the average homogeneity by 1%. Similarly, the average homogeneity is essentially unchanged (less than 1%) after 10% links are randomly added. Thus, the MC^2 algorithm is robust against possible false positives and false negatives.

Furthermore, the obtained modules provide a useful resource to analyze genes (either *known* or *unknown*) that are graph-theoretically close to the modules. The recent work by Krauthammer et al. showed that it was effective to discover novel candidate Alzheimer's disease (AD) genes by looking for genes tightly connected to multiple known AD genes in a literature-derived molecular network.¹¹ Similarly, genes involved in the same specific cellular process in a discovered module may be used as the seeds to discover additional genes related to the same process. To this end, we have built an online server for researchers to further explore any genes of their interest.

Several other methods for discovering modules are also developed. For example, Wu et al.²⁶ predicted functional modules based on the strength of gene functional relationship (measured by combined information from phylogenetic profile, gene neighborhood, and GO assignments). In our work, GO information is used only as an independent information to test modules discovered. Gagneur et al.¹⁰ proposed a modular decomposition method for analyzing the pro-inflammatory tumor necrosis factor- α (TNF- α)/NF κ B transcription factor

Table 3. Overlap between Discovered Modules with Known Complexes

| | predicted ^a | actual ^b | overlap ^c | accuracy (%) ^d | coverage (%) ^e |
|-----------------|------------------------|---------------------|----------------------|---------------------------|---------------------------|
| SM ^f | 716 | 1007 | 432 | 60 (68) | 43 (60) |
| MC ² | 890 | 1581 | 599 | 67 (72) | 38 (56) |

^a The total number of genes covered by the modules predicted by computational methods. ^b The total number of genes in the complexes matched to the modules. ^c The number of genes overlapped between computational and experimental methods. ^d The fraction of overlapped genes in the total number genes of predicted modules. The number in parentheses is the averaged fraction of overlapped genes per complex. ^e The results of SM³ are obtained from <http://web.mit.edu/leonid/modules/allOurClusters.html>. The total number of modules provided by SM is 90.

pathway, and Tornor and Mewes⁷ identified modules via collective and multibody correlations in a genetic network. Hu et al.²⁷ developed CODENSE and applied it to 39 coexpression networks. They found that 76% discovered modules, after second-order clustering, had more than 40% of member genes in a specific GO term at least 5 levels below the root. By comparison, 87% modules discovered by MC^2 have more than 40% of its member genes with a specific GO term at least 5 levels below the root.

Spirin and Mirny (SM)³ developed a method that identifies modules via a combination of clique enumeration, superparamagnetic clustering, Monte Carlo optimization, and further cleaning and merging according to statistical significance. We use the same data set of experimental complexes (compiled by Spirin and Mirny³) to compare predicted modules with the experimentally derived protein complexes. We also compare our results with the merged set of all modules predicted by their methods (<http://web.mit.edu/leonid/modules/allOurCluster>). We map a computationally predicted module to an experimental complex by minimizing the P -value of overlap, P_{overlap} , as defined by Spirin and Mirny.³ Only one mapping is arbitrarily chosen if more than one experimental complexes map to a predicted module with the same P_{overlap} . The mapping between a predicted module to an experimental complex is assessed in terms of accuracy and coverage. Accuracy is the percentage of the number of genes in an experimental complex overlapped with the genes in a predicted module. Coverage is the percentage of the number of modules genes overlapped with the genes in an experimental complex. Table 3 shows the comparison between MC^2 and the SM method. Although SM gives a higher average coverage for each module (60% by SM vs 56% by MC^2), MC^2 has a higher average accuracy in locating correct interacting proteins in a module (72% by MC^2 vs 68% by SM). It is not surprising that MC^2 has a smaller coverage because MC^2 is based on clique merging. The requirement of $Q = 1$ as an initial seed for modules is likely to be overly strict. Another parameter to compare the accuracy of different methods is the fraction of modules which has more than 50% genes overlapping with the genes in the experimental complexes.⁹ This number is 60% for SM and 62% for MC^2 , respectively. In contrast, this success rate for significant overlapping between modules and experimental complexes is low for MCODE developed by Bader and Hogue,⁹ which detects densely connected regions (or module) by a graph theoretic clustering.

Finally, it should be mentioned that MC^2 is a simple, efficient, purely graph-theoretic method. It takes only about 5

min to identify modules from the yeast gene interaction network with a typical PC (Intel Xeon 2.5 GHz CPU and 512 MB memory). Even if we randomly increase the number of interactions from 19 614 by more than 10 times to 200 000, a CPU time of 12 h remains affordable.

