International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:2, 2008

Weighted Clustering Coefficient for Identifying Modular Formations in Protein-Protein Interaction Networks

Zelmina Lubovac, Björn Olsson, and Jonas Gamalielsson

***Abstract*—**This paper describes a novel approach for deriving modules from protein-protein interaction networks, which combines functional information with topological properties of the network. This approach is based on weighted clustering coefficient, which uses weights representing the functional similarities between the proteins. These weights are calculated according to the semantic similarity between the proteins, which is based on their Gene Ontology terms. We recently proposed an algorithm for identification of functional modules, called SWEMODE (Semantic WEights for MODule Elucidation), that identifies dense sub-graphs containing functionally similar proteins. The rational underlying this approach is that each module can be reduced to a set of triangles (protein triplets connected to each other). Here, we propose considering semantic similarity weights of all triangle-forming edges between proteins. We also apply varying semantic similarity thresholds between neighbours of each node that are not neighbours to each other (and hereby do not form a triangle), to derive new potential triangles to include in module-defining procedure. The results show an improvement of pure topological approach, in terms of number of predicted modules that match known complexes.

***Keywords*—**Modules, systems biology, protein interaction networks, yeast.

I. Introduction

MOLECULAR biology is becoming a highly modular science where functional modules are considered to be a critical level of biological organization. The term "module", as understood in molecular biology, was originally defined as a discrete unit with a function that is separable from those of other modules [1]. Furthermore, modularity refers to clusters of elements that work in a co-operative fashion to achieve some defined function. Protein complexes constitute one example type of module, since the proteins within a complex interact functionally and physically to form a robust unit, which in its turn carries out some biological function [2].

The clustering coefficient measures the local cohesiveness around a node, and it is defined, for any node *i*, as the fraction of neighbours of *i* that are connected to each other [3]. Simply stated, the clustering coefficient *clust*(*i*) reflects the presence of 'triangles' which have a corner at *i* (see the triangle with dashed sides in Fig. 1). The clustering coefficient is also useful in measuring the global density of triangles in the network as a whole. In previous work, we have compared the clustering coefficient with its weighted counterpart, to characterize global properties of the network [4]. Weight of the link between a pair of proteins reflects the functional strength of the interaction, defined as semantic function similarity between those proteins. Here, we apply a novel approach of combining functional information with topological properties of the network to reveal modular formations.

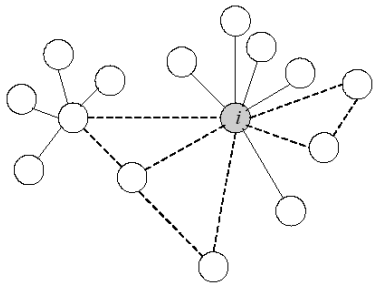

Fig. 1 The clustering coefficient *clust*(*i*) corresponds to the number of "triangles" (see triangles with dashed sides) incident to node *i* (filled circle)

We define modules as dense regions of the PIN (Protein Interaction Network), which contain functionally related proteins. In SWEMODE [4], we proposed semantic function-weighted clustering coefficient, which takes into consideration the functional similarity between interacting proteins. In this study we employ a novel approach for weighting nodes, by considering semantic similarity weights of all of the triangle-forming edges. To our best knowledge, there exist no other methods for identifying modules with semantic similarity-weighted clustering coefficient, apart from our earlier publication [4].

II. Related Work

Although the clustering coefficient is a good measure of the density of interactions in a protein interaction sub-graph, it is

Manuscript received June 30, 2006.

Z. Lubovac is with the School of Humanities and Informatics, University of Skövde, P.O. Box 408, 541 28 Skövde, Sweden (correspondence: phone: +46(0)500 448305; fax: 046(0)500 448399; e-mail: zelmina.lubovac@ his.se).

B. Olsson, is with the School of Humanities and Informatics, University of Skövde, P.O. Box 408, 541 28 Skövde, Sweden (e-mail: bjorn.olsson@his.se).

