IDENTIFICATION OF FUNCTIONAL MODULES IN PROTEIN COMPLEXES VIA HYPERCLIQUE PATTERN DISCOVERY

HUI XIONG[a,*], XIAOFENG HE[b], CHRIS DING[b], YA ZHANG[c], VIPIN KUMAR[a], STEPHEN R. HOLBROOK[b]

[a] *Computer Science & Engineering, University of Minnesota, MN, USA*
E-mail: {huix, kumar}@cs.umn.edu

[b] *Computational Research Division and Physical Biosciences Division*
Lawrence Berkeley National Laboratory, Berkeley, CA, USA
E-mail: {xhe, chqding, srholbrook}@lbl.gov

[c] *Information Sciences and Technology, Penn State University, PA, USA*
E-mail: yzhang@ist.psu.edu

Proteins usually do not act isolated in a cell but function within complicated cellular pathways, interacting with other proteins either in pairs or as components of larger complexes. While many protein complexes have been identified by large-scale experimental studies [7,8], due to a large number of false-positive interactions existing in current protein complexes [10], it is still difficult to obtain an accurate understanding of functional modules, which encompass groups of proteins involved in common elementary biological function. In this paper, we present a hyperclique pattern discovery approach for extracting functional modules (hyperclique patterns) from protein complexes. A hyperclique pattern is a type of association pattern containing proteins that are *highly affiliated* with each other. The analysis of hyperclique patterns shows that proteins within the same pattern tend to present in the protein complex together. Also, statistically significant annotations of proteins in a pattern using the Gene Ontology suggest that proteins within the same hyperclique pattern more likely perform the same function and participate in the same biological process. More interestingly, the 3-D structural view of proteins within a hyperclique pattern reveals that these proteins physically interact with each other. In addition, we show that several hyperclique patterns corresponding to different functions can participate in the same protein complex as independent modules. Finally, we demonstrate that a hyperclique pattern can be involved in different complexes performing different higher-order biological functions, although the pattern corresponds to a specific elementary biological function.

1 Introduction

Complex cellular processes are modular and are accomplished by proteins in complex multi-protein assemblies. Often these multi-protein complexes act as highly efficient protein machines and perform activities related to complex

*Corresponding Author

biological phenomena, such as DNA replication, transcription, metabolism, and signal transduction. A variety of experimental and computational approaches have been employed to deduce the constituents of protein macromolecular complexes. Experimental approaches such as the yeast two-hybrid genetic screen [14,9] yield binary interaction data while more recent large-scale methods [7,8] combine tagged "bait" proteins and protein-complex purification schemes with mass spectrometric measurements to identify protein complexes that contain three or more components.

While proteomic studies [7,8] have generated large amount of interesting protein complex data, much remains to be learned before we have a comprehensive knowledge of functional modules - groups of proteins involved in common elementary biological function. Along this line, an important issue is the effective extraction of functional modules. Previous research on this topic can be grouped into two approaches. One approach is targeted on extraction of densely connected subgraphs from the protein interaction network, such as fully connected subgraphs (cliques)[13] and almost fully connected subgraphs (k-cores)[2]. However, algorithms for finding cliques and k-core are typically quite expensive. Another approach for detection of functional modules is through clustering analysis [5,12], which divide proteins into groups (clusters) in the way such that similar proteins are in the same cluster and dissimilar proteins are in different clusters.

In this paper, we present a hyperclique pattern discovery approach for identifying functional modules (hyperclique patterns) from protein complex data. A hyperclique pattern is a type of association pattern containing proteins that are *highly affiliated* with each other; that is, every pair of proteins within a hyperclique pattern is guaranteed to have the cosine similarity (uncentered Pearson correlation coefficient [†]) above certain level. As a result, our method is more robust than related approaches in the presence of large number of false-positive protein interactions. Indeed, a significant number of false-positive protein interactions are present in current experimentally identified protein complexes. Gavin *et al.* [7] estimate that 30% of the protein interactions they detect may be spurious, as inferred from duplicate analyses of 13 purified protein complexes. Finally, please note that clustering analysis finds related proteins with a global constraint, while hyperclique patterns capture relationships among proteins on a local level and thus are more compact representations of proteins.

