Assessing Significance of Connectivity and Conservation in Protein Interaction Networks*

Mehmet Koyutürk, Ananth Grama, and Wojciech Szpankowski
Department of Computer Sciences
Purdue University
W. Lafayette, IN 47907, U.S.A.
{koyuturk,ayg,spa}@cs.purdue.edu

Abstract

Computational and comparative analysis of protein-protein interaction (PPI) networks enable understanding of the modular organization of the cell through identification of functional modules and protein complexes. These analysis techniques generally rely on topological features such as connectedness, based on the premise that functionally related proteins are likely to interact densely and that these interactions follow similar evolutionary trajectories. Significant recent work in our lab, and in other labs has focused on efficient algorithms for identification of modules and their conservation. Application of these methods to a variety of networks has yielded novel biological insights.

In spite of algorithmic advances, development of a comprehensive infrastructure for interaction databases is in relative infancy compared to corresponding sequence analysis tools such as BLAST and CLUSTAL. One critical component of this infrastructure is a measure of the statistical significance of a match or a dense subcomponent. Corresponding sequence-based measures such as E-values are key components of sequence matching tools. In the absence of an analytical measure, conventional methods rely on computer simulations based on ad-hoc models for quantifying significance. This paper presents the first such effort, to the best of our knowledge, aimed at analytically quantifying statistical significance of dense components and matches in reference model graphs. We consider two reference graph models – a $G(n,p)$ model in which each pair of nodes 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 occurring in PPI networks. We argue that by choosing conservatively the value of p, the $G(n,p)$ model will dominate that of the power-law graph that is often used to model PPI networks. We also propose a method for evaluating statistical significance based on the results derived from this analysis, and demonstrate the use of these measures for assessing significant structures in PPI networks. Experiments performed on a rich collection of PPI networks show that the proposed model provides a reliable means of evaluating statistical significance of dense patterns in these networks.

*This work is supported in part by the NIH Grant R01 GM068959-01, and the NSF Grants CCR-0208709, CCF-0513636, DMS-0202950.

1 Introduction

Availability of high-throughput methods for identifying protein-protein interactions has resulted in a new generation of extremely valuable data [2, 36]. Effective analysis of the interactome holds the key to functional characterization, phenotypic mapping, and identification of pharmacological targets, among other important tasks. Computational infrastructure for supporting analysis of the interactome is in relative infancy, compared to its sequence counterparts [34]. 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) [3, 6, 18, 19, 22, 26]. These methods are based on the premise that functionally related proteins generally manifest themselves as dense components in the network [31]. The hypothesis that proteins performing together a particular cellular function 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 on these superposed graphs [11, 14, 15, 16, 24, 25].

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 it is to observe the pattern in the reference model that often is a probabilistic measure for scoring patterns. 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 (expected) behavior. One such score, in the context of sequences is the E-value returned by BLAST matches [35]. 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, there are, to the best of our knowledge, no analytical methods for quantifying the significance of such patterns. For this reason, existing algorithms for detecting patterns generally adopt simple ad-hoc measures (such as relative density, diameter) to assess the significance of identified patterns. This paper represents the first such effort at analytically quantifying the statistical significance of a pattern 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.

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. 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 assure that our $G(n, p)$ model stochastically dominates that of a power-law graph. Our two-level $G(n, p)$ model is shown 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 [10]

can be modified to incorporate statistical significance in identification of dense patterns. We also generalize these 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 [2] and DIP [36]. 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. The framework proposed here can also be extended to include more general networks that capture the degree distribution of PPI networks more accurately, namely power-law [33, 37], geometric [20], or exponential [8] degree distributions.

The rest of this manuscript is organized as follows: In the next section, we first 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. We present and discuss experimental results in Section 3. We present proofs of important analytical 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 [31], many commonly used methods focus on discovering dense regions of the network for identification of functional modules or protein complexes [3, 6, 18, 22, 26]. Subgraph density is also central for many algorithms that target identification of conserved modules and complexes [11, 15, 24]. 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 feasible theoretical and computational analysis.