J. Gamalielsson is with the School of Humanities and Informatics, University of Skövde, P.O. Box 408, 541 28 Skövde, Sweden (e-mail: jonas.gamalielsson@his.se).

strongly dependent on the size of the sub-graph. This makes it very difficult to use clustering coefficient to discern sub-graphs for which the density is statistically significant. Spirin and Mirny [2] observed that the majority of cliques of size four or greater are statistically significant in protein networks compared with random graphs. Such enrichment in the number of cliques reveals essential modularity in the network structure, suggesting that many of these protein interactions are responsible for the formation of complexes and functional modules [2].

Several other methods of network clustering have been applied to reveal modular organization in PINs.[5-7] Furthermore, an algorithm for Molecular COmplex DEtection (MCODE), based on a local network density function named core-clustering coefficient, has been applied earlier to find clusters corresponding to molecular complexes [8]. However, those methods have mostly been focused on topological properties of the network. On the other hand, there are successful approaches for functional grouping of genes based solely on their functional annotation from Gene Ontology [9]. We developed a module-identifying algorithm, SWEMODE (Semantic WEights for MODule Elucidation) [4], based on a weighting scheme according to semantic similarity between the proteins. SWEMODE takes advantage of two aspects of functional annotation encoded in Gene Ontology, molecular function and biological process, and combines these with topological properties of the protein network. In this work, we develop a weighted counterpart, i.e. weighted core-clustering coefficient, which takes into consideration functional weights of all triangle forming edges. Weighted clustering coefficient that we employ here has been proposed in [10] for characterizing weighted financial and metabolic networks with motif intensity scores. We employ this clustering coefficient in a novel way by combining it with semantic similarity weights. *K*-cores have been proposed earlier for detection of protein complexes from protein interaction network.[8, 11] It has also been found recently that proteins that participate in central cores have more vital functions and higher probability of being evolutionary conserved than the proteins that participate in more peripheral cores [12], which motivated the use of this aspect in SWEMODE.

III. Method

A. Protein Interaction Network

Information on protein interactions was downloaded from the Database of Interacting Proteins (DIP[1]),[13] which contains experimentally determined interactions between proteins in *Saccharomyces cerevisiae*, the majority of which were identified with high-throughput Y2H.[14] In Y2H technology, a bait protein, fused to a DNA-binding domain, is used to attract a potential binding protein (prey), fused to a transcriptional activation domain. If the bait and the prey protein interact, their DNA-binding domain and activation domain will combine to form a transcriptional activator, resulting in the expression of a reporter gene.

B. Semantic Similarity Weights

The Gene Ontology (GO)[15] is becoming a *de facto* standard for annotation of gene products. Several methods have used GO to predict the function of hypothetical proteins from protein-protein interaction graphs [16, 17]. We use this measure to assign two weights to each protein-protein interaction, corresponding to semantic similarities between the interacting proteins. One weight is based on annotation from the GO sub-ontology for molecular function, and the other on the sub-ontology covering biological process.

We calculate semantic similarity using the information theoretic measure originally proposed by Lin,[18] which is here calculated using the GO terms assigned to the proteins in the *Saccharomyces* Genome Database (SGD[2]).[19] To calculate the semantic similarity between two gene products, the probability of each term assigned to any of the gene products is first derived. This probability is calculated by counting the number of times the term or any of its descendants occur in SGD annotations, divided by the total number of GO term annotations in SGD. The probability increases as we move towards the root of GO, has probability 1. Given these probabilities, there are several ways to calculate semantic similarity [18, 20, 21].