Hyperclique pattern discovery is especially effective on protein complex data, because protein complex data can be viewed as a bipartite graph[5] (a ma-

[†]When computing Pearson correlation coefficient, the data mean is not subtracted.

trix in which rows represent protein complex and column represents proteins). In contrast, previous approaches are usually based on a graph of pairwise similarities. A bipartite graph representation of protein complexes allows us to efficiently compute hyperclique patterns, much faster than finding cliques or k-cores in a graph.

The analysis of discovered hyperclique patterns from protein complexes using the Gene Ontology suggests that proteins within the same hyperclique pattern more likely perform the same function and participate in the same biological process. For example, all proteins of an identified pattern {`Pre2, Pre4, Pre5, Pre6, Pre8, Pre9, Pup3, Scl1`} corresponds to the same function annotation "*endopeptidase activity*" and the 3-D structural view of these proteins reveals that they physically interact with each other. Furthermore, we show that several hyperclique patterns with different functions can participate in the same protein complex as independent modules. Finally, we demonstrate that a hyperclique pattern can be involved in different protein complexes performing different higher-order biological functions, although the pattern corresponds to a specific biological function.

2 Hyperclique Pattern Discovery

In this section, we describe the concept of hyperclique patterns [15] after first introducing the concept on which it is based: the association rule [1].

2.1 Association Rules

We present the concept of association rules [1] within the context of biology. Let $P = \{p_1, p_2, \ldots, p_n\}$ be a set of proteins and $C = \{c_1, c_2, \ldots, c_l\}$ be the set of protein complexes, where each complex c_i is a set of proteins and $c_i \subseteq P$. A pattern is a set of proteins $X \subseteq P$, and the **support** of X, $supp(X)$, is the fraction of protein complexes containing X. For example, in Table 1, the support of the pattern $\{p_3, p_4\}$ is $3/5 = 60\%$, since three protein complexes (c2, c3, c4) contain both p_3 and p_4.

An association rule is of the form $X \rightarrow Y$, which means the presence of pattern X implies the presence of pattern Y in the same protein complex, where $X \subseteq P$, $Y \subseteq P$, and $X \cap Y = \phi$. The **confidence** of the association rule $X \rightarrow Y$ is written as $conf(X \rightarrow Y)$ and is defined as $conf(X \rightarrow Y) = supp(X \cup Y)/supp(X)$. For instance, for protein complex data shown in Table 1, the confidence of the association rule $\{p_3\} \rightarrow \{p_4\}$ is $conf(\{p_3\} \rightarrow \{p_4\}) = \text{supp}(\{p_3, p_4\})/\text{supp}(\{p_3\}) = 60\% \ / \ 80\% = 75\%$. In biology domain, there are many interesting patterns occuring at low levels of support, such as the ones

Table 1. A Sample Protein Complex Data Set.

Protein Complex	Proteins
c1	p_1, p_2
c2	p_1, p_3, p_4, p_5
c3	p_2, p_3, p_4, p_6
c4	p_1, p_2, p_3, p_4
c5	p_1, p_2, p_3, p_6

identified in this paper. However, existing association-rule mining algorithms often have difficulties in finding patterns at low levels of support. Also, many patterns discovered by association-rule mining algorithms contain proteins which are poorly correlated with each other.

2.2 Hyperclique Patterns

A hyperclique pattern is a new type of association pattern that contains proteins that are *highly affiliated* with each other; that is, every pair of proteins within a pattern is guaranteed to have the cosine similarity (uncentered Pearson correlation coefficient) above a certain level. Indeed, the presence of a protein in one protein complex strongly implies the presence of every other protein that belongs to the same hyperclique pattern. The h-confidence measure is specifically designed to capture the strength of this association.

