Assessing Significance of Connectivity and Conservation in Protein Interaction Networks

Mehmet Koyutürk, Wojciech Szpankowski, and Ananth Grama

Department of Computer Sciences, Purdue University, West Lafayette, IN 47907.

Corresponding Author:

Mehmet Koyutürk

Lawson Computer Science Building

West Lafayette, IN 47907, USA.

e-mail: koyuturk@cs.purdue.edu

Fax: +1 765-494-0739

Abstract

Comparative analyses of cellular interaction networks enable understanding of the cell's modular organization through identification of functional modules and complexes. These techniques often rely on topological features such as connectedness and density, based on the premise that functionally related proteins are likely to interact densely and that these interactions follow similar evolutionary trajectories. Significant recent work has focused on efficient algorithms for identification of such functional modules and their conservation. In spite of algorithmic advances, development of a comprehensive infrastructure for interaction databases is in relative infancy compared to corresponding sequence analysis tools. One critical, and as yet unresolved aspect of this infrastructure is a measure of the statistical significance of a match, or a dense subcomponent. In the absence of analytical measures, conventional methods rely on computationally expensive simulations based on ad-hoc models for quantifying significance. In this paper, we present techniques for analytically quantifying statistical significance of dense components in reference model graphs. We consider two reference models – a $G(n, p)$ model in which each pair of nodes in a graph has an identical likelihood, p, of sharing an edge, and a two-level $G(n, p)$ model, which accounts for high-degree hub nodes generally observed in interaction networks. Experiments performed on a rich collection of protein interaction (PPI) networks show that the proposed model provides a reliable means of evaluating statistical significance of dense patterns in these networks. We also adapt existing state-of-the-art network clustering algorithms by using our statistical significance measure as an optimization criterion. Comparison of the resulting module identification algorithm, SIDES, with existing methods shows that SIDES outperforms existing algorithms in terms of sensitivity and specificity of identified clusters with respect to available GO annotations.

1 Introduction

Effective analysis of the interactome holds the key to functional characterization, phenotypic mapping, and identification of pharmacological targets, among other important tasks [3, 42]. Computational infrastructure for supporting analysis of the interactome is in relative infancy, compared to its sequence counterparts [40]. A large body of work on computational analysis of these graphs has focused on identification of dense components (proteins that densely interact with each other) [4, 7, 22, 23, 26, 31]. These methods are based on the premise that functionally related proteins generally manifest themselves as dense components in the network [36]. The hypothesis that proteins performing a particular cellular function together are expected to be conserved across several species along with their interactions, is also used to guide the process of identifying conserved networks across species. Based on this observation, PPI network alignment methods superpose PPI networks that belong to different species and search for connected, dense, or heavy subgraphs in these superposed graphs [14, 18, 20, 19, 29, 30].

There are two critical aspects of identifying meaningful structures in data – the algorithm for the identification and a method for scoring an identified pattern. In this context, the score of a pattern corresponds to its significance. A score is generally computed with respect to a reference model – i.e., given a pattern and a reference model, how likely is it to observe the pattern in the reference model. The less likely such an occurrence is in the reference model, the more interesting it is, since it represents a significant deviation from the reference (nominal) behavior. One such score, in the context of sequences is the E-value returned by BLAST matches [41]. This score broadly corresponds to the likelihood that a match between two sequences is generated by a random process. The lower this value, the more meaningful the match. It is very common in a variety

of applications to use a threshold on E-values to identify homologies across sequences. It is reasonable to credit E-value as one of the key ingredients of the success of sequence matching algorithms and software.

While significant progress has been made towards developing algorithms on graphs for identifying patterns (motifs, dense components), conservation, alignment, and related problems, analytical methods for quantifying the significance of such patterns are limited. Existing algorithms for detecting general patterns typically adopt simple ad-hoc measures (such as frequency or relative density) [4, 18], compute z-scores for *the* observed pattern based on simplifying assumptions [20, 29, 30], or rely on Monte-Carlo simulations [29] to assess the significance of identified patterns. Itzkovitz et al. [15] analyze the expected number of occurrences of specific *topological motifs* in a variety of random networks. This paper represents the first effort at analytically quantifying the statistical significance of the *existence* of *a* pattern with observed property, with respect to a reference model. Specifically, it presents a framework for analyzing the occurrence of dense patterns in randomly generated graph-structured data (based on the underlying model) with a view to assessing the significance of a pattern based on the statistical relationship between subgraph density and size. This result generalized in a straightforward manner to the problem of assessing statistical significance of matches between two interaction networks.

The selection of an appropriate reference model for data and the method of scoring a pattern or match, are important aspects of quantifying statistical significance. Using a reference model that fits the data very closely makes it more likely that an experimentally observed biologically significant pattern is generated by a random process drawing data from this model. Conversely, a reference model that is sufficiently distinct from observed data is likely to tag most patterns as being significant. Clearly, neither extreme is desirable for good coverage and accuracy. In this

paper, we consider two reference models (i) a $G(n,p)$ model of a graph with n nodes, where each pair of nodes has an identical probability, p, of sharing an edge, and (ii) a two level $G(n,p)$ model in which the graph is modeled as two separate $G(n,p)$ graphs with intervening edges. The latter model captures the heavy nodes corresponding to hub proteins, typically observed in PPIs. For these models, we analytically quantify the behavior of the largest dense subgraph and use this to derive a measure of significance. We show that a simple $G(n,p)$ model can be used to assess the significance of dense patterns in graphs with arbitrary degree distribution, with a conservative adjustment of parameters so that the model stochastically dominates a graph generated according to a given distribution. In particular, by choosing p to be maximal, we ensure that the largest dense subgraph in our $G(n,p)$ model stochastically dominates that of a power-law graph. Our two-level $G(n,p)$ model is devised to mirror key properties of the underlying topology of PPI graphs, and consequently yields a more conservative estimate of significance. Finally, we show how existing graph clustering algorithms [13] can be modified to incorporate statistical significance in identification of dense patterns, resulting in an effective module identification algorithm, SIDES. (SIDES is available as a standalone application and as a plugin to Cytoscape over the public domain from our lab.) We also generalize our results and methods to the comparative analysis of PPI networks and show how the significance of a match between two networks can be quantified in terms of the significance of the corresponding dense component in a suitable specified product graph.

Our analytical results are supported by extensive experimental results on a large collection of PPI networks derived from BIND [3] and DIP [42]. These results demonstrate that the proposed model and subsequent analysis provide reliable means for evaluating the statistical significance of highly connected and conserved patterns in PPI networks. We also compare the resulting al-

gorithmic technique, SIDES, with the module identification algorithm, MCODE [4] and show that SIDES outperforms this algorithm in terms of specificity and sensitivity of identified clusters with respect to GO annotations. The framework proposed here can be extended to include more general networks that capture the degree distribution of PPI networks more accurately, namely power-law [38, 43], geometric [24], or exponential [9] degree distributions.

The rest of this manuscript is organized as follows: In the next section, we discuss graph models for PPI networks. We then analyze the behavior of the largest dense subgraph and derive measures for assessing statistical significance of highly connected as well as highly conserved subgraphs in PPI networks. In Section 3, we introduce the SIDES algorithm. We present and discuss experimental results in Section 4 and conclude our discussion in Section 5.

2 Probabilistic Analysis of Dense Subgraphs

Since proteins that are part of a functional module are likely to densely interact with each other, while being somewhat isolated from the rest of the network [36], many commonly used methods focus on discovering dense regions of the network for identification of functional modules or protein complexes [4, 7, 22, 26, 31]. Subgraph density is also central to many algorithms that target identification of conserved modules and complexes [14, 20, 29]. In order to assess the statistical significance of such dense patterns, we analyze the distribution of the largest "dense" subgraph generated by an underlying reference model. Using this distribution, we estimate the probability that an experimentally observed pattern will occur in the network by chance. The reference model must mirror the basic characteristics of experimentally observed networks in order to capture the underlying biological process correctly, while being simple enough to facilitate theoretical and

computational analysis.

2.1 Modeling PPI Networks

With the increasing availability of high-throughput interaction data, there has been significant effort aimed at modeling PPI networks. The key observation on these networks is that a few central proteins interact with many proteins, while most proteins in the network have few interacting partners [16, 25]. A commonly accepted model that confirms this observation is based on power-law degree distribution [5, 37, 38, 43]. In this model, the number of nodes in the network that have d neighbors is proportional to $d^{-\gamma}$, where γ is a network-specific parameter. It has also been shown that there exist networks that do not possess a power-law degree distribution [12, 35]. In this respect, alternative models that are based on geometric [24] or exponential [9] degree distribution have been also proposed.