In order to calculate the similarity between two proteins *i* and *j*, we need to calculate the similarity between the terms belonging to the term sets T_i and T_j that are used to annotate these proteins. We use Lin's similarity measure for calculating term-term similarity. Given the ontology terms $t_k \in T_i$ and $t_l \in T_j$, the semantic similarity is defined as: [18]

$$sim(t_k, t_l) = 2 \ln p_{ms}(t_k, t_l) / \ln p(t_k) + \ln p(t_l) \qquad (1)$$

where $p(t_k)$ is the probability of term t_k and $p_{ms}(t_k,t_l)$ is the probability of the *minimum subsumer* of t_k and t_l, which is defined as the lowest probability found among the parent terms shared by t_k and t_l.[22] We use the average term-term similarity [22] since each protein can be annotated by several terms, and since we are here interested in the overall similarity between the pair of proteins rather than between pairs of individual ontology terms. Given two proteins, *i* and *j*, with T_i and T_j containing *m* and *n* terms, respectively, the protein-protein similarity is defined as the average inter-set similarity between terms from T_i and T_j:

$$ss_{ij} = \frac{1}{m \times n} \sum_{t_k \in T_i, t_l \in T_j} sim(t_k, t_l) \qquad (2)$$

where $sim(t_k, t_l)$ is calculated using (1).

[1] http://dip.doe-mbi.ucla.edu

[2] http://genome-www.stanford.edu/Saccharomyces/

C. Clustering Coefficients

Consider an undirected protein-protein interaction graph with binary weights $w_{ij} = \{0,1\}$ where 1 denotes an interaction and 0 denotes non-interaction. The clustering coefficient for node *i* is defined as [3]:

$$clust(i) = 2n_i / k_i(k_i - 1) \quad (3)$$

where n_i is the number of triangles incident to node *i*, and *k* is the number of direct neighbours of node *i*. This measure, although providing a signature of structural organisation of networks, is based solely on topological grounds. However, it has been shown that inclusion of weights may change our view of structural organisation [4, 23]

Recently, a few extensions of the topological clustering coefficient have emerged for weighted networks. The weighted clustering coefficient, proposed by Barrat et al. [23], has been applied to two types on networks, the world-wide airport network and the scientist collaboration network. It is defined as [23]:

$$wclust_B(i) = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2} a_{ij} a_{ih} a_{jh} \quad (4)$$

where $s_i = \sum_j w_{ij} a_{ij}$ denotes the strength of nodes in terms of the total weight of their interactions, and a_{ij} is an element in the underlying adjacency matrix. We introduced the notion of combining semantic similarity weights with topological protein-protein interactions by using this measure for the purpose of identifying modular formations in protein networks [4]. Hence, the weight w_{ij} equals semantic similarity ss_{ij}, and strength *s* is defined as functional strength of a node, i.e. the sum of all semantic similarities between a protein and its immediate neighbours.

According to the definition, $wclust_B(i)$ only considers the weights of the triangle forming edges adjacent to node *i*, but not the edges connecting the neighbours of *i*.

There are several reasons for considering all triangle-forming edges in the analysis of protein interaction networks. Data obtained from high-throughput Y2H screens is prone to errors, and may contain large numbers of false positives. Furthermore, as mentioned earlier, small cliques, (see examples of such graphs in Fig. 2) are more likely to emerge by chance than large ones [2]. However, as the weighted clustering coefficient by Bader et al. [8] does not differ from the general clustering coefficient for small sub-graphs, we have employed a novel approach for combining semantic similarity weights with topological information, which considers all three edges of the triangles.

The approach is based on the weighted clustering coefficient $wclust_O(i)$ which has been proposed in [10] for characterising weighted financial and metabolic networks with motif intensity scores. We propose combining the semantic similarity weights