Definition 2.1 *The* **h-confidence** *of a pattern $X = \{p_1, p_2, \cdots, p_m\}$, denoted as $hconf(X)$, is a measure that reflects the overall affinity among proteins within the pattern. This measure is defined as $min(conf(\{p_1\} \rightarrow \{p_2, \ldots, p_m\}), conf(\{p_2\} \rightarrow \{p_1, p_3, \ldots, p_m\}), \ldots, conf(\{p_m\} \rightarrow \{p_1, \ldots, p_{m-1}\}))$, where $conf$ is the confidence of association rule as given above.*

Example 2.1 *For the sample protein complex data set shown in Table 1, let us consider a pattern $X = \{p_2, p_3, p_4\}$. We have $supp(\{p_2\}) = 80\%$, $supp(\{p_3\}) = 80\%$, $supp(\{p_4\}) = 60\%$, and $supp(\{p_2, p_3, p_4\}) = 40\%$. Then,*

$$conf(\{p_2\} \rightarrow \{p_3, p_4\}) = supp(\{p_2, p_3, p_4\})/supp(\{p_2\}) = 50\%$$
$$conf(\{p_3\} \rightarrow \{p_2, p_4\}) = supp(\{p_2, p_3, p_4\})/supp(\{p_3\}) = 50\%$$
$$conf(\{p_4\} \rightarrow \{p_2, p_3\}) = supp(\{p_2, p_3, p_4\})/supp(\{p_4\}) = 66.7\%$$

Therefore, $hconf(X) = min(conf(\{p_2\} \rightarrow \{p_3, p_4\}), conf(\{p_3\} \rightarrow \{p_2, p_4\}), conf(\{p_4\} \rightarrow \{p_2, p_3\})) = 50\%$.

Definition 2.2 *A pattern X is a* **hyperclique pattern** *if $hconf(X) \geq h_c$, where h_c is a user-specified minimum h-confidence threshold. A hyperclique pattern is a* **maximal hyperclique pattern** *if no superset of this pattern is also a hyperclique pattern.*

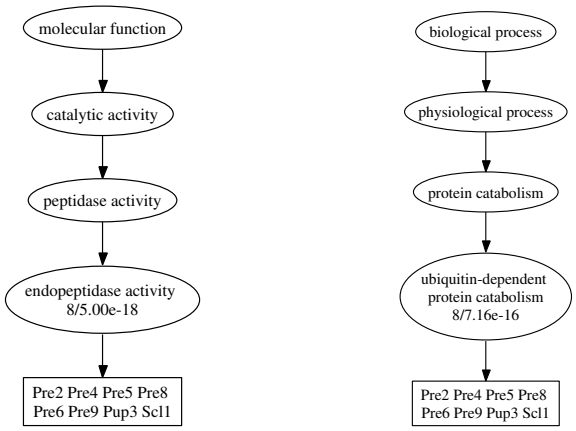

Figure 1. Gene Ontology annotations of pattern {Pre2, Pre4, Pre5, Pre6, Pre8, Pre9, Pup3, Scl1}. Figure on the left shows subgraph of the function annotation. Figure on the right shows subgraph of process annotation. Proteins are shown in square box and significant nodes are labeled with the number of proteins annotated directly or indirectly to that term and the p-value for the term.

Table 2. Examples of Hyperclique Patterns from Yeast Protein Complex Data.

Yeast Protein Complex Data		
Hyperclique patterns	Supp	Hconf
{Cus1, Msl1, Prp3, Prp9, Sme1, Smx2, Smx2, Smx3, Yhc1}	1.25%	100%
{Pre2, Pre4, Pre5, Pre6, Pre8, Pre9, Pup3, Scl1}	1.7%	66.7%
{Cwc2, Ecm2, Hsh155, Prp19, Prp21, Snt309}	1.7%	100%
{Emg1, Imp3, Imp4, Kre31, Mpp10, Nop14, Sof1, Utp15, Noc4}	1.25%	100%

Let us consider the sample protein complex data in Table 1. For the h-confidence threshold 0.5, the pattern $\{p_2, p_3, p_4\}$ is a hyperclique pattern. Furthermore, since no superset of this pattern is a hyperclique pattern at the threshold 0.5, this pattern is also a maximal hyperclique pattern.