While assessing the statistical significance of identified patterns, existing methods that target identification of highly connected or conserved patterns in PPI networks generally rely on the assumption that interactions in the network are independent of each other [17, 20, 29]. Since degree distribution is critical to the generation of interesting patterns, these methods estimate the probability of each interaction based on the degree distribution of the underlying network. These probabilities can be estimated computationally by generating several random graphs with the same degree distribution via repeated edge swaps and counting the occurrence of each edge in this large collection of random graphs [29]. Alternately, they can be estimated analytically, by relying on a simple random graph model that is based on a given degree distribution [8, 15]. In this model, each node $u \in V(G)$ of graph $G = (V, E)$ is associated with expected degree d_u and the probability

of existence of an edge between u and v is defined as $P(uv \in E(G)) = d_u d_v / \sum_{u \in V(G)} d(u)$. In order for this function to be a well-defined probability measure for simple graphs, we must have $d_{\max}^2 \leq \sum_{u \in V(G)} d(u)$, where $d_{\max} = \max_{u \in V(G)} d_u$. However, available protein interaction data generally does not conform to this assumption. For example, based on the PPI networks we derive from BIND [3] and DIP [42] databases, yeast *Jsn1* protein has 298 interacting partners, while the total number of interactions in the *S. cerevisiae* PPI network is 18193. Similarly, the *D. Melanogaster* PPI network with 28830 interactions contains a protein (CG12470-PA ORF) with 207 interacting partners. Such problems complicate the analysis of the significance of certain structures for models that are based on arbitrary degree distribution.

While models that assume power-law [38, 43], geometric [24], or exponential [9] degree distributions may capture the topological characteristics of PPI networks accurately, they require more involved analysis and may also require extensive computation for assessment of significance. To the best of our knowledge, the distribution of dense subgraphs, even maximum clique, which forms a special case of this problem, has not been studied for power-law graphs. In this paper, we first build a framework for the simple and well-studied $G(n, p)$ model and attempt to generalize our results to more complicated models that assume heterogeneous degree distribution.

2.2 Largest dense subgraph

Given graph G, let $F(U) \subseteq E(G)$ be the set of edges in the subgraph induced by node subset $U \subseteq V(G)$. The density of this subgraph is defined as $\delta(U) = |F(U)|/|U|^2$. Note here that we assume directed edges and allow self-loops for simplicity. PPI networks are undirected graphs and they contain self-loops in general, but any undirected network can be easily modeled by a directed

graph and this does not impact the asymptotic correctness of the results. We define a ρ-dense subgraph to be one with density *larger* than pre-defined threshold ρ, *i.e.,* U induces a ρ-dense subgraph if $F(U) \geq \rho|U|^2$. For any ρ, we are interested in the number of nodes in the largest ρ-dense subgraph. This is because any ρ-dense subgraph in the observed PPI network with size larger than this value will be "unusual", *i.e.*, statistically significant. Note that maximum clique is a special case of this problem with $\rho = 1$.

We first analyze the behavior of the largest dense subgraph for the $G(n, p)$ model of random graphs. We subsequently generalize these results to the piecewise degree distribution model in which there are two different probabilities of generating edges. In the $G(n, p)$ model, a graph G contains n nodes and each edge occurs independently with probability p.

Let random variable $R_n(\rho)$ be the size of the maximum subset of vertices that induce a ρ-dense subgraph, *i.e.,*

$$R_n(\rho) = \max_{U \subseteq V(G): \delta(U) \geq \rho} |U|. \tag{1}$$

The behavior of $R_n(1)$, which corresponds to maximum clique, is well studied for the $G(n, p)$ model and its typical value is shown to be $O(\log_{1/p} n)$ [6]. In the following theorem, we derive a general result for the typical value of $R_n(\rho)$ for any $\rho > p$.

Theorem 1 *If G is a random graph with n vertices, where every edge exists with probability p and $\rho > p$, then*

$$\lim_{n \to \infty} \frac{R_n(\rho)}{\log n} = \frac{1}{\kappa(p, \rho)} \qquad (pr.), \tag{2}$$

where

$$\kappa(p, \rho) = -H_p(\rho) = \rho \log \frac{\rho}{p} + (1 - \rho) \log \frac{1 - \rho}{1 - p}. \tag{3}$$

Here, $H_p(\rho)$ denotes weighted entropy. More precisely,

$$P(R_n(\rho) \geq r_0) \leq O\left(\frac{\log n}{n^{1/\kappa(p,\rho)}}\right), \tag{4}$$

where

$$r_0 = \frac{\log n - \log\log n + \log \kappa(p,\rho) - \log e + 1}{\kappa(p,\rho)} \tag{5}$$

for large n.

Proof. We first prove the upper-bound. Let $X_{r,\rho}$ denote the number of subgraphs of size r with density at least ρ, *i.e.*, $X_{r,\rho} = |\{U \subseteq V(G) : |U| = r \wedge |F(U)| \geq \rho r^2\}|$. From first moment method, we obtain $P(R_n(\rho) \geq r) \leq P(X_{r,\rho} \geq 1) \leq \mathbf{E}[X_{r,\rho}]$.

Let Y_r denote the number of edges induced by r vertices. Then, $\mathbf{E}[X_r] = \binom{n}{r} P(Y_r \geq \rho r^2)$. Moreover, since Y_r is a Binomial r.v. $B(r^2, p)$ and $\rho > p$, we have

$$P(Y_r \geq \rho r^2) \leq (r^2 - \rho r^2) P(Y_r = \rho r^2) \leq \binom{r^2}{\rho r^2} (r^2 - \rho r^2) p^{\rho r^2} (1-p)^{r^2 - \rho r^2}. \tag{6}$$

Hence, we get $P(R_n(\rho) \geq r) \leq \binom{n}{r}\binom{r^2}{\rho r^2}(r^2 - \rho r^2) p^{\rho r^2}(1-p)^{r^2-\rho r^2}$.

Using Stirling's formula, we find the following asymptotics for $\binom{n}{r}$:

$$\binom{n}{r} \sim \begin{cases} \frac{1}{\sqrt{2\pi r}} \frac{n^r}{r^r} e^{-r} & \text{if } r = o(\sqrt{n}) \\ \frac{1}{\sqrt{2\pi\alpha(1-\alpha)n}} 2^{nH(\alpha)} & \text{if } r = \alpha n \end{cases} \tag{7}$$

where $H(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$ denotes the binary entropy.

Let $Q = 1/p^\rho (1-p)^{1-\rho}$. Plugging the above asymptotics into (2.2), we obtain

$$P(R_n(\rho) \geq r) \leq \frac{r\sqrt{1-\rho}}{2\pi\sqrt{\rho}} \exp_2(-r^2 \log Q + r \log n - r \log r + r^2 H(\rho) - r \log e) \tag{8}$$

Defining $\kappa(p,\rho) = \log Q - H(\rho)$, we find $P(R_n(\rho) \geq r_0) \leq \frac{r_0\sqrt{1-\rho}}{2\pi\sqrt{\rho}} \exp_2(f(r_0))$, where $f(r_0) = -r_0(r_0\kappa(p,\rho) - \log n + \log r + \log e)$. Plugging in (5) and working out the algebra, we

obtain $f(r_0) = -r_0\left(1 - O\left(\frac{\log\log n}{\log n}\right)\right)$. Hence, $P(R_n(\rho) \geq r_0) \leq O\left(2^{-r_0}\right) = O\left(\frac{\log n}{n^{1/\kappa(p,\rho)}}\right)$. This completes the proof for the upper-bound.

For the lower bound, we have

$$P(R_n(\rho) < r) = P(X_{r,\rho} = 0) \leq \frac{\mathbf{E}[X_{r,\rho}^2]}{\mathbf{E}[X_{r,\rho}]^2}. \tag{9}$$

from second moment method [33]. Letting $m = \rho r^2$, we obtain $\mathbf{E}[X_{r,\rho}] = \binom{n}{r}\binom{r}{m}p^m q^{r^2-m}$and

$$\mathbf{E}[X_{r,\rho}^2] = \binom{n}{r}\sum_{l=0}^{r}\binom{r}{l}\binom{n-r}{r-l}\sum_{k\in I_l}\binom{l^2}{k}p^k(1-p)^{l^2-k}\left[\binom{r^2-l^2}{m-k}p^{m-k}(1-p)^{r^2-l^2-(m-k)}\right]^2 \tag{10}$$

where $I_l = \{k : \max(0, l^2 + m - r^2) \leq k \leq \min(l^2, m)\}$. Here, for two node subsets U_r and V_r, l denotes the number of nodes at the intersection of U_r and V_r, *i.e.,* $l = |U_r \cap V_r|$. On the other hand, k denotes the number of edges at the intersection of the subgraphs induced by U_r and V_r, *i.e.,* $k = |F(U_r) \cap F(V_r)|$. Hence,

$$\frac{\mathbf{E}[X_{r,\rho}^2]}{\mathbf{E}[X_{r,\rho}]^2} = \sum_{l=0}^{r}\sum_{k\in I_l} f(r,l,k) \tag{11}$$

where

$$f(r,l,k) = \frac{\binom{n-r}{r-l}\binom{r}{l}\binom{l^2}{k}\binom{r^2-l^2}{m-k}^2 p^{-k}(1-p)^{k-l^2}}{\binom{n}{r}\binom{r^2}{m}^2}. \tag{12}$$

Therefore,

$$P(R_n(\rho) < r) \leq \sum_{l=0}^{r}\sum_{k\in I_l} f(r,l,k) \leq r^3 \max_{l,k} f(r,l,k). \tag{13}$$