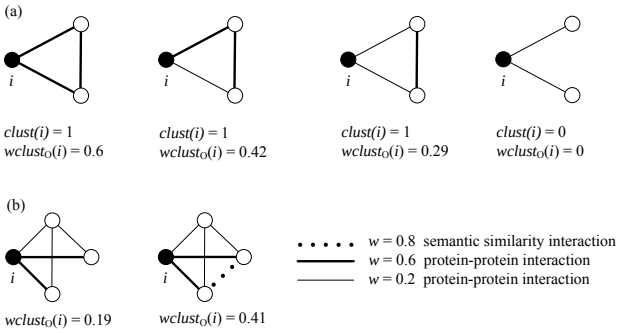

Fig. 2 Illustration of differences between clustering coefficients

between protein-protein interactions with an underlying adjacency matrix, which may vary depending on different semantic similarity thresholds. The original definition by [10] has been adopted accordingly:

$$wclust_O(i) = \frac{2}{k_i(k_i - 1)} \sum_{j,h} (ss_{ij} ss_{ih} ss_{jh})^{1/3} a_{ij} a_{ih} a_{jh} \quad (5)$$

where ss_{ij} is semantic similarity as defined in (2). The properties of a graph are expressed by its adjacency matrix a_{ij}, the entries of which are assigned with 1 if there is an edge between nodes *i* and *j* (corresponding to a protein-protein interaction), otherwise 0.

The advantage of weighted clustering coefficient compared to its topological counterpart is illustrated in Fig. 2. Fig. 2b shows gradually decreasing weights between triplets of proteins. The values of general clustering coefficient $clust(i)$ drop from 1, which is the maximum value, to 0 for the fourth triplet, where there is no link between neighbours adjacent to *i*. Furthermore, for a node *i*, if there is no edge between a pair of neighbours of *i*, and their semantic similarity exceeds a given threshold *t*, we consider this as an edge. For example, in Fig. 2b, there is a missing edge between a pair of neighbours of *i*, and the semantic similarity (see stretched line) between those proteins is 0.8. By setting *t* to 0.5, for example, the adjacency matrix is assigned with 1, assuming an edge between those proteins. This modification results in a considerable increase of $wclust_O(i)$ (from 0.19 to 0.41 in the example).

It should also be noted that we calculate two semantic similarity values for each node, one based on GO molecular function, and the second based on GO biological process. We then use the higher of the two as the final weight of the interaction. This gives the added advantage of taking both aspects into consideration.

D. The Algorithm for Module Identification

The aim of SWEMODE is to identify highly interconnected sub-graphs with high functional homogeneity. We call those sub-graphs core modules. In previous work, [8] Bader and

Hogue developed an algorithm for finding complexes in large-scale networks, called MCODE, which is based on the weighting of nodes with a core-clustering coefficient, which is the density of the highest *k*-core of the closed neighbourhood $N[i]$. The highest *k*-core of a graph is the central most densely connected sub-graph. In previous work, we have proposed an alternative method, called SWEMODE (Semantic WEights for MODule Elucidation), that is used for deriving functional modules, based on the weighted cohesiveness of the sub-graphs. [4] Here, we use a novel weighting scheme, based on weighted core-clustering coefficient of node *i*, $core\_wclust_O(i)$, which is defined as the weighted clustering coefficient of the highest *k*-core of the closed neighbourhood $N[i]$. The use of weighted core-clustering, instead of weighted clustering coefficient, is advantageous since it amplifies the importance of tightly interconnected regions, while removing many less connected nodes that are usually present in protein neetworks.[8] The relative weight assigned to node *i*, based on this measure, is the product of the weighted core-clustering coefficient and the highest *k*-core level of the immediate neighbourhood of *i*.

The second stage of the algorithm, i.e. core module prediction, is similar to the molecular complex prediction step of MCODE.[8] It uses the node weights, seeds a module with the highest weighted node, and then traverse the immediate neighbourhood of the seed node, identifying neighbours whose weights satisfy the node weight percentage (*NWP*) requirement in the module. This module prediction procedure is repeated using the node with the second highest weight as seed for a new module, and so on until the end of the node ranking. The requirement for inclusion of the neighbours in a module is that their weights are higher than a threshold, which is a given *NWP* of the seed node.[8] At this stage, ones the node has been visited and added to the complex, they can not be added to other complex.[4] However, in post-processing step, some overlap is allowed.