Acknowledgment. This work was supported by NIH (Grant Nos. R01 GM 966049 and R01 GM 068530), a grant from HHMI to SUNY Buffalo and by the Center for Computational Research and the Keck Center for Computational Biology at SUNY Buffalo. Y.Z. was also supported in part by a two-base grant from National Science Foundation of China.

Supporting Information Available: The number of modules (Figure S1), the number of proteins per module (Figure S2), the number of genes involved in the modules (Figure S3), average locational homogeneity (Figure S4), and the significance of enrichment of coexpressed genes in predicted modules over the whole network graph (Figure S5). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Hartwell, L. H.; Hopfield, J. J.; Leibler, S.; Murray, A. W. From molecular to modular cell biology. *Nature* **1999**, *402*, C47–C52.
- (2) Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; Alon, U. Network motifs: Simple building blocks of complex networks. *Science* **2002**, *298*, 824–827.
- (3) Spirin, V.; Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12123–12128.
- (4) Watts, D. J.; Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **1998**, *393*, 440–442.
- (5) Goldberg, D. S.; Roth, F. P. Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *100*, 4372–4376.
- (6) Barabasi, A. L.; Oltvai, Z. N. Network biology: understanding the cell’s functional organization. *Nat. Rev. Genet.* **2004**, *5*, 101–113.
- (7) Tornow, S.; Mewes, H. W. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res.* **2003**, *31*, 6283–6289.
- (8) Rives, A. W.; Galitski, T. Modular organization of cellular networks. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 1128–1133.
- (9) Bader, G. D.; Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* **2003**, *4*, 2.
- (10) Gagneur, J.; Krause, R.; Bouwmeester, T.; Casari, G. Modular decomposition of protein–protein interaction networks. *Genome Biol.* **2004**, *5*, R57.
- (11) Krauthammer, M.; Kaufmann, C.; Gilliam, T.; Rzhetsky, A. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer’s disease. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15148–15153.
- (12) Salwinski, L.; Miller, C. S.; Smith, A. J.; Pettit, F. K.; Bowie, J. U.; Eisenberg, D. The Database of Interacting Proteins. *Nucleic Acids Res.* **2004**, *32*, D449–D451.
- (13) Kelley, R.; Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* **2005**, *23*, 561–566.
- (14) Bron, C.; Kerbosch, J. Algorithm 457: finding all cliques of an undirected graph [H]. *Comm. ACM* **1973**, *16*, 575–577.
- (15) Radicchi, F.; Castellano, C.; Ceconi, F.; Loreto, V.; Parisi, D. Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 2658–2663.
- (16) Boyle, E. I.; Weng, S.; Gollub, J.; Jin, H.; Botstein, D.; Cherry, J. M.; Sherlock, G. GO:: TermFinder – open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **2004**, *20*, 3710–3715.
- (17) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **2000**, *25*, 25–29.
- (18) Zhou, X. H.; Kao, M. C. J.; Wong, W. H. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12783–12788.
- (19) Hughes, T. R.; Marton, M. J.; Jones, A. R.; Roberts, C. J et al. Functional discovery via a compendium of expression profiles. *Cell* **2000**, *102*, 109–126.
- (20) Deng, M.; Metah, S.; Sun, F.; Chen, T. Inferring domain-domain interactions from protein–protein interactions. *Genome Res.* **2000**, *12*, 1540–1548.
- (21) Albers, M.; Diment, A.; Muraru, M.; Russell, C. S.; Beggs, J. D. Identification and characterization of Prp45p and Prp46p, essential pre-mRNA splicing factors. *RNA* **2003**, *9*, 138–150.
- (22) Haendler, B.; Keller, R.; Hiestand, P. C.; Kocher, H. P.; Wegmann, G.; Movva, N. R. Yeast cyclophilin: isolation and characterization of the protein, cDNA and gene. *Gene* **1989**, *83*, 39–46.
- (23) Arevalo-Rodriguez, M.; Heitman, J. Cyclophilin A is localized to the nucleus and controls meiosis in *Saccharomyces cerevisiae*. *Eukaryot. Cell* **2005**, *4*, 17–29.
- (24) von Mering, C.; Krause, R.; Snel, B.; Cornell, M.; Oliver, S. G.; Fields, S.; Bork, P. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **2002**, *417*, 399–403.
- (25) Deng, M.; Sun, F.; Chen, T. Assessment of the reliability of protein–protein interactions and protein function prediction. *Pac. Symp. Biocomput.* **2003**, *8*, 140–151.
- (26) Wu, H.; Su, Z.; Mao, F.; Olman, V.; Xu, Y. Prediction of functional modules based on comparative genome analysis and gene ontology application. *Nucleic Acids Res.* **2005**, *33*, 2822–2837.
- (27) Hu, H.; Yan, X.; Huang, Y.; Han, J.; Zhou, X. J. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics Suppl. 1* **2005**, *21*, i213–i221.

PR050366G