Table 2 shows some hyperclique patterns identified from a yeast protein complex data set [7]. One hyperclique pattern in that table is {`Pre2, Pre4, Pre5, Pre6, Pre8, Pre9, Pup3, Scl1`}. Figure 1 shows the function and process subgraphs of the Gene Ontology corresponding to this pattern. One observation is that all proteins in this pattern perform the same biological function, *endopeptidase activity*. Also, all proteins in the pattern involve in the same biological process, *ubiquitin-dependent protein catabolism*. In other words, proteins in the same hyperclique pattern are highly-affiliated with each other. Indeed, the following Theorem 2.1 guarantees if a hyperclique pattern

has an h-confidence value above the h-confidence threshold, h_c, then every pair of proteins within the pattern must have a cosine similarity (uncentered Pearson correlation coefficient) greater than or equal to h_c.

Theorem 2.1 *Given a hyperclique pattern $X = \{p_1, p_2, \ldots, p_m\}$ at the h-confidence threshold h_c, for two proteins p_l and p_k such that $\{p_l, p_k\} \subset X$, we have $cosinesim(p_l, p_k) \geq h_c$, where $cosinesim(p_l, p_k) = \frac{supp(\{p_l, p_k\})}{\sqrt{supp(\{p_l\})supp(\{p_k\})}}$, which is the cosine similarity between p_l and p_k.*

2.3 Computation Algorithm

In a nutshell, the process of searching hyperclique patterns can be viewed as the generation of a level-wise pattern tree. Every level of the tree contains patterns with the same number of proteins. If the level is increased by one, the pattern size (number of proteins) is also increased by one. Every pattern has a branch (sub-tree) which contains all the superset of this pattern. Our algorithm for finding hyperclique patterns is breath-first. We first check all the patterns at the first level. If a pattern is not satisfied with the user-specifed support and h-confidence thresholds, the whole branch corresponding to this pattern can be pruned without further checking. This is due to the the anti-monotone property of support and h-confidence measures. Consider the h-confidence measure, the anti-monotone property guarantees that the h-confidence value of a pattern is greater than or equal to that of any superset of this pattern. Following this manner, the pattern tree is growing level-by-level until all the patterns have been generated. This algorithm is very efficient for handling large-scale datasets [15].

3 Protein Complex Data and Analysis Tools

Protein Complex Data: Two datasets [7,8] summarizing large-scale experimental studies of multi-protein complexes are available for the yeast *Saccharomyces cerevisiae*. Coupling different purification (immunoprecipitation and tandem affinity purification (TAP)) and labeling schemes with mass spectrometry (MS), both studies used bait proteins to identify physiologically intact protein complexes. Independent research[4,11] showed that the TAP-MS dataset by Gavin, *et al.* [7] has a relatively better accuracy for predicting protein functions, therefore we take this dataset to illustrate our method. In this TAP-MS dataset, there are a total of 1,440 distinct proteins within 232 multi-protein complexes, and the data format is illustrated in Table 1.

Analysis Tools: The Gene Ontology (http://www.geneontology.org) was used to annotate the proteins of hyperclique patterns identi-

fied in the TAP-MS dataset. A graph drawing package GraphViz (http://www.research.att.com/sw/ tools/graphviz/) was used to produce the graph representation of the annotation. The functional description of each protein (if available) was obtained from the Saccharomyces Genome Database (SGD)[6]. The 3-D structure information of yeast proteins was obtained from the Protein Data Bank (PDB) (http://www.rcsb.org/pdb), and PyMOL[3] was used for visualizing the 3-D structure of proteins within a hyperclique pattern.