In a post-processing step, modules may be filtered according to their connectivity, i.e. the user can choose to remove modules both before and after applying so called "fluffing" step. We perform filtering of all modules containing less than 2 elements before and after fluffing. Fluff parameter that is used to introduce overlapping modules, and can vary between 0.0 and 1.0. [8] For every node in the module, its immediate neighbours are added to the module, if they have not been visited and if their neighbourhood weighted cohesiveness is higher that the given fluff threshold *f*. Fluffing step has been applied both on filtered modules and modules where no filtering parameter was applied. Analysis is based on the results from approximately 440 different parameter settings.

IV. Results

A. Evaluation of SWEMODE Using MIPS Complexes

SWEMODE was used to predict functional modules in the CORE data set. Resulting modules were then compared to the MIPS data set of known protein complexes. The MIPS[3] protein complex catalogue is a curated set of manually annotated yeast protein complexes derived from literature scanning. After removal of 44 complexes that contain only one member, 212 complexes were left in the data set. MIPS complex data set is however incomplete, which may have affected the presented outcome in terms of the number of matched complexes. For example, the complex containing Lsm-proteins, which has the highest rank in our evaluation (see section *Module Ranking with Density Score*), is not present in the MIPS complex data set, although it is a well-known complex.[24] Furthermore, a module may consist of a protein complex and some additional proteins that interact with the complex to perform a distinct function. Even though the MIPS complex data set is incomplete, it is currently the best available resource for protein complexes that we are aware of.

SWEMODE was run using the weighting scheme $core\_wclust_O(i)$ based on the combination of two GO aspects, GO molecular function and GO biological process, over a range of 20 *NWP* parameter values (0 to 0.95 in increments of 0.05). Fluff threshold parameter was also varied between 0 and 1 (in increments of 0.1). In previous work, it has been found that combination of GO biological process and GO molecular function was most suitable for prediction of modules [4], which is why we have not considered each aspect separately here.

To evaluate the performance of SWEMODE and choose the best parameter settings when using $core\_wclust_O(i)$, we used the overlap score [8]. Overlap score, *O*, is defined as [8]:

$$O_{ij} = \left|M_i \cap M_j\right|^2 \Big/ \left|M_i\right|\left|M_j\right| \qquad (6)$$

where M_i is the predicted module, and M_j is a module from the MIPS complex data set. The *O* measure assigns a score of 0 to modules that have no intersection with any known complex, while modules that exactly match a known complex get the score 1. The measure is not so sensitive to a size of modules, meaning that predicted module that fully overlap with a MIPS complex, but is much larger or smaller than MIPS complex will get a low *O*. The best choice of parameters for SWEMODE is the one that predicts the largest number of modules that match the largest number of MIPS protein complexes. Hence, the overlap score may be seen as a measure of biological significance of the module prediction, assuming that the set of complexes obtained from MIPS is biologically plausible.

We have first analysed the effect of using $core\_wclust_O(i)$ with varying semantic similarity threshold *t* on the number of predicted and matched modules. We tested following thresholds *t* on semantic similarity values: 0.1, 0.3, 0.5, 0.7, and 0.9 (Fig. 3). The numbers of matches at each threshold level are not based on the best parameter setting, but are the

[3] http://mips.gsf.de/proj/yeast/

average numbers from 440 different parameter settings. As *O* increases, fewer predicted complexes match known complexes.