2.1 Modeling PPI Networks

With the increasing availability of high-throughput interaction data, there has been significant effort on 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 [12, 21]. A commonly accepted model that confirms this observation is based on power-law degree distribution [4, 32, 33, 37]. 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 [9, 30]. In this respect, alternative models that are based on geometric [20] or exponential [8] 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 the interactions in the network are independent of each other [13, 15, 24]. Since degree distribution is critical for 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 many 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 [24]. Alternately, they can be estimated analytically, by relying on a simple random graph model that is based on a given degree distribution [7]. 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 $P(uv \in E(G)) = d_u d_v / |E(G)|$. In

order for this function to be a well-defined probability measure, we must have $d_{\max}^2 \leq |E(G)|$, where $d_{\max} = \max_{u \in V(G)} d_u$. However, available protein interaction data generally does not confirm this assumption. For example, based on the PPI networks we derive from BIND [2] and DIP [36] 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 [33, 37], geometric [20], or exponential [8] 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 on the simple and well-studied $G(n,p)$ model and attempt to generalize our results to more complicated models that assume heterogeneous degree distribution. In the forthcoming work we turn our attention such graphs.

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 $\delta(U) = |F(U)|/|U|^2$. 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 $F(U) \geq \rho|U|^2$, where $\rho > p$. 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. We assume that the edges are directed and self-loops are allowed. Note that PPI networks are undirected graphs and they contain self-loops in general, but any undirected network can be easily modeled by a directed graph.

Let random variable R_ρ be the size of the maximum subset of vertices that induce a ρ-dense subgraph, *i.e.,*

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

The behavior of R_1, which corresponds to maximum clique, is well studied on $G(n,p)$ model and its typical value is shown to be $O(\log_{1/p} n)$ [5]. In the following theorem, we present a general result for the typical value of R_ρ for any ρ.

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

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

where

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

More precisely,

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

where

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

for large n.

The proof of this theorem is presented in Section 4. 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 or not.

For a graph of arbitrary distribution, let $d_{\max}$ denote the maximum expected degree as defined in Section 2.1. Let $p_{\max} = d_{\max}/n$. It can be easily shown that the largest dense subgraph in the $G(n,p)$ graph with $p = p_{\max}$ stochastically dominates that in the random graph generated according to the given degree distribution (e.g., power-law graphs). Hence, by estimating the edge probability conservatively, we can use the above result to determine whether a dense subgraph identified in a PPI network of arbitrary degree distribution is statistically significant. Moreover, 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, as we show (cf. (14)) in the proof of Theorem 1 in Section 4, the probability of observing this subgraph in a graph generated according to the reference model is bounded by

$$P(R_{\hat{\rho}} \geq (1+\epsilon)\log n/\kappa(\hat{\rho},p)) \leq \frac{\sqrt{1-\rho}}{2\pi\sqrt{\rho}} \frac{(1+\epsilon)\log n}{n^{\epsilon(1+\epsilon)\log n/\kappa(\hat{\rho},p)}}. \tag{6}$$

While these results on $G(n,p)$ 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:

$$P(uv \in E(G)) = \begin{cases} p_h & \text{if } u,v \in V_h \\ p_l & \text{if } u,v \in V_l \\ p_b & \text{if } u \in V_h, v \in V_l \text{ or } u \in V_l, v \in V_h \end{cases} \tag{7}$$

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 [12, 21]. 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.

Note that the random graph model described above can be generalized to an arbitrary number of node classes to capture the underlying degree distribution more accurately. Indeed, with appropriate adjustment of certain parameters, this model will converge to power-law or exponential degree distribution at the limit with increasing number of node classes. In fact, our experiments indicate that the piecewise graph model is better suited than the power-law model. However, in order to get a better fit we need to introduce three or four classes in our piecewise model.