4 Analysis of Hyperclique Pattern using Gene Ontology

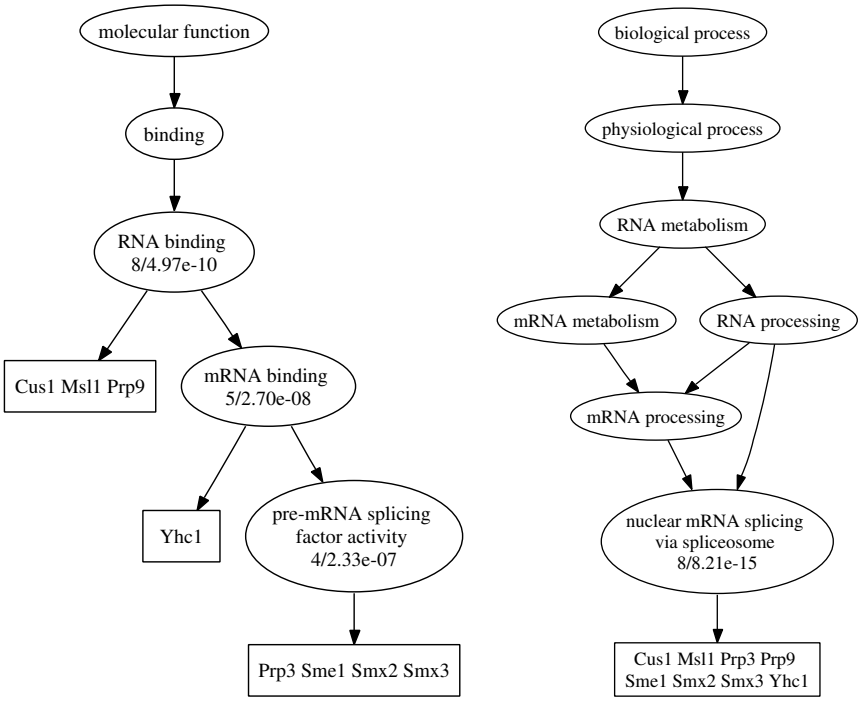

Figure 2. The Gene Ontology annotations of pattern {Cus1, Msl1, Prp3, Prp9, Sme1, Smx2, Smx3, Yhc1}. Figure on the left shows subgraph of function annotation of the pattern. Figure on the right shows subgraph of process annotation. Proteins are listed in square box. Significant nodes are labeled with the number of proteins annotated directly or indirectly to that term and the p-value for the term.

Setting a support threshold to be 0 and an h-confidence threshold to be 0.6, we obtained 60 maximal hyperclique patterns. Limited by space, we

analyze some of the patterns obtained. Detailed results are available at our project web site [‡] .

The proteins within the same hyperclique pattern have strong association with each other. To investigate this, we analyze the annotations of the patterns using the terms from the Gene Ontology. Figure 2 shows the subgraphs of the Gene Ontology corresponding to pattern {`Cus1, Msl1, Prp3, Prp9, Sme1, Smx2, Smx3, Yhc1`}. The left subgraph in the figure is the molecular function annotation of the proteins in the pattern. Note that all 8 proteins from this pattern are annotated to the term *RNA binding* with p-value 4.97e-10. The p-value is calculated as the probability that n or more proteins would be assigned to that term if proteins from the entire genome are randomly assigned to that pattern. The smaller the p-value, the more significant the annotation. Among the pattern, 4 proteins {`Prp3, Sme1, Smx2, Smx3`} are annotated to a more specific term *pre-mRNA splicing factor activity* with p-value 2.33e-07. The annotation of these proteins confirms that each pattern form a module performing specific function. The right subgraph in Figure 2 shows the biological process this pattern is involved. The proteins are annotated to the term *nuclear mRNA splicing via spliceosome* with p-value 8.21e-15 which is statistically significant.

Table 3. The Hyperclique pattern {Pre2, Pre4, Pre5, Pre8, Pup3, Pre6, Pre9, Scl1} contained in four protein complexes. All proteins in the pattern are in bold.