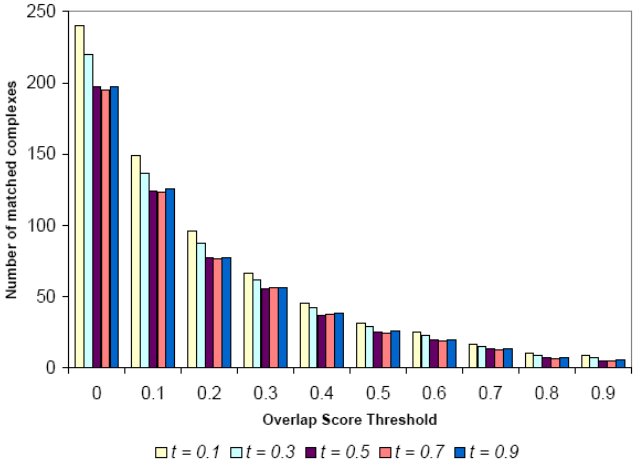

Fig. 3 The effect of varying semantic similarity threshold on the number of matched complexes

The best result is obtained at lowest threshold value $t = 0.1$, implying that even low semantic similarity between proteins is useful to consider in the weighting scheme presented here.

Furthermore, we compared the best result from *core_wclust*$_O$, when threshold was varied, with two other weighting schemas, *core_clust* and *core_wclust*$_O$(*top*). As mentioned earlier, *core_clust* is topological core clustering coefficient, whereas *core_wclust*$_O$(*top*) corresponds to *wclust*$_O$ applied to topological network, when no semantic similarity thresholds were considered. The best result is obtained with *core_wclust*$_O$ (Fig. 4). Also *core_wclust*$_O$(*top*) is considered to perform better then *core_clust*, in spite of the fact that this topological weighting schema results in larger number of modules. With increasing overlap score threshold (O = 0.2), fewer modules pass the thresholds compared to the other schema, meaning that biological significance of those modules may be lower.

B. Module Ranking with Density Score

Further evaluation of the obtained modules, was focused on choosing best parameter setting by using *core_wclust*$_O$, which is the one that resulted in the largest number of modules that match MIPS complexes. The parameter setting that gave best result is when we did not perform filtering in prior to fluffing, and fluff parameter f was higher than 0, meaning that all neighbours of the original modules have been added to modules, no matter if they belong to several modules. *PWD* was set to 0.95. This parameter setting resulted in 521 modules with connectivity $k \geq 2$.

The obtained modules are ranked according to the density score. Given a module graph $G = (V, E)$, where the number of proteins is denoted by n and the number of interactions is denoted by m, the density is defined as m divided by the theoretical maximum number of edges possible for the module graph, m_{max},[8] defined as $n(n-1)/2$.

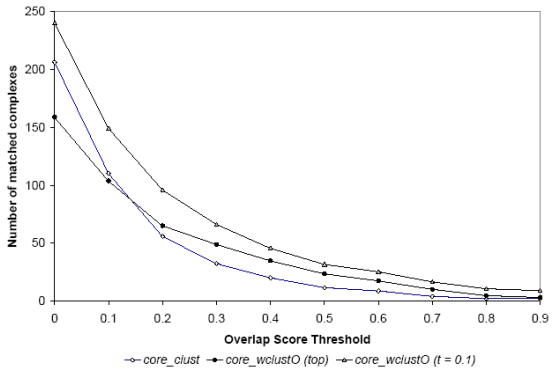

Fig. 4 Comparison between topological weighting scheme, *core_clust*, its weighted analogue *core_wclust*$_O$(*top*) and *core_wclust*$_O$ when semantic similarity threshold t was set to 0.1

Table I shows a list of the 10 top-ranked modules. The score in column five corresponds to the density of the module multiplied with its number of members. In this way, larger and more densely connected modules are given higher scores.

The functional module with highest rank corresponds to the Lsm complex. All eight Lsm-proteins are correctly predicted by the algorithm. Sm-like (Lsm) proteins participate in a variety of RNA processing events. For example, Lsm1-Lsm7 are involved in mRNA degradation and splicing.[25] Besides Lsm-proteins, this module contain Pat1, which is a decapping activator that interacts with Lsm1-Lsm7. In this way, the Lsm-proteins may promote mRNA decapping, which is necessary for mRNA degradation.[26] Other examples of identified complexes are Oligosaccharyl Transferase Complex[27] (rank 6) and pore complex (rank 2).