For $r = \frac{(1-\epsilon)\log n}{\kappa(\rho,p)}$, $0 \leq l \leq r$ and $k \in I_l$, we will show that

$$f\left(\frac{(1-\epsilon)\log n}{\kappa(\rho,p)}, l, k\right) \leq n^{\frac{-\epsilon(1-\epsilon)\log n}{\kappa(\rho,p)}} \tag{14}$$

to conclude that $P(R_n(\rho) < r) \leq \frac{(\log n)^3}{n^{\frac{\epsilon(1-\epsilon)\log n}{\kappa(\rho,p)}}}$. To achieve this, let $\alpha = l^2/r^2$ and $\beta = k/l^2$. Then, assuming $\rho > 1/2$ without loss of generality, the interval corresponding to I_l for $0 \leq \alpha \leq 1$

becomes

$$J_\alpha = \left\{ \begin{array}{ll} 0 \le \beta \le 1 & \text{if } 0 \le \alpha \le 1-\rho \\ \beta: \ \frac{\alpha+\rho-1}{\alpha} \le \beta \le 1 & \text{if } 1-\rho \le \alpha \le \rho \\ \frac{\alpha+\rho-1}{\alpha} \le \beta \le \frac{\rho}{\alpha} & \text{if } \rho \le \alpha \le 1. \end{array} \right\} \tag{15}$$

Inserting $l = \sqrt{\alpha} r$ and $k = \alpha\beta r^2$ in (12), we obtain

$$f_{\alpha,\beta}(r) = \frac{\binom{r}{\sqrt{\alpha}r}\binom{n-r}{(1-\sqrt{\alpha})r}\binom{\alpha r^2}{\alpha\beta r^2}\binom{(1-\alpha)r^2}{(\rho-\alpha\beta)r^2}^2 p^{-\alpha\beta r^2}(1-p)^{\alpha(\beta-1)r^2}}{\binom{n}{r}\binom{r^2}{\rho r^2}^2}. \tag{16}$$

Plugging Stirling's approximation (7) for appropriate regimes, we get

$$\begin{aligned} \log(f_{\alpha,\beta}(r)) \sim \ & -r(\sqrt{\alpha}\log n) \\ & + r^2(\alpha H(\beta) - \alpha(\beta \log p + (1-\beta)\log(1-p)) + 2(1-\alpha)H\left(\tfrac{\rho-\alpha\beta}{1-\alpha}\right) - 2H(\rho)). \end{aligned} \tag{17}$$

Hence, for $r = \frac{(1-\epsilon)\log n}{\kappa(p,\rho)}$, we have

$$\log\left(f_{\alpha,\beta}\left(\frac{(1-\epsilon)\log n}{\kappa(p,\rho)}\right)\right) \sim \frac{1-\epsilon}{\kappa(p,\rho)}(\log n)^2\left[-\sqrt{\alpha} + \frac{1-\epsilon}{\kappa(p,\rho)} g(\alpha,\beta)\right] \tag{18}$$

where

$$g(\alpha,\beta) = \alpha H(\beta) - \alpha(\beta\log p + (1-\beta)\log(1-p)) + 2(1-\alpha)H\left(\frac{\rho-\alpha\beta}{1-\alpha}\right) - 2H(\rho). \tag{19}$$

Working out the algebra, we observe that

$$\max_{0\le\alpha\le 1, \beta\in J_\alpha} g(\alpha,\beta) = g(1,\rho) = \kappa(\rho,p) \tag{20}$$

where the maximum corresponds to the boundary point $l = r$ and $k = \rho r^2$. Hence, it immediately follows from (18) that $\log(f) \le \frac{-\epsilon(1-\epsilon)}{\kappa(p,\rho)}(\log n)^2$ for $0 \le \alpha \le 1$ and $\beta \in J_\alpha$, which leads to (14).

□

Observe that, if n is large enough, the probability that a dense subgraph of size r_0 exists in the subgraph is very small. Consequently, r_0 may provide a threshold for deciding whether an observed dense pattern is statistically significant.

For a graph of arbitrary degree distribution, let $d_{\max}$ denote the maximum expected degree as defined in Section 2.1. Estimating the probability of observing an edge between any two nodes in the $G(n,p)$ model by $p = d_{\max}/n$, it is possible to conservatively assess the significance of a dense subgraph using the above results. The above result also provides a means for quantifying the significance of an observed dense subgraph. For a subgraph with size $\hat{r} > r_0$ and density $\hat{\rho}$, let $\epsilon = \frac{\hat{r} - \log n/\kappa(\hat{\rho},p)}{\log n/\kappa(\hat{\rho},p)}$. Then, it follows from (8) that the probability of observing this subgraph in a graph generated according to the reference model is bounded by

$$P(R_n(\hat{\rho}) \geq (1+\epsilon)\log n/\kappa(\hat{\rho},p)) \leq \frac{\sqrt{1-\rho}}{2\pi\sqrt{\rho}} \frac{(1+\epsilon)\log n}{n^{\epsilon(1+\epsilon)\log n/\kappa(\hat{\rho},p)}}. \tag{21}$$

While these results for the $G(n,p)$ model provide a simple yet effective way of assessing statistical significance of dense subgraphs, we extend our analysis to a more complicated model, which takes into account the degree distribution to capture the topology of the PPI networks more accurately.

2.3 Piecewise degree distribution model

In the piecewise degree distribution model, nodes of the graph are divided into two classes, namely high-degree and low-degree nodes. More precisely, we define random graph G with node set $V(G)$ that is composed of two disjoint subsets $V_h \subset V(G)$ and $V_l = V(G) \setminus V_h$, where $n_h = |V_h| \ll |V_l| = n_l$ and $n_h + n_l = n = |V(G)|$. In the reference graph, the probability of an edge is defined based on the classes of its incident nodes as:

$$P(uv \in E(G)) = \begin{cases} p_h & \text{if } u,v \in V_h \\ p_l & \text{if } u,v \in V_l \\ p_b & \text{if } u \in V_h, v \in V_l \text{ or } u \in V_l, v \in V_h \end{cases} \tag{22}$$

Here, $p_l < p_b < p_h$. This model captures the key lethality and centrality properties of PPI networks in the sense that a few nodes are highly connected while most nodes in the network have low degree [16, 25]. Observe that, under this model, G can be viewed as a superposition of three random graphs G_l, G_h, and G_b. Here, G_h and G_l are $G(n,p)$ graphs with parameters (n_h, p_h) and (n_l, p_l), respectively. G_b, on the other hand, is a random bipartite graph with node sets V_l, V_h, where each edge occurs with probability p_b. Hence, we have $E(G) = E(G_l) \cup E(G_h) \cup E(G_b)$. This facilitates direct employment of the results in the previous section for analyzing graphs with piecewise degree distribution.

We now show that the high-degree nodes in the piecewise degree distribution model contribute a constant factor to the typical size of the largest dense subgraph as long as n_h is bounded by a constant.

Theorem 2 *Let G be a random graph with piecewise degree distribution, as defined by (22). If $n_h = O(1)$, then*

$$P(R_n(\rho) \geq r_1) \leq O\left(\frac{\log n}{n^{1/\kappa(p_l,\rho)}}\right), \tag{23}$$

where

$$r_1 = \frac{\log n - \log\log n + 2n_h \log B + \log \kappa(p_l,\rho) - \log e + 1}{\kappa(p_l,\rho)} \tag{24}$$

and $B = \frac{p_b q_l}{p_l} + q_b$, where $q_b = 1 - p_b$ and $q_l = 1 - p_l$.

Proof. Let $X^h_{r,\rho}$, $X^l_{r,\rho}$ be the number of ρ-dense subgraphs induced by only nodes in G_h or G_l, respectively. Let $X^b_{r,\rho}$ be the number of these induced by nodes from both sets. Clearly, $X_{r,\rho} = X^h_{r,\rho} + X^l_{r,\rho} + X^b_{r,\rho}$. The analysis for $G(n,p)$ directly applies for $\mathbf{E}[X^h_{r,\rho}]$ and $\mathbf{E}[X^l_{r,\rho}]$, hence we emphasize on $\mathbf{E}[X^b_{r,\rho}]$. Since $n_h = O(1)$, we have $\mathbf{E}[X^b_{r,\rho}] \leq (1-\rho) r^2 \sum_{k=0}^{n_h} \binom{n_h}{k} \binom{n_l}{r-k} \sum_{l=0}^{2k(r-k)} \binom{2k(r-k)}{l} \binom{(r-k)^2}{\rho r^2 - l} p_b^l q_b^{2k(r-k)-l} p_l^{\rho r^2 - l}\ q_l^{(r-k)^2 - \rho r^2 + l}$,where $q_b = 1 - p_b$

and $q_l = 1 - p_l$. Then,

$$\mathbf{E}[X^b_{r,\rho}] \leq c(1-\rho)r^2 n_h \binom{n_l}{r} \sum_{l=0}^{2n_h r} \binom{2n_h r}{l} \binom{r^2}{\rho r^2 - l} p_b^l q_b^{2n_h r - l} p_l^{\rho r^2 - l} q_l^{r^2 - \rho r^2 + l}, \tag{25}$$

where c is a constant. Since $l = o(\rho r^2)$, we have $\binom{r^2}{\rho r^2 - l} \leq \binom{r^2}{\rho r^2}$ for $0 \leq l \leq 2n_h r$. Therefore,

$$\mathbf{E}[X^b_{r,\rho}] \leq (1-\rho)r^2 \binom{n}{r} \binom{r^2}{\rho r^2} p_l^{\rho r^2} q_l^{r^2 - \rho r^2} \sum_{l=0}^{2n_h r} \binom{2n_h r}{l} \left(\frac{p_b q_l}{p_l}\right)^l q_b^{2n_h r - l}. \tag{26}$$