CID	Protein Complexes	Function Category
106	Blm3 Dam1 Dbp9 Ecm29 Est3 Gfa1 Ino4 Kap95 Lys12 Mds3 Nud1 Pda1 Pdb1 Pre10 **Pre2** Pre3 **Pre4 Pre5 Pre6 Pre8 Pre9** Pse1 **Pup3** Rgr1 Rpt3 Rpt5 **Scl1** Spa2 Srp1 Ulp1 YFL006W YGR081C YMR310C YPL012W Yra1	Protein Synthesis and Turnover
148	Cdc6 Ecm29 Gfa1 Mlh2 Nas6 Pgk1 Pre1 **Pre2** Pre3 **Pre4 Pre5 Pre6** Pre7 **Pre8 Pre9 Pup3** Rpn10 Rpn11 Rpn12 Rpn13 Rpn3 Rpn5 Rpn6 Rpn7 Rpn8 Rpn9 Rpt1 Rpt2 Rpt3 Rpt4 Rpt5 Rpt6 **Scl1** Ubp6	Protein Synthesis and Turnover
157	Blm3 Cdc6 Ecm29 Mlh2 Pgk1 Pre1 Pre10 **Pre2** Pre3 **Pre4 Pre5 Pre6** Pre7 **Pre8 Pre9 Pup3** Rgr1 Rpn10 Rpn11 Rpn12 Rpn13 Rpn3 Rpn5 Rpn6 Rpn7 Rpn8 Rpn9 Rpt1 Rpt2 Rpt3 Rpt4 Rpt5 Rpt6 **Scl1** Ubp6 YFL006W	Protein Synthesis and Turnover
151	Blm3 Cdc55 Cin1 Erg13 Hhf2 Hos2 Iml1 Kap95 Kel1 Lte1 Myo5 Pfk1 Pph21 Pph22 Pre1 Pre10 **Pre2 Pre4 Pre5 Pre6** Pre7 **Pre8 Pre9** Pup1 Pup2 **Pup3** Rrd2 Rts1 **Scl1** Sif2 Srp1 Tdh2 Tdh3 Tef4 Tpd3 YBL104C YCR033W YGL245W YGR161C YIL112W YKR029C Yef3 Yor1 Yra1 Zds1 Zds2	Signalling