V. Discussion and Conclusions

We have proposed a method for analysis of protein networks using a measure based on a novel combination of topological and functional information of the proteins. The algorithm takes advantage of this integrated measure to identify locally dense regions with high functional similarity. In the evaluation of the method, we found modules with high functional homogeneity, in many cases corresponding to sets of proteins that constitute known molecular complexes and some additional interacting proteins which share high functional similarity with the complex but are not part of it. Together, such sets of interacting proteins form functional modules that control or perform particular cellular functions, without necessarily forming a macromolecular complex. Thus, the method may be used for the prediction of unknown proteins which participate in the identified modules. We have demonstrated that adding additional knowledge by considering semantic similarity between the proteins, even at low similarity generates modules that are more biologically plausible that those generated solely based on topological information.

It is also important to mention the MIPS database that we used in our evaluation covers complexes rather than modules,

TABLE I
MODULES WITH TOP 10 RANKS

Rank	# proteins	Protein Names	Density	[a]Score	[b]Cellular component
1	9	Lsm3, Lsm2, Lsm8, Lsm5, Lsm6, Lsm7, Lsm4, Lsm1, Pat1	1.00	9.00	Ribonuclopprotein complex
2	15	Kap95, Nup145, Srp1, Gsp1, Nup100, Nup116, Nup1, Nup57, Nup42, Nup49, Nup60, Nup2, Pse1, Crm1, Msn5	0.59	8.86	Pore complex
3	9	Bet1, Ret2, Bos1, Cop1, Sec21, Sec22, Sec26, Sec27, , Ret3	0.92	8.25	COPI vesicle coat
4	12	Rpn11, Rpn12, Rpt3, Ecm29, Rpt2, Rpt6, Rpt1, Pre1, Rad23, Pre5, Rpt2, Rpt5	0.67	8.00	Proteasome complex
5	9	Taf9, Gcn4, Ngg1, Taf1, Ada2, Taf5, Taf6, Spt7, Taf10	0.86	7.75	SLIK complex
6	8	Ost5, Ost2, Ost3, Ost1, Ost4, Stt3, Swp1, Wbp1	0.96	7.71	Oligosaccharyl transferase complex
7	17	Mak21, Mak5, Rlp7, Nop4, Has1, Nop7, Cic1, Nop15, Rrp12, Ytm1, Erb1, Nog1, Nop2, Puf6, Tif6, Nsa2, Sda1	0.45	7.63	Nucleolus
8	9	Caf130, Not3, Ccr4, Caf40, Cdc39, Mot2, Not5, Pop2, Taf1	0.83	7.50	CCR4-NOT complex
9	9	Mpe1, Cft2, Pap1, Pta1, Ref2, Pfs2, Pti1, Pcf11, Rna14	0.81	7.25	mRNA cleavage factor complex
10	17	Rpn10, Rpt3, Rad23, Ecm29, Pre1, Pre2, Pre4, Pre5, Pre6, Pre8, Pre9, Pup3, Rpt2, Rpt4, Rpt6, Scl1, Rpt1	0.43	7.25	Proteasome complex

[a] Density multiplied with number of members of the module, [b] Most significantelly shared GO term from sub-ontology describing cellular component.

and it may therefore only partially describe what we expect to see in the results from a module prediction method. A module may include more than just a complex. MIPS is currently the best source available, but it can not be considered a benchmark in its current form. We hope that future applications of this work will contribute to developing a benchmark which can be used for a more thorough evaluation of prediction accuracy. Future work will also include investigating other weighting functions, for example based on the GO cellular component annotation. We will also compare our method more systematically with other methods for sub-graph identification.