Using $B = \frac{p_b q_l}{p_l} + q_b$ as defined in Theorem 2, we find $P(R_n(\rho) > r) \leq O(2^{f_1(r)})$, where $f_1(r) = -r(r\kappa(\rho) - \log n + \log r - \log e + 2n_h \log B)$.Hence, $P(R_n(\rho) > r_1) \leq O(2^{f_1(r_1)}) \leq O\left(\frac{\log n}{n^{1/\kappa(p_l,\rho)}}\right)$ for large n. □

Note that the above result is based on asymptotic behavior of r_1, hence the $\log n$ term dominates as $n \to \infty$. However, if n is not large enough, the $2n_h \log B$ term may cause over-estimation of the critical value of the largest dense subgraph. Therefore, the application of this theorem is limited for smaller n and the choice of n_h is critical.

A heuristic approach for estimating n_h is as follows. Assume that the underlying graph is generated by a power-law degree distribution, where the number of nodes with degree d is given by $nd^{-\gamma}/\zeta(\gamma)$ [1]. Here, $\zeta(.)$ denotes the Riemann zeta-function. If we divide the nodes of this graph into two classes where high-degree nodes are those with degree $d \geq (n/\zeta(\gamma))^{1/\gamma}$ so that the expected number of nodes with degree d is at most one, then $n_h = \sum_{d=(n/\zeta(\gamma))^{1/\gamma}}^{\infty} nd^{-\gamma}/\zeta(\gamma)$ is bounded, provided the above series converges.

2.4 Conservation of dense subgraphs

Comparative methods that target identification of conserved subnets in PPI networks induce a cross-product, or superposition, of several networks in which each node corresponds to a group

of orthologous proteins [17, 20, 19, 29, 30]. Here, we rely on ortholog groups available in the COG database [34] to relate proteins in different PPI networks [19]. Labeling each node in the PPI network with the COG family of the protein it represents, we obtain an intersection of two PPI networks by inserting an edge between two COG families only if proteins that belong to these families interact in both graphs. In the case of the $G(n, p)$ model, the above framework directly applies to the identification of dense subgraphs in this intersection graph, where the probability of observing a conserved interaction is estimated as $p_I = p_1 p_2$. Here p_1 and p_2 denote the probability of observing an edge in the first and second networks, respectively. For the piecewise degree distribution model, on the other hand, we have to assume that the orthologs of high-degree nodes in one graph are high-degree nodes in the other graph as well. If this assumption is removed, it can still be shown that the low-degree nodes dominate the typical behavior of the largest conserved subgraph. Note that the reference model assumes that the orthology relationship between proteins in the two networks is already established and the model estimates the conditional probability that the interactions between these given ortholog proteins are densely conserved.

3 SIDES: An Algorithm for the Identification of Significantly Dense Subgraphs

We use the above results to modify an existing state-of-the-art graph clustering algorithm, HCS [13], in order to incorporate statistical significance in identification of interesting dense subgraphs. HCS is a recursive algorithm that is based on decomposing the graph into dense subgraphs by recursive application of min-cut partitioning. A min-cut partition of the nodes of a graph

$G = (V, E)$ is a disjoint partition of V into V_0 and V_1 such that the cut

$$C(V_0, V_1) = |\{uv \in E : u \in V_0, v \in V_1 \vee u \in V_1, v \in V_0\}| \tag{27}$$

is minimized. In the original HCS algorithm, the density of any subgraph found in this recursive decomposition is compared with a pre-defined density threshold. If a subgraph is dense enough, it is reported as a highly-connected cluster of nodes, else it is partitioned again. While this algorithm provides a strong heuristic that is well suited to the identification of densely interacting proteins in PPI networks [23], the selection of density threshold poses an important problem. In other words, it is hard to provide a biologically justifiable answer to the question "How dense must a subnetwork of a PPI network be to be considered biologically interesting?". Our framework provides an answer to this question from a statistical point of view by establishing the relationship between subgraph size and density as a stopping criterion for the algorithm.

For any subgraph encountered during the course of the algorithm, we estimate the critical size of the subgraph to be considered interesting, by plugging in its density in (5) or (24). If the size of the subgraph is larger than this probabilistic upper-bound, we report the subgraph as being statistically significant. Otherwise, we continue partitioning the graph.

An important problem relating to the use of min-cut partitioning is that min-cut partitioning tends to single out a node in one part, since no balance constraint is imposed. Hence, recursive application of min-cut on large graph is likely to result in many clusters containing a single node, which indeed is not significant. This problem is particularly important in PPI networks because of their characteristic degree distribution, *i.e.,* most proteins in the network are low-degree nodes, which are likely to be singled out by min-cut partitioning. We resolve this problem by an additional modification to the HCS algorithm and we partition the network to minimize the ratio cut rather

than the edge cut. Ratio cut partitioning is a well-studied problem in various contexts. It targets minimization of the edge cut while maintaining balance implicitly, without imposing any strict balance constraints [11]. Although being NP-hard, in contrast to the min-cut problem [39], the problem can be solved effectively by heuristic methods and is very well suited for partitioning of PPI networks since no strict balance is required but single-node partitioning needs to be avoided. In our implementation, we define ratio-cut as

$$R(V_0, V_1) = \frac{C(V_0, V_1)}{\min(|V_0|, |V_1|)} \tag{28}$$

and adopt a simple min-cut algorithm [32] to heuristically solve this problem. The underlying algorithm considers $|V|$ partitions, which are locally optimal and chooses the one that induces minimum edge-cut, which is shown to be the global optimum. In our implementation, we consider the same $|V|$ partitions, but choose the one that minimizes the ratio cut of (28) to heuristically favor a more balanced partition.

The resulting *si*gnificant *de*nse *s*ubgraph identification algorithm, SIDES, is shown in Figure 1. Details of the recursive algorithm and the min-cut algorithm can be found in [13] and [32], respectively. Note that this algorithm only identifies disjoint subgraphs, but can be easily extended to obtain overlapping dense subgraphs by greedily growing each of the resulting subgraphs until significance is lost. The C source code and a Java implementation as a Cytoscape [28] plug-in for SIDES are available as open source at `http://www.cs.purdue.edu/homes/koyuturk/sides/`.

4 Results and Discussion

In this section, we first compare the behavior of dense subgraphs in experimentally available network data with the theoretical results presented in this paper. Then, we present experimental results on the performance of SIDES, which uses statistical significance as an optimization criterion, and demonstrate the excellent performance of SIDES in identifying biologically relevant protein clusters as compared to existing algorithms. We do this by quantifying the biological significance of identified clusters in terms of specificity and sensitivity.

4.1 Behavior of Largest Dense Subgraph

We experimentally analyze connectivity and conservation in PPI networks of 11 species gathered from BIND [3] and DIP [42] databases. These networks vary significantly in size and comprehensiveness and cover a broad range of organisms. Relatively large amounts of interaction data is available for *S.cerevisiae* (18192 interactions between 5157 proteins), *D. melanogaster* (28829 among 8577), *H. sapiens* (7393 among 4541), *C. elegans* (5988 among 3345), *E. coli* (1329 among 1079), while the networks for other organisms are restricted to a small portions of their networks. With a view to assessing the impact on performance of merging networks from diverse data sources, we also consider a network in which interactions from these databases are merged to obtain network with binary interactions, i.e., there is an edge between two proteins in the network if these proteins interact in at least one of the databases.

In Figure 2, we examine the behavior of largest subgraph with respect to number of nodes in the PPI network for two different values of density threshold (ρ). Note that, in the context of the experimental results reported in this section, the term *largest dense subgraph* refers to the dense

subgraph of maximal size identified by our algorithm, and does not necessarily correspond to the largest dense subgraph of the underlying graph. In the figure, each organism corresponds to a sample point, which is marked by its name. Since the sparsity and degree distribution of these networks vary significantly across different organisms, the estimated values of edge probabilities vary accordingly. Hence, the curves for r_0 ($G(n,p)$ model) and r_1 (piecewise degree distribution model) do not show a linear behavior. As seen in the figure, piecewise degree distribution model provides a more conservative assessment of significance. This is primarily because of the constant factor in the critical value of r_1. The observed size of the largest dense subgraph in smaller networks is not statistically significant, while larger and more comprehensive networks contain subgraphs that are twice as large as the theoretical estimate, with the exception of the *D. melanogaster* PPI network. The lack of dense subnets in the *D. melanogaster* network may be due to differences in experimental techniques (*e.g.,* two hybrid vs AP/MS) and/or the incorporation of identified interactions in the interaction network model (*e.g.,* spoke vs matrix) [27]. In order to avoid problems associated with such variability, it may be necessary to revise the definition of subgraph density or preprocess the PPI networks to standardize the topological representation of protein complexes in the network model.