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 (7). If $n_h = O(1)$, then*

$$P(R_\rho \geq r_1) \leq O\left(\frac{\log n}{n^{1/\kappa(p_l,\rho)}}\right), \tag{8}$$

where

$$r_1 = \frac{\log n - \log\log n + 2n_h \log B + \log \kappa(p_l,\rho) + \log e + 1}{\kappa(p_l,\rho)} \tag{9}$$

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

Note that the above result is based on asymptotic behavior of r_1, hence the $\log n/\kappa(p_l,\rho)$ 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.

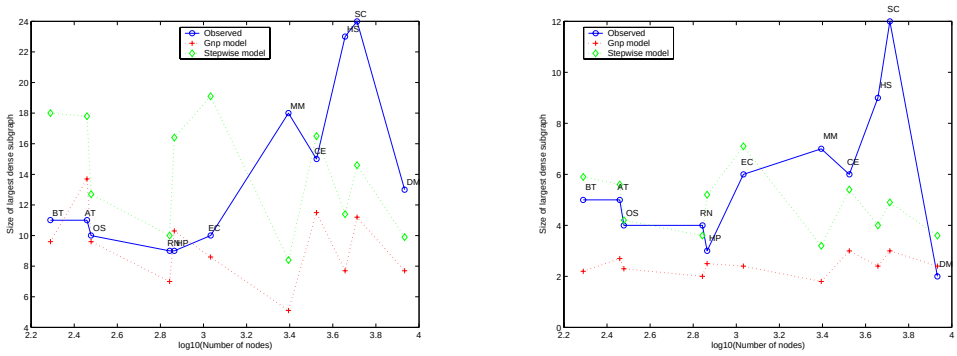
2.4 Identifying significant dense subgraphs

We use the above results to modify an existing state-of-the-art graph clustering algorithm, HCS [10], 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 repeated application of min-cut partitioning. The density of any subgraph found in this recursive decomposition is compared with a pre-defined density threshold. If the 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 [19], 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 (9). If the size of the subgraph is larger than this probabilistic upper-bound, then we report the subgraph as being statistically significant. Otherwise, we continue partitioning the graph. Note that this algorithm only identifies disjoint subgraphs, but can be easily extended to obtain overlapping dense subgraphs by greedily growing the resulting graphs until significance is lost.

2.5 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 [13, 15, 16, 24, 25]. Here, we rely on ortholog groups available in the COG database [29] to relate proteins in different PPI networks [16]. 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 putting 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



$\rho = 0.5$ $\rho = 1.0$

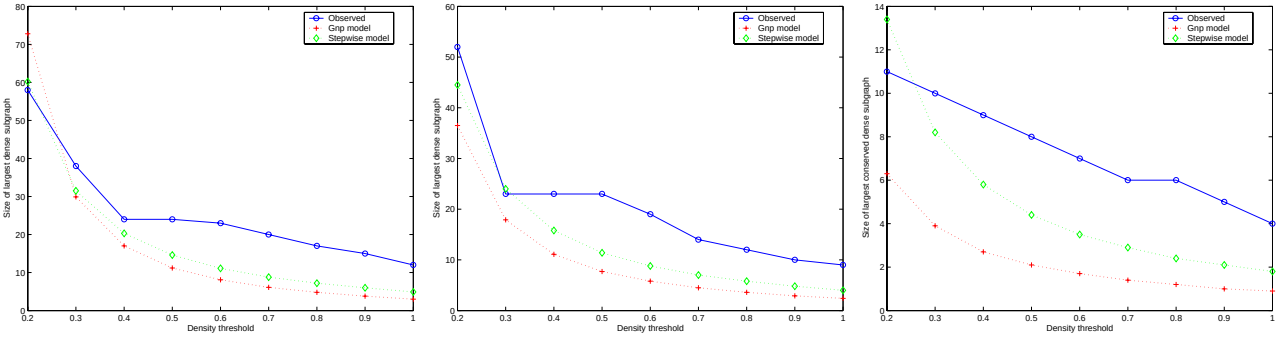
Figure 1: The behavior of the size of largest dense subgraph with respect to number of proteins in the network where a subgraph is considered dense if $\rho = 0.5$ and $\rho = 1.0$ (clique), respectively. Each sample point corresponds to the PPI network of a particular species, as marked by the initials of its name. The typical values of largest dense subgraph size based on $G(n, p)$ and piecewise degree distribution models are also shown.