[‡]http://www.cs.umn.edu/~huix/pfm/pfm.html

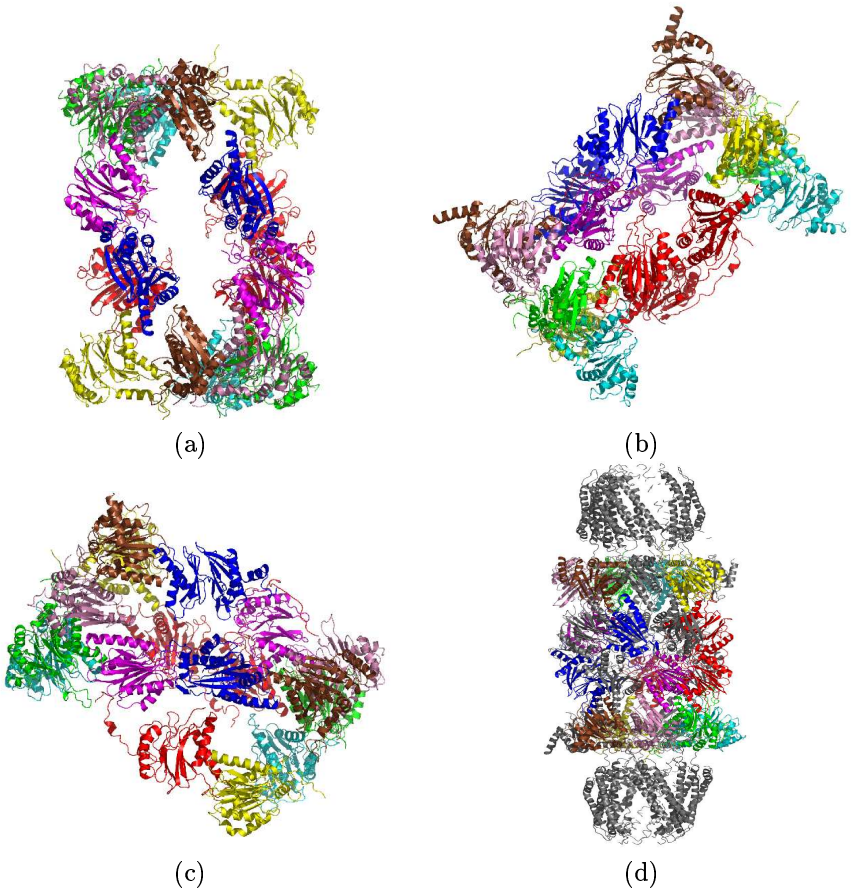

Figure 3. The 3-D structure of the yeast proteasome including all proteins in the hyperclique pattern {Pre2, Pre4, Pre5, Pre8, Pup3, Pre6, Pre9, Scl1}. In the figure, (a), (b), and (c) show 3-D structures of proteins only in the pattern. (Pre2(blue), Pre4(red), Pre5(yellow), Pre6(brown), Pre8(green), Pre9(pink), Pup3(magenta), Scl1(cyan)). In contrast, (d) shows 3-D structures of all proteins in the proteasome complex.

5 Hyperclique Patterns as Functional Modules

Gene Ontology annotations reveal that proteins in the same hyperclique pattern tend to perform a common function and be involved in the same biological process. In this subsection, we describe the role of hyperclique patterns as functional modules.

Consider the hyperclique pattern {`Pre2, Pre4, Pre5, Pre6, Pre8,`

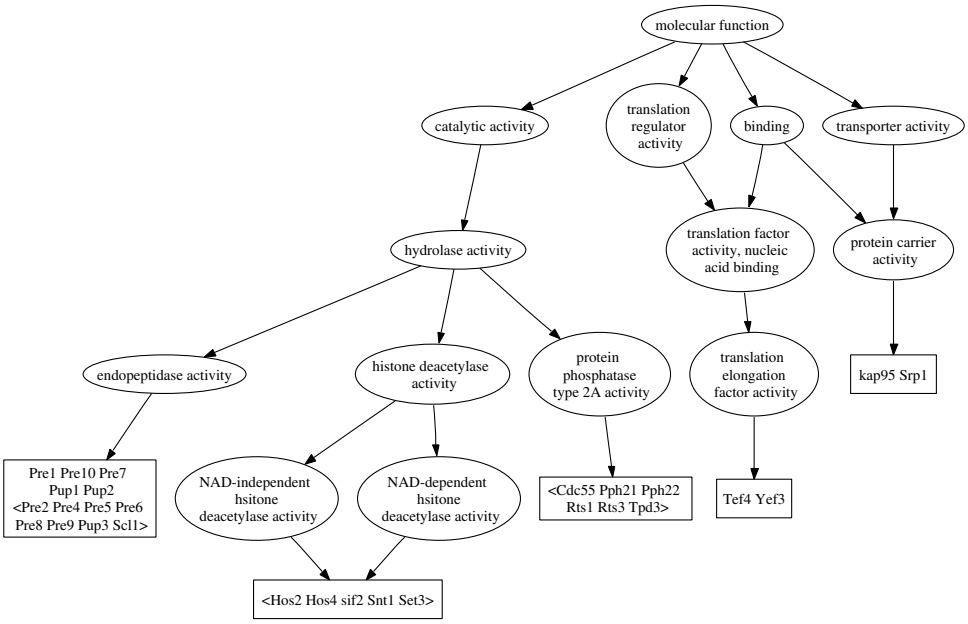

Figure 4. Subgraph of the Gene Ontology (function) corresponding to the protein complex 151. Proteins within a pair of $< >$ form a hyperclique pattern.

`Pre9, Pup3, Scl1`}. All proteins in this pattern are components of the proteasome complex which destroys the proteins no longer in use or the failures of the translation products. The proteasome constitutes nearly 1% of cellular proteins. The 3-D structure of this complex (PDB ID: 1fnt) is available in the Protein Data Bank (PDB). Figure 3 (d) shows the 3-D structure of all proteins in the proteasome. Figures 3 (a), (b), and (c) show the 3-D structure of all proteins in the hyperclique pattern from different view angles. As can be seen, proteins in the pattern have physical interactions with each other. This is compelling physical evidence implying that proteins in the same hyperclique pattern tend to physically interact together to form a compact structure and perform a common molecular function. Figure 1 illustrates the molecular function and biological process of this pattern. It is also interesting to observe that this hyperclique pattern is contained in four protein complexes in the TAP-MS data set, as shown in Table 3. According to Gavin's function category, these four protein complexes belong to two different function categories: *protein synthesis and turnover* and *signalling*. In other words, this hyperclique pattern acts as a functional module participating in protein complexes which perform different high-order functions.