The behavior of largest dense subgraph size with respect to density threshold is shown in Figure 3 for *S. Cerevisiae* and *H. Sapiens* PPI networks and their intersection. It is evident from the figure that the observed size of the largest dense subgraph follows a similar trajectory with the theoretical values estimated by both models. Moreover, in both networks, the largest dense subgraph turns out to be significant for a wide range of density thresholds. For lower values of ρ, the observed subgraphs are either not significant or are marginally significant. This is a desirable characteristic of significance-based analysis since identification of very large sparse subgraphs should

be avoided while searching for dense patterns in PPI networks. Observing that the $G(n, p)$ model becomes more conservative than the piecewise degree distribution model for lower values of ρ, we conclude that this model may facilitate fine-grain analysis of modularity in PPI networks.

4.2 Performance of SIDES

In this section we demonstrate the performance of SIDES in identification of significantly dense subgraphs on the available yeast PPI network derived from DIP and BIND databases and compare it with an existing complex identification algorithm, MCODE [4]. Both algorithms work on a set of interactions modeled as a simple graph and return a set of protein clusters, each of which induce unusually dense subgraphs in the network. MCODE associates each cluster with a score defined as the ratio of number of interactions to the number of proteins in the cluster. SIDES, on the other hand, associates each cluster with a p-value, which estimates the likelihood of observing the number of interactions between an identical number of proteins in a graph generated by the reference model, as discussed in Section 2.

We evaluate the biological relevance of identified clusters based on Gene Ontology [2]. We estimate the statistical significance of the enrichment of each GO term in the cluster using Ontologizer [10]. For a given cluster, Ontologizer associates each GO term with a p-value, which estimates the probability of the observed enrichment of the GO term in a set of randomly chosen proteins conditioned on the enrichment of the parents of the term in GO hierarchy, based on a reference model that assumes hypergeometric distribution of GO terms among proteins. In our experiments, we use the "Parent-Child" option of Ontologizer for characterization of over-representation of GO Terms. Note that the p-values reported in this section are corrected for multiple clusters

using the Bonferroni correction option provided by Ontologizer.

The distribution of the p-value for the most significant annotation with respect to cluster size for clusters identified by SIDES and MCODE on the yeast PPI network is shown in Figure 4(a). Since each cluster is generally associated with more than one significant GO term, we report the p-value that corresponds to the most significant term(s). From a statistical perspective, this term(s) correspond(s) to the most biologically meaningful annotation. On the *S. cerevisiae* PPI network, SIDES identifies 73 significantly dense subgraphs, while MCODE discovers 103 dense clusters. As evident in the figure, SIDES tends to discover smaller clusters as compared to MCODE and preserves specificity of identified clusters in terms of GO annotations irrespective of cluster size.

In order to quantify the quality of the clusters with respect to GO annotations, we use two metrics measuring the *specificity* and *sensitivity* of a cluster with respect to the associated GO term. Assume that a cluster C containing n_C proteins is associated with a term T that is attached to n_T proteins in the set of all proteins in the network. Then, if n_{CT} of the proteins in C are attached to T, we define specificity as

$$specificity = 100 \times \frac{n_{CT}}{n_C}, \tag{29}$$

measuring the purity of the cluster with respect to the corresponding term. Similarly, sensitivity is defined as

$$sensitivity = 100 \times \frac{n_{CT}}{n_T}, \tag{30}$$

measuring the extent to which the cluster represents the corresponding term.

Since a single cluster is generally associated with more than one significant annotation, we define the specificity and sensitivity of a cluster as the maximum among all significant annotations. Note that the maximum specificity and sensitivity do not necessarily correspond to the same GO

term, *i.e.*, we evaluate both methods optimistically, considering each significantly enriched term as potentially of biological relevance, since a dense cluster may indeed correspond to multiple processes. Therefore, specificity of a cluster measures the functional purity of a cluster, while sensitivity measures the ability of the cluster to represent a functional annotation alone. The scatter-plot of specificity vs. sensitivity for all clusters discovered by the two algorithms is shown in Figure 4(b). As evident in the figure, only four of the 73 SIDES clusters have specificity less than 70%. Most (62%) of the blue circles (corresponding to SIDES clusters) reside on the upper right quarter of the plane, illustrating SIDES's ability to accurately identify most of the proteins taking part in a specific process, while maintaining specificity of the enrichment of clusters. The behavior of cluster specificity and sensitivity with respect to cluster size is shown in Figure 5.

A comparison of clusters identified by SIDES and MCODE in terms of biological specificity and sensitivity is shown on Table 1. As seen in the table, SIDES is about 20% more specific and 15% more sensitive than MCODE on the yeast network on average. For a cluster, zero specificity or sensitivity corresponds to the case where no significant annotation for the cluster is found. Note that, for all of the 73 SIDES clusters, at least one GO term is significantly enriched in the cluster.

We also evaluate the performance of SIDES and MCODE on a probabilistic interaction network that is obtained through integration of various sources of interaction data [21]. In this network, each protein pair is assigned an interaction likelihood score based on statistical aggregation of experimentally observed interactions and computationally predicted functional linkages. This network is expected to be more comprehensive and less noisy compared to those that rely on a single data source. We run the two algorithms on the network of $\approx$34000 interactions with highest likelihood. These are highlighted as *confident* interactions (ConfidentNet) by Lee et al. Interestingly, the specificity and sensitivity of dense clusters identified by both algorithms on this network are sig-

nificantly lower than that on the network obtained from BIND and DIP. Namely, SIDES provides 73% specificity and 46% sensitivity on ConfidentNet, while MCODE provides 69% specificity and 37% sensitivity. This difference in the purity of identified dense clusters with respect to GO annotations may be explained as follows. While most of the interactions in BIND and DIP databases correspond to some form of physical binding, ConfidentNet integrates various forms of interaction, from physical binding to higher level functional association such as co-citation and co-evolution, which may also include indirect interactions. Therefore, the dense subgraphs in this network may correspond to higher level functional modularity, including crosstalk between various processes - resulting in functionally heterogeneous clusters. The evaluation of functional enrichment in Ontologizer, as well as the specificity and sensitivity measures, however, evaluate the homogeneity of clusters. Consequently, we speculate that the biological semantics of subgraph density may depend on the nature of interactions in the network. Specifically, dense subgraphs on networks of physical interactions are more likely to correspond to lower level modules with functional homogeneity than those on higher level networks.

As would be expected, this significant increase in accuracy comes at the price of increased computation time. In other words, MCODE is faster than SIDES since it adapts a greedy heuristic with local optimization, while SIDES solves a more expensive min-cut algorithm repeatedly and the resulting recursion tree is generally imbalanced. However, it should be noted that both algorithms are fast enough to allow online application with real-time performance for small networks and offline application with reasonable performance on larger networks. Namely, for networks of a few thousand interactions, both algorithms work in seconds, while for ConfidentNet with $\approx$34000 interactions, both MCODE and the C implementation of SIDES provide results in a few minutes.

The most significant dense subgraphs identified by SIDES in the yeast PPI network are shown

in Table 2. As seen in the table, SIDES is able to capture many protein complexes, including transcription factor complex, mRNA cleavage factor complex, proteasome complex, nuclear ubiquitin ligase complex, mediator complex, schistoseome complex, exosome, oligosaccharyl transferase complex, TRAPP complex, eukaryotic transcription initiation factor 2B complex, hydrogen-translocating V-type ATPase complex, CCR4-NOT complex, HOPS complex, and transcription export complex. The modularity of many fundamental processes is also captured by SIDES. For example, 12 nuclear ubiquitin ligase complex proteins that induce a subgraph of 62 interactions make up 91.7% of the proteins that take part in cyclin metabolism. A complete list of protein clusters that induce significantly dense subgraphs, which may be regarded as putative functional modules, are also available at the SIDES website.

Significant dense subgraphs that are conserved in *S. cerevisiae* and *H. sapiens* PPI networks are shown in Table 3. Most of these dense components are involved in fundamental processes and the proteins that are parts of these components share a particular function. Among these, the 7-protein conserved subnet that consists of 6 Exosomal 3'-5' exoribonuclease complex subunits and Succinate dehydrogenase is interesting. As in the case of dense subgraphs in a single network, the conserved dense subgraphs provide an insight into the crosstalk between proteins that perform different functions. For example, the largest conserved subnet of 11 proteins contains Mismatch repair proteins, Replication factor C subunits, and RNA polymerase II transcription initiation/nucleotide excision repair factor TFIIH subunits, which are all involved in DNA repair. The conserved subnets identified by SIDES are small and appear to be partial, since we employ a strict interpretation of conserved interaction here. In particular, limiting the ortholog assignments to proteins that have a COG assignment and considering only matching direct interactions as conserved interactions, limits the ability of the algorithm to identify a comprehensive set of conserved dense

graphs. Algorithms that rely on sequence alignment scores and consider indirect or probable interactions [19, 29, 30] coupled with adaptation of the statistical framework presented in this paper have the potential of increasing the coverage of identified patterns, while correctly evaluating the interestingness of observed patterns.