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 here assumes that the orthology relationship between proteins in the two networks is already established and estimates the conditional probability that the interactions between these given ortholog proteins are densely conserved.

3 Results and Discussion

In this section, we experimentally analyze connectivity and conservation in PPI networks of 11 species gathered from BIND [2] and DIP [36] 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 portion of their proteins.

In Figure 1, we inspect the behavior of largest subgraph with respect to number of nodes in the PPI network for two different values of density threshold (ρ). In the figure, each organism corresponds to a sample point, which is marked with the initials of 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 mainly 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 *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) [23]. 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



S. cerevisiae *H. sapiens* *S. cerevisiae & H. sapiens*

Figure 2: 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. Typical values of largest dense subgraph size based on $G(n, p)$ and piecewise degree distribution models are also shown.

complexes in the network model.

The behavior of largest dense subgraph size with respect to density threshold is shown in Figure 2 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 they 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.

We implement the modified HCS heuristic described in Section 2.4 using a simple min-cut algorithm [27]. A selection of most significant dense subgraphs discovered on *S. cerevisiae* PPI network are shown in Table 1. In the table, as well as the size, density and significance of identified subgraphs, we list the GO annotations that are shared by most of the proteins in the dense subgraph. The GO annotations may refer to function [F], process [P], or component [C]. For most of the significant dense subgraphs, most of the proteins that induce the subgraph are involved in the same cellular process. As an extreme case, the algorithm also identifies proteins that share a common function or that are part of a particular complex. For example, the dense subgraph of 7 proteins in the last row corresponds to the mitochondrial outer membrane translocase (TOM) complex, which mediates recognition, unfolding, and translocation of preproteins [17]. On the other hand, some dense subgraphs correspond to proteins that are involved in a range of processes but localize in the same cellular component, such as the largest dense subgraph identified by modified HCS, which contains 24 proteins.

The significant dense subgraphs that are conserved in *S. cerevisiae* and *H. sapiens* PPI networks are shown in Table 2. Most of these dense components are involved in fundamental processes and the proteins that are part 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 on 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 the modified HCS algorithm are small and appear to be partial, since we employ a strict understanding of conserved interaction here. In particular,

Table 1: Seven most significant dense subgraphs identified in *S. cerevisiae* PPI network by the modified HCS algorithm.

# Proteins	# Interactions	Significance	GO Annotation
10	45	$p < 1e-200$	[P] ER to Golgi transport (90%) [C] TRAPP complex (90%)
20	138	$p < 8e-187$	[P] ubiquitin-dependent protein catabolism (80%) [F] endopeptidase activity (50%)
24	165	$p < 4e-175$	[C] nucleolus (54%) [C] nucleus (46%)
16	104	$p < 3e-173$	[P] histone acetylation (62%) [C] SAGA complex (56%) [P] chromatin modification (56%)
15	90	$p < 8e-145$	[F] RNA binding (80%) [C] mRNA cleavage & polyadenylation specificity fac. comp. (80%) [P] mRNA polyadenylylation (80%)
14	79	$p < 3e-127$	[P] mRNA catabolism (71%) [F] RNA binding (64%) [P] nuclear mRNA splicing, via spliceosome (57%)
7	20	$p < 9e-30$	[C] mitochondrial outer membrane translocase complex (100%) [F] protein transporter activity (100%) [P] mitochondrial matrix protein import (100%)

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 [24, 25, 16] coupled with adaptation of the statistical framework in this paper have the potential of increasing the coverage of identified patterns, while correctly evaluating the interestingness of observed patterns.

4 Proof of Theorems

In this section we prove Theorems 1 and 2.