REFERENCES

[1] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, pp. C47-52, 1999.
[2] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proc Natl Acad Sci U S A*, vol. 100, pp. 12123-8, 2003.
[3] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440-2, 1998.
[4] Z. Lubovac, J. Gamalielsson, and B. Olsson, "Combining functional and topological properties to identify core modules in protein interaction networks," *Proteins*, 2006.
[5] A. W. Rives and T. Galitski, "Modular organization of cellular networks," *Proc Natl Acad Sci U S A*, vol. 100, pp. 1128-33, 2003.
[6] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis, "Detection of functional modules from protein interaction networks," *Proteins*, vol. 54, pp. 49-57, 2004.
[7] J. F. Poyatos and L. D. Hurst, "How biologically relevant are interaction-based modules in protein networks?," *Genome Biol*, vol. 5, pp. R93, 2004.
[8] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, pp. 2, 2003.
[9] N. Speer, H. Fröhlich, C. Spieth, and A. Zell, "Functional Grouping of Genes Using Spectral Clustering and Gene Ontology," presented at IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2005), San Diego, USA, 2005.
[10] J.-P. Onella, J. Saramäki, J. Kertész, and K. Kaski, "Intensity and coherence of motifs in weighted complex networks," *Physical Reviews E*, vol. 71, pp. 065103, 2005.
[11] A. H. Tong, B. Drees, G. Nardelli, G. D. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. W. Hogue, S. Fields, C. Boone, and G. Cesareni, "A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules," *Science*, vol. 295, pp. 321-4, 2002.
[12] S. Wuchty and E. Almaas, "Peeling the yeast protein network," *Proteomics*, vol. 5, pp. 444-9, 2005.
[13] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "DIP: the database of interacting proteins," *Nucleic Acids Res*, vol. 28, pp. 289-91, 2000.
[14] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proc Natl Acad Sci U S A*, vol. 98, pp. 4569-74, 2001.
[15] "Creating the gene ontology resource: design and implementation," *Genome Res*, vol. 11, pp. 1425-33, 2001.
[16] M. Deng, Z. Tu, F. Sun, and T. Chen, "Mapping Gene Ontology to proteins based on protein-protein interaction data," *Bioinformatics*, vol. 20, pp. 895-902, 2004.
[17] U. Karaoz, T. M. Murali, S. Letovsky, Y. Zheng, C. Ding, C. R. Cantor, and S. Kasif, "Whole-genome annotation by using evidence integration in functional-linkage networks," *Proc Natl Acad Sci U S A*, vol. 101, pp. 2888-93, 2004.
[18] D. Lin, "An information-theoretic definition of similarity," presented at The 15th International Conference on Mashine Learning, Madison, WI, 1998.
[19] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlock, A. Sethuraman, S. Weng, D. Botstein, and J. M. Cherry, "Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO)," *Nucleic Acids Res*, vol. 30, pp. 69-72, 2002.
[20] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," presented at International Conference on Research in Computational Linguistics, Taiwan, 1998.
[21] P. Resnik, "Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language," *Journal of Artificial Intelligence Research*, vol. 11, pp. 95-130, 1999.
[22] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble, "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation," *Bioinformatics*, vol. 19, pp. 1275-83, 2003.
[23] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proc Natl Acad Sci U S A*, vol. 101, pp. 3747-52, 2004.
[24] E. Bouveret, G. Rigaut, A. Shevchenko, M. Wilm, and B. Seraphin, "A Sm-like protein complex that participates in mRNA degradation," *Embo J*, vol. 19, pp. 1661-71, 2000.
[25] W. He and R. Parker, "Functions of Lsm proteins in mRNA degradation and splicing," *Curr Opin Cell Biol*, vol. 12, pp. 346-50, 2000.
[26] S. Tharun, W. He, A. E. Mayes, P. Lennertz, J. D. Beggs, and R. Parker, "Yeast Sm-like proteins function in mRNA decapping and decay," *Nature*, vol. 404, pp. 515-8, 2000.
[27] R. Knauer and L. Lehle, "The oligosaccharyltransferase complex from Saccharomyces cerevisiae. Isolation of the OST6 gene, its synthetic interaction with OST3, and analysis of the native complex," *J Biol Chem*, vol. 274, pp. 17249-56, 1999.