5 Conclusion

In this paper, we present a technique for analytically assessing statistical significance of connectivity and conservation in PPI networks. Specifically, we examine the occurrence of *dense* subgraphs, which forms one of the most well-studied pattern structures in extracting biologically novel information from PPI networks. While the analysis based on the $G(n, p)$ model and its extension provides a good way of assessing significance, models that mirror the topological characteristics of PPI networks should be further analyzed. This paper provides a stepping stone for the analysis of such complicated models.

Acknowledgments

This work is supported in part by the NIH Grant R01 GM068959-01, and the NSF Grants CCR-0208709, CCF-0513636, DMS-0503742.

References

[1] W. Aiello, F. Chung, and L. Lu. A random graph model for power law graphs. In *Proc. ACM Symp. Theory of Computing*, pages 171–180, 2000.

[2] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.

[3] G. D. Bader, I. Donalson, C. Wolting, B. F. Quellette, T. Pawson, and C. W. Hogue. BIND-The Biomolecular Interaction Network Database. *Nuc. Acids Res.*, 29(1):242–245, 2001.

[4] G. D. Bader and C. W. V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(2), 2003.

[5] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[6] B. Bollobás. *Random Graphs*. Cambridge University Press, Cambridge, UK, 2001.

[7] C. Brun, C. Herrmann, and A. Guénoche. Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics*, 5(95), 2004.

[8] F. Chung, L. Lu, and V. Vu. Spectra of random graphs with given expected degrees. *PNAS*, 100(11):6313–6318, 2003.

[9] A. del Sol, H. Fujihashi, and P. O'Meara. Topology of small-world networks of protein-protein complex structures. *Bioinformatics*, 21(8):1311–1315, 2005.

[10] S. Grossmann, S. Bauer, P. N. Robinson, and M. Vingron. An improved statistic for detecting over-represented gene ontology annotations in gene sets. In *10th International Conference on Research in Computational Molecular Biology (RECOMB'06)*, pages 85–98, 2006.

[11] L. Hagen and A. B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 11(9):1074–85, 1992.

[12] J.-D. J. Han, D. Dupuy, N. Bertin, M. E. Cusick, and M. Vidal. Effect of sampling on topology predictions of protein interaction networks. *Nat. Biotech.*, 23(7):839–844, 2005.

[13] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76:171–181, 2000.

[14] H. Hu, X. Yan, Y. Huang, J. Han, and X. J. Zhou. Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, 21:i213–i221, 2005.

[15] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon. Subgraphs in random networks. *Physical Review E*, 68(026127), 2003.

[16] H. Jeong, S. P. Mason, A. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

[17] B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *PNAS*, 100(20):11394–11399, 2003.

[18] M. Koyutürk, A. Grama, and W. Szpankowski. An efficient algorithm for detecting frequent subgraphs in biological networks. In *Bioinformatics (ISMB'04)*, pages i200–i207, 2004.

[19] M. Koyutürk, Y. Kim, S. Subramaniam, W. Szpankowski, and A. Grama. Detecting conserved interaction patterns in biological networks. *Journal of Computational Biology*, 13(7):1299–1322, 2006.

[20] M. Koyutürk, Y. Kim, U. Topkara, S. Subramaniam, W. Szpankowski, and A. Grama. Pairwise alignment of protein interaction networks. *Journal of Computational Biology*, 13(2):182–189, 2006.

[21] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte. A probabilistic functional network of yeast genes. *Science*, 306(5701):1555–1558, 2004.

[22] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis. Detection of functional modules from protein interaction networks. *Proteins*, 54(1):49–57, 2004.

[23] N. Pržulj. Graph theory analysis of protein-protein interactions. In I. Jurisica and D. Wigle, editors, *Knowledge Discovery in Proteomics*. CRC Press, 2004.

[24] N. Pržulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric?. *Bioinformatics*, 20(18):3508–3515, 2004.

[25] N. Pržulj, D. A. Wigle, and I. Jurisica. Functional topology in a network of protein interactions. *Bioinformatics*, 20(3):340–348, 2004.

[26] A. W. Rives and T. Galitski. Modular organization of cellular networks. *PNAS*, 100(3):1128–1133, 2003.

[27] D. Scholtens, M. Vidal, and R. Gentleman. Local modeling of global interactome networks. *Bioinformatics*, 21(17):3548–57, 2005.

[28] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–504, 2003.

[29] R. Sharan, T. Ideker, B. P. Kelley, R. Shamir, and R. M. Karp. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. In *RECOMB'04*, pages 282–289, 2004.

[30] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *PNAS*, 102(6):1974–1979, 2005.

[31] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *PNAS*, 100(21):12123–12128, 2003.

[32] M. Stoer and F. Wagner. A simple min-cut algorithm. *J. ACM*, 44(4):585–591, 1997.

[33] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, New York, 2001.

[34] R. Tatusov, N. Fedorova, J. Jackson, A. Jacobs, B. Kiryutin, and E. Koonin. The COG database: An updated version includes eukaryotes. *BMC Bioinformatics*, 4(41), 2003.

[35] A. Thomas, R. Cannings, N. A. Monk, and C. Cannings. On the structure of protein-protein interaction networks. *Biochem Soc Trans.*, 31(6):1491–6, 2003.

[36] S. Tornow and H. W. Mewes. Functional modules by relating protein interaction networks and gene expression. *Nuc. Acids Res.*, 31(21):6283–6289, 2003.

[37] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Bio Evol*, 18(7):1283–92, 2001.

[38] A. Wagner. How the global structure of protein interaction networks evolves. *Proc. R. Soc. Lond. Biol. Sci.*, 270(1514):457–466, 2003.

[39] S. Wang and J. M. Siskind. Image segmentation with ratio cut. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6):675–690, 2003.

[40] M. Waterman. *Introduction to Computational Biology*. Chapman & Hall, London, 1995.

[41] M. S. Waterman and M. Vingrons. Rapid and accurate estimates of statistical significance for sequence data base searches. *PNAS*, 91:4625–28, 1994.

[42] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. Kim, and D. Eisenberg. DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. *Nuc. Acids Res.*, 30:303–305, 2002.

[43] S. H. Yook, Z. N. Oltvai, and A. L. Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, April 2004.

procedure MINCUTPHASE(Subgraph S, Node $s \in V(S)$)
▷ grows graph starting from seed node s by adding most heavily connected nodes
▷ **returns** the last two nodes and the cut between last node and others
▷ $w(uv)$: number of edges between nodes represented by u and those represented by v
$\mathcal{V} \leftarrow \{s\}$
while $|\mathcal{V}| < |V(S)| - 1$ **do**
$v \leftarrow \text{argmax}_{v' \in V(S)} \sum_{u' \in \mathcal{V}} w(u'v')$
$\mathcal{V} \leftarrow \mathcal{V} \cup \{v\}$
$u \leftarrow V(S) \setminus \mathcal{V}$
return $\{v, u, \sum_{u' \in \mathcal{V}} w(u'u)\}$

procedure RATIOCUTPARTITION(Subgraph S)
▷**returns** partition that locally minimizes ratio-cut
▷ $w(u)$: number of nodes represented by u
for $u \in V(S)$ **do**
$w(u) \leftarrow 1$
$W \leftarrow |V(S)|$
$\bar{R} \leftarrow E(S) + 1$
pick arbitrary seed node $s \in V(S)$
while $|V(S)| > 1$ **do**
$\{v, u, C\} \leftarrow$ MINCUTPHASE(S, s)
$R = C / \min(w(u), W - w(u))$
if $R < \bar{R}$ **then** $\bar{R} \leftarrow R$
merge u into v, $w(v) \leftarrow w(v) + w(u)$
return partition that corresponds to $\bar{R}$

procedure RECURSIVERATIOCUT(Subgraph S, Integer n, Real p)
▷**returns** set of dense subgraphs of S that are significant w.r.t. n and p
$\rho \leftarrow |E(S)|/|V(S)|^2$
Estimate r_0 as given by (5)
if $|V(S)| > r_0$ **then**
Estimate significance of S as given by (21)
return $\{S\}$
else
$\{S_0, S_1\} \leftarrow$ RATIOCUTPARTITION(S)
return RECURSIVERATIOCUT(S_0, n, p) $\bigcup$ RECURSIVERATIOCUT(S_1, n, p)

procedure SIDES(Network G)
▷**returns** set of significantly dense subgraphs of G
$p \leftarrow \max_{u \in V} |\{v \in V(G) : uv \in E(G)\}|/|V(G)|$
return RECURSIVERATIOCUT(G, $|V(G)|$, p)

Figure 1: SIDES algorithm for identifying significantly dense subgraphs in a network, based on recursive ratio-cut partitioning.