Proof of Theorem 1: 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

$$P(R_\rho \geq r) \leq P(X_{r,\rho} \geq 1) \leq \mathbf{E}[X_{r,\rho}]. \tag{10}$$

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

$$P(Y_r \geq \rho r^2) \leq (r^2 - \rho r^2) P(Y_r = \rho r^2) \leq \binom{r^2}{\rho r^2} (r^2 - \rho r^2) p^{\rho r^2} (1-p)^{r^2 - \rho r^2}. \tag{11}$$

Hence, we get

$$P(R_\rho \geq r) \leq \binom{n}{r} \binom{r^2}{\rho r^2} (r^2 - \rho r^2) p^{\rho r^2} (1-p)^{r^2 - \rho r^2}. \tag{12}$$

Table 2: Seven most significant conserved dense subgraphs identified in *S. cerevisiae* and *H. sapiens* PPI networks by the modified HCS algorithm.

# Proteins	# Conserved Interactions	Significance	COG Annotation
10	17	$p < 1.7e-69$	RNA polymerase (100%)
11	11	$p < 4.2e-27$	Mismatch repair (33%) RNA polymerase II TI/nucleotide excision repair factor TFIIH (33%) Replication factor C (22%),
7	7	$p < 9.9e-26$	Exosomal 3'-5' exoribonuclease complex (86%)
4	4	$p < 2.9e-26$	Single-stranded DNA-binding replication protein A (50%) DNA repair protein (50%)
5	4	$p < 3.9e-13$	Small nuclear ribonucleoprotein(80%) snRNP component (20%)
5	4	$p < 3.9e-13$	Histone (40%) Histone transcription regulator (20%) Histone chaperone (20%)
3	3	$p < 1.4e-10$	Vacuolar sorting protein (33%) RNA polymerase II transcription factor complex subunit (33%) Uncharacterized conserved protein (33%)

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

$$\binom{n}{r} \sim \begin{cases} \frac{1}{\sqrt{2\pi r}} \frac{n^r}{r^r} e^{-r} & \text{if } r = o(\sqrt{n}) \\ \frac{1}{\sqrt{2\pi\alpha(1-\alpha)n}} 2^{nH(\alpha)} & \text{if } r = \alpha n \end{cases} \tag{13}$$

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 (12), we obtain

$$P(R_\rho \geq r) \leq \frac{r\sqrt{1-\rho}}{2\pi\sqrt{\rho}} \exp_2(-r^2 \log Q + r \log n - r \log r + r^2 H(\rho) - r \log e\}) \tag{14}$$

Defining $\kappa(p,\rho) = \log Q - H(\rho)$ as in Section 2, we find

$$P(R_\rho \geq r_0) \leq \frac{r_0\sqrt{1-\rho}}{2\pi\sqrt{\rho}} \exp_2(f(r_0)), \tag{15}$$

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,

$$P(R_\rho \geq r_0) \leq O\left(2^{-r_0}\right) = O\left(\frac{\log n}{n^{1/\kappa(p,\rho)}}\right). \tag{16}$$

This completes the proof for the upper-bound.

The lower-bound is not of a particular interest in terms of statistical significance, but we provide a sketch of the proof for completeness. By the second moment method [28], we have

$$P(R_\rho < r) \leq P(X_{r,\rho} = 0) \leq \frac{\mathbf{Var}[X_{r,\rho}]}{\mathbf{E}[X_{r,\rho}]^2} = \frac{1}{\mathbf{E}[X_{r,\rho}]} + \frac{\sum_{U_r \neq V_r} \mathbf{Cov}[X_\rho^{U_r}, X_\rho^{V_r}]}{\mathbf{E}[X_{r,\rho}]^2}, \tag{17}$$

where $X_\rho^{U_r}$ is the indicator r.v. for the subgraph induced by the vertex set U_r being ρ-dense. Letting $r = (1-\epsilon)\log n/\kappa(\rho)$, we observe that $\frac{1}{\mathbf{E}[X_{r,\rho}]} \to 0$ as $n \to \infty$. We split the sum $\sum_{U_r,V_r} \mathbf{Cov}[X_\rho^{U_r}, X_\rho^{V_r}] = g(r)+h(r)$, where $g(r)$ spans the set of node subsets U_r, V_r with intersection of cardinality at most $O(\rho r^2)$. Observe that when U