Furthermore, we observed that three identified hyperclique patterns are contained in the protein complex 151 (refer to Table 3). Figure 4 shows the subgraph of the Gene Ontology (function) corresponding to the protein complex 151. As can be seen, three hyperclique patterns correspond to three different functions: the pattern {`Pre2, Pre4, Pre5, Pre6, Pre8, Pre9, Pup3, Scl1`} corresponds to the function *endopeptiase activity*, the pattern {`Hos2, Hos4, Sif2, Snt1, Set3`} corresponds to the function *histone deacetylase activity*, and the pattern {`Cdc55, Php21, Php22, Rts1, Rts3, Tpd3`} corresponds to the function *Protein phosphatase type 2A activity*. This indicates that hyperclique patterns can serve as different functional modules to participate in a common protein complex.

6 Discussion

In this paper, we describe a hyperclique pattern discovery approach to identify functional modules in protein complex data. The tight threshold in the definition of hyperclique patterns ensures the strong associations among the proteins in the same functional module. Analysis using the Gene Ontology indicates that the computationally discovered hyperclique patterns are biologically significant. Our approach can not only effectively identify the basic functional modules in protein complexes, but also is robust in the presence of large number of false-positive protein interactions, due to the strong associations among the constituent proteins.

Our work discovered several interesting protein functional modules. For example, one discovered protein functional module {`Pre2, Pre4, Pre5, Pre8, Pup3, Pre6, Pre9, Scl1`} focuses on the function *endopeptidase activity* by the Gene Ontology. This hyperclique pattern with specific function is also found to exist in four experimentally determined protein complexes performing different higher level biological functions.

Acknowledgments

This work was supported by U.S. Department of Energy, Office of Science, Office of Laboratory Policy and Infrastructure, through an LBNL LDRD, under contract DE-AC03-76SF00098. This work was also supported by NSF grant # IIS-0308264, NSF ITR #0325949, and by Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory cooperative agreement number DAAD19-01-2-0014. The content of this work does not necessarily reflect the position or policy of the government and no official endorsement should be inferred.

References

1. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases., In *ACM SIGMOD*, 1993.
2. G.D. Bader and C.W. Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnology*, 2002.
3. W.L. Delano. *The PyMOL User's Manual.* DeLano Scientific, 2002.
4. M. Deng, F. Sun, and T. Chen. Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac Symp Biocomput*, pages 140–151, 2003.
5. C. Ding, X. He, R.F. Meraz, and S.R. Holbrook. A unified representation of multi-protein complex data for modeling interaction networks. *Proteins: Structure, Function, and Genetics, to appear*, 2004.
6. S. S. Dwight, M. A. Dolinski, and K. Ball et al. Saccharomyces genome database (sgd) provides secondary gene annotation using the gene ontology (go). *Nucleic Acids Research*, 2002.
7. A. Gavin et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.
8. Y. Ho et al. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415:180–183, 2002.
9. T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two hybrid analysis to explore the yeast protein interaction. in *Proc. Natl. Acad. Sci.*, 98(8):4569–4574, 2001.
10. A. Kumar and M. Snyder. Protein complexes take the bait. *Nature*, 415:123–124, 2002.
11. C. Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417, 2002.
12. J.B. Pereira-Leal, A.J. Enright, and C.A. Ouzounis. Detection of functional modules from protein interaction networks. *PROTEINS: Structure, Function, and Bioinformatics*, 54:49–57, 2004.
13. V. Spirin and L.A. Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, 100(21):12123–12128, October 2003.
14. P. Uetz, L. Cagney, and G. Mansfield et al. A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*. *Nature*, 403:623–627, 2000.
15. H. Xiong, P. Tan, and V. Kumar. Mining strong affinity association patterns in data sets with skewed support distribution. In *Proc. of the IEEE International Conference on Data Mining*, pages 387–394, 2003.