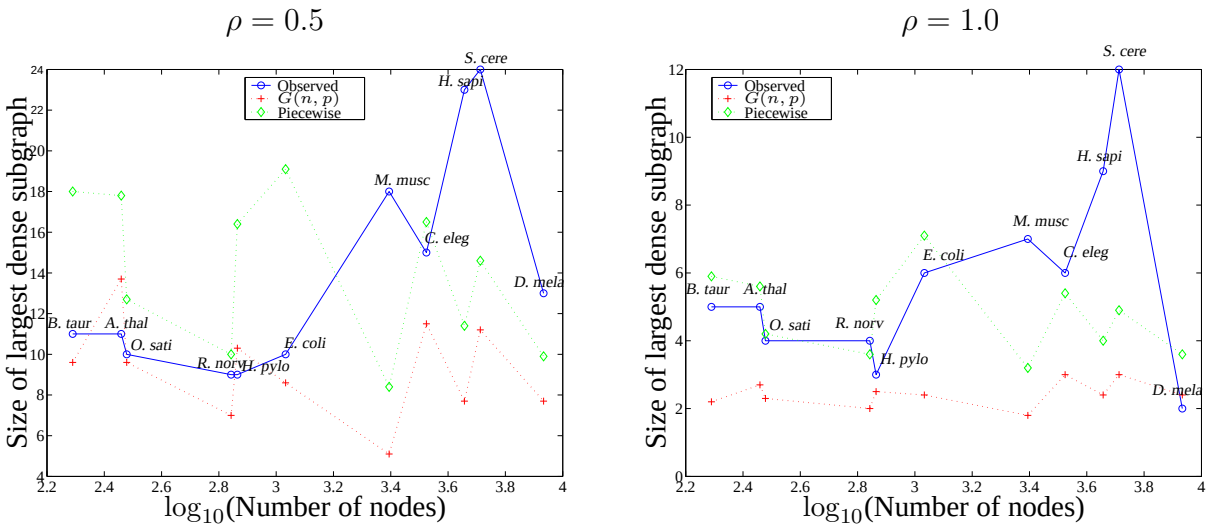


Figure 2: The behavior of the size of largest dense subgraph with respect to number of proteins in the network where a subgraph is considered dense if $\rho = 0.5$ and $\rho = 1.0$ (clique), respectively. Each sample point corresponds to the PPI network of a particular species, as marked by its name. The critical values of largest dense subgraph size based on $G(n, p)$ and piecewise degree distribution models are also shown.

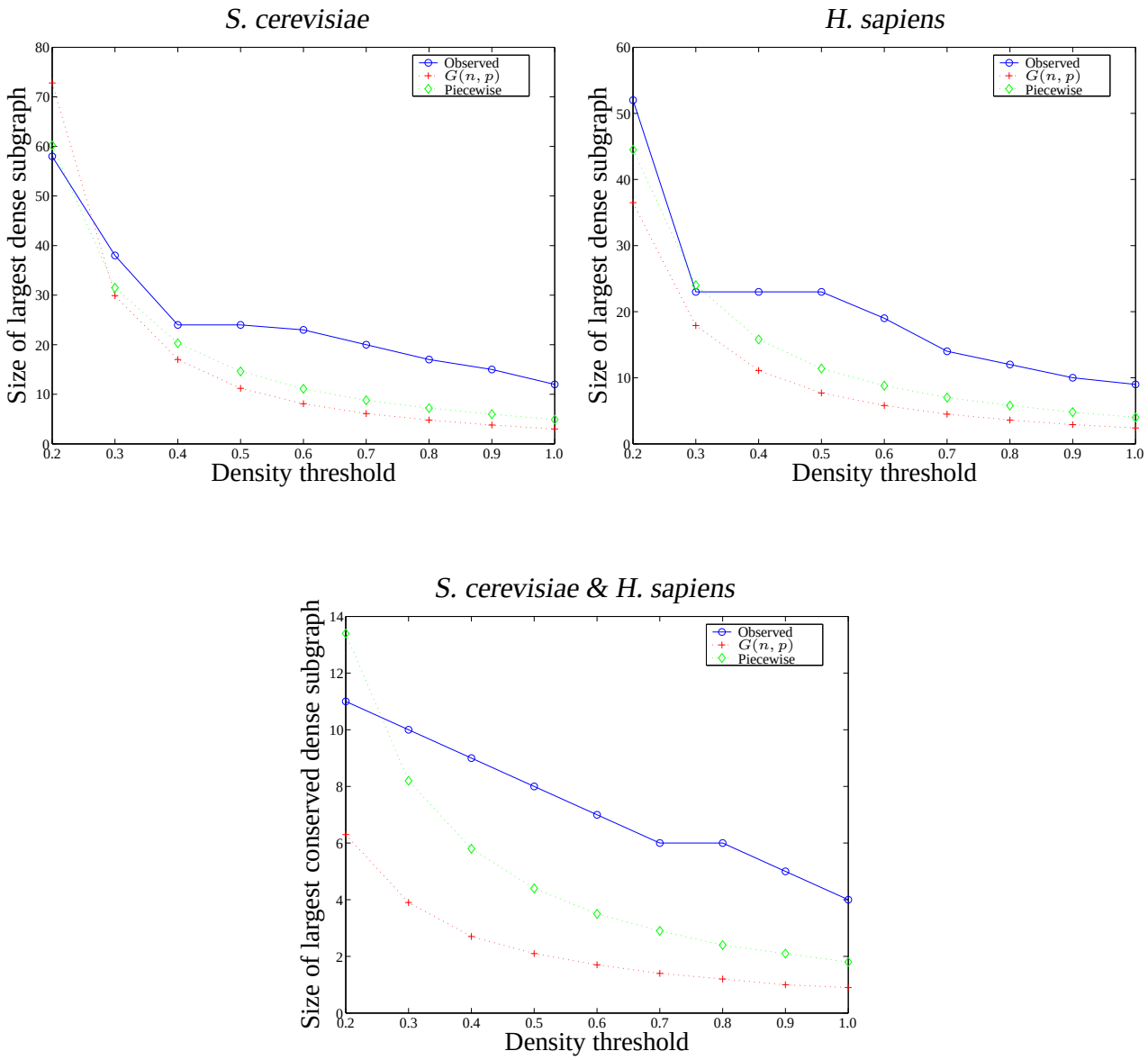
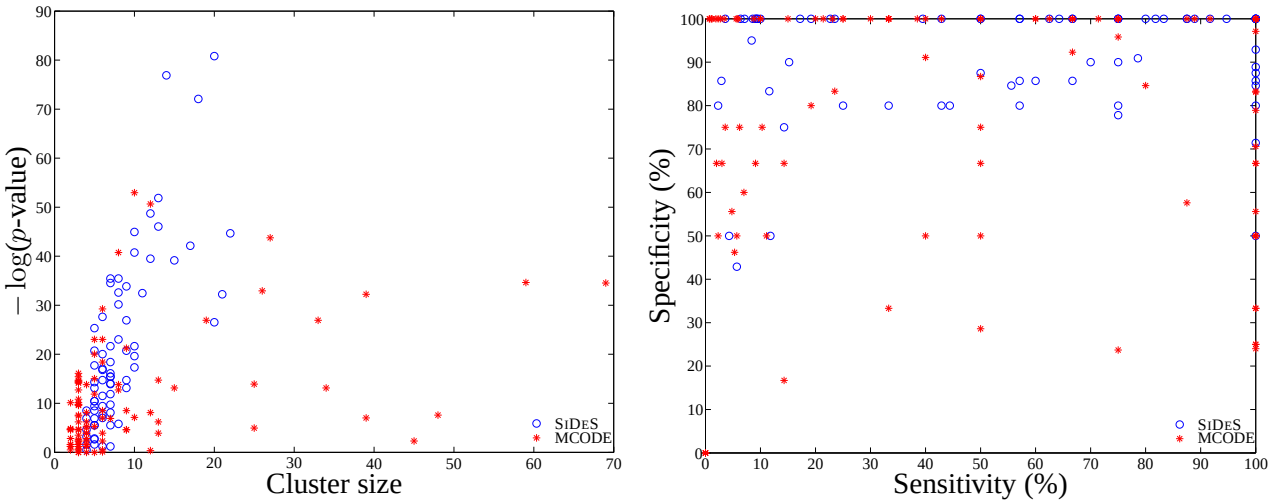
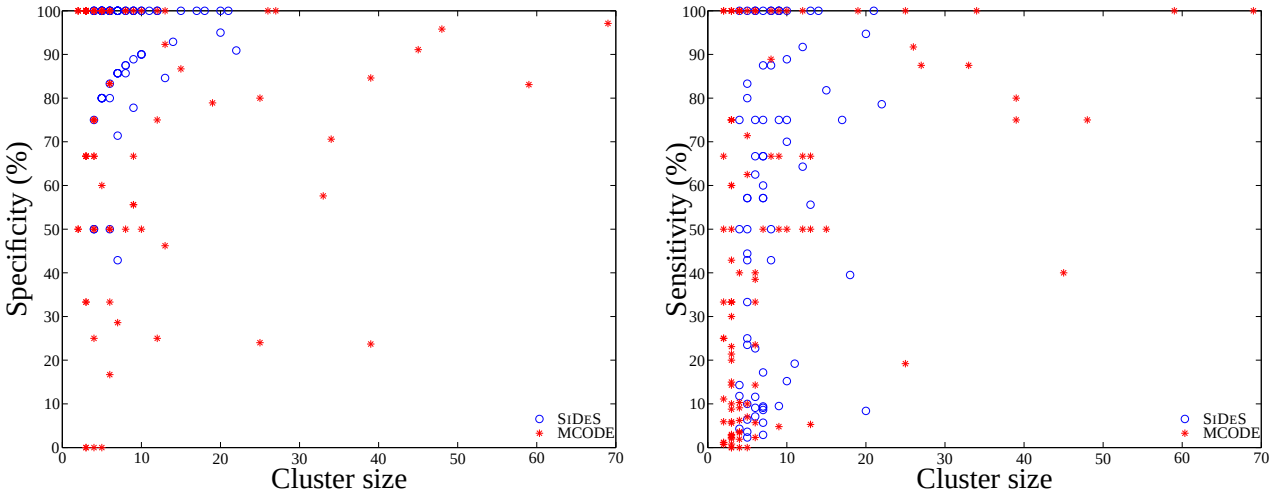


Figure 3: Behavior of the size of the largest dense subgraph and largest conserved dense subgraph with respect to density threshold (ρ) for *S. cerevisiae* and *H. sapiens* PPI networks. Critical values of largest dense subgraph size based on $G(n,p)$ and piecewise degree distribution models are also shown.



(a) (b)

Figure 4: (a) The behavior of the significance of attached GO annotation with minimum p-value with respect to cluster size for the dense clusters identified by the SIDES and MCODE algorithms. Cluster size and significance of GO annotation are significantly correlated (0.76, $p < 9e - 15$) for SIDES, showing that SIDES is able to tune the size of cluster to accurately capture the "meaning". The correlation of size and significance for MCODE is 0.43 ($p < 5e - 06$). (b) Sensitivity vs Specificity of clusters identified by the two algorithms. Only four of the 73 SIDES clusters have specificity less than 70%. Most (62%) of the blue circles are clustered on the upper right quarter of the plane, illustrating SIDES's ability to accurately identify most of the proteins taking part in a specific process, while maintaining specificity of the enrichment of clusters.



(a) (b)

Figure 5: The behavior of specificity and sensitivity with respect to cluster size for dense clusters identified by the SIDES and MCODE algorithms. (a) Size vs Specifity, correlation for SIDES is 0.22 ($p < 0.06$), while it is -0.02 ($p < 0.83$) for MCODE. Note that if the clusters were constructed randomly, size and specificity would be negatively correlated. The positive correlation for SIDES's clusters is illustrative of SIDES's ability of tuning cluster size to optimize specificity. (b) Size vs Sensitivity, correlation for SiDeS is 0.27 ($p < 0.02$), while it is 0.36 ($p < 2e - 04$) for MCODE. If the clusters were constructed at random, one would expect strong positive correlation between size and sensitivity.

Table 1: Comparison of SIDES and MCODE algorithms in terms of their specificity and sensitivity with respect to GO annotations.

	SIDES			MCODE		
	Min.	Max.	Avg.	Min.	Max.	Avg.
Specificity(%)	43.0	100.0	91.2	0.0	100.0	77.8
Sensitivity(%)	2.0	100.0	55.8	0.0	100.0	47.6

Table 2: The most significant protein clusters that induce dense subgraphs on the *S. cerevisiae* PPI network and their annotation.

Size (#P, #I)	Density p-value	Annotation	% Spec.	% Sens.	Annotation p-value
(22, 145)	2e-234	[F] transcription regulator activity	90.9	6.9	4e-20
		[C] transcription factor complex	90.9	17.1	6e-20
		[P] protein amino acid acylation	63.6	32.6	1e-11
(21, 123)	3e-181	[C] cytoplasmic mRNA processing body	36.8	100.0	1e-14
		[P] mRNA metabolism	94.7	10.2	2e-05
(20, 114)	1e-169	[P] cytoplasm organization and biogenesis	90.0	8.4	3e-12
		[C] nucleolus	80.0	7.7	1e-09
(20, 112)	4e-163	[C] mRNA cleavage factor complex	90.0	94.7	8e-36
		[P] RNA 3'-end processing	80.0	69.6	2e-16
(18, 94)	5e-138	[C] proteasome complex (sensu Eukaryota)	94.4	39.5	5e-32
		[P] proteolysis	94.4	11.0	3e-10
		[F] peptidase activity	83.3	15.5	1e-09
(12, 62)	2e-134	[C] nuclear ubiquitin ligase complex	100.0	47.8	2e-20
		[P] cyclin catabolism	100.0	91.7	2e-14
		[F] ligase activity	90.9	9.9	8e-11
(17, 82)	2e-114	[F] transcription regulator activity	100.0	5.9	5e-19
		[C] mediator complex	88.2	75.0	6e-10
		[P] transcription	100.0	3.8	3e-06
(15, 64)	6e-85	[C] spliceosome complex	93.3	18.9	1e-17
		[F] binding	100.0	1.7	2e-09
		[P] mRNA processing	100.0	11.8	1e-05
(14, 55)	5e-69	[C] exosome (RNase complex)	92.9	100.0	4e-34
		[P] mRNA catabolism	92.9	25.5	2e-06
(10, 38)	1e-66	[C] oligosaccharyl transferase complex	100.0	88.9	2e-18
		[P] glycoprotein metabolism	100.0	15.1	9e-09
		[F] oligosaccharyl transferase activity	100.0	88.9	3e-07
(13, 48)	5e-59	[C] proteasome complex (sensu Eukaryota)	84.6	25.6	1e-20
		[P] biopolymer catabolism	76.9	4.5	1e-05
(13, 48)	5e-59	[C] TRAPP complex	76.9	100.0	3e-23
		[C] Golgi cis-face	76.9	76.9	2e-11
		[P] ER to Golgi vesicle-mediated transport	76.9	15.2	1e-03
(10, 35)	7e-54	[C] Golgi apparatus	80.0	5.4	3e-08
		[C] cytoplasmic membrane-bound vesicle	70.0	8.0	3e-07
		[P] Golgi vesicle transport	90.0	6.9	2e-04
(12, 42)	1e-51	[C] hydrogen-translocating V-type ATPase complex	75.0	64.3	2e-15
		[P] vacuolar transport	100.0	19.7	2e-09
		[P] regulation of pH	100.0	50.0	2e-06
(9, 30)	8e-49	[C] eukaryotic TIF 2B complex	55.6	100.0	2e-12
		[F] translation regulator activity	77.8	15.2	3e-11
		[P] macromolecule biosynthesis	88.9	2.0	2e-03
(10, 32)	4e-42	[C] CCR4-NOT complex	90.0	75.0	3e-09
		[P] regulation of RNA metabolism	60.0	27.3	4e-06
(11, 33)	3e-33	[C] cell cortex	100.0	11.6	8e-15
		[P] cytoskeleton organization and biogenesis	72.7	4.0	3e-04
(9, 26)	2e-32	[C] spliceosome complex	77.8	9.5	4e-07
		[P] RNA splicing	88.9	7.0	2e-02
(10, 29)	4e-31	[C] proton-transporting ATP synthase complex	90.0	56.2	3e-20
		[P] hydrogen transport	90.0	50.0	2e-14
(8, 22)	2e-29	[C] histone methyltransferase complex	87.5	87.5	7e-15
		[P] protein amino acid alkylation	87.5	36.8	2e-07
		[F] protein methyltransferase activity	87.5	50.0	8e-04
(7, 17)	1e-21	[F] DNA clamp loader activity	57.1	57.1	5e-05
		[P] DNA replication	100.0	6.9	3e-03
		[C] replication fork	85.7	15.4	4e-03
(8, 20)	2e-21	[C] HOPS complex	75.0	100.0	8e-14
		[P] vacuole organization and biogenesis	75.0	17.6	1e-07
(8, 19)	1e-17	[C] transcription export complex	71.4	71.4	1e-10
		[P] establishment of RNA localization	85.7	8.5	5e-05
(7, 15)	3e-13	[C] exocyst	100.0	87.5	4e-16
		[P] exocytosis	100.0	20.0	1e-05

Table 3: Seven most significant conserved dense subgraphs identified in *S. cerevisiae* and *H. sapiens* PPI networks by the modified HCS algorithm and their functional enrichment according to COG functional annotations.

# Prot	# Cons Int	$p <$	COG Annotation
10	17	10^{-68}	RNA polymerase (100%)
11	11	10^{-26}	Mismatch repair (33%)
			RNA polymerase II TI/nucleotide excision repair factor TFIIH (33%)
			Replication factor C (22%),
7	7	10^{-25}	Exosomal 3'-5' exoribonuclease complex (86%)
4	4	10^{-24}	Single-stranded DNA-binding replication protein A (50%)
			DNA repair protein (50%)
5	4	10^{-12}	Small nuclear ribonucleoprotein(80%)
			snRNP component (20%)
5	4	10^{-12}	Histone (40%)
			Histone transcription regulator (20%)
			Histone chaperone (20%)
3	3	10^{-9}	Vacuolar sorting protein (33%)
			RNA polymerase II transcription factor complex subunit (33%)
			Uncharacterized conserved protein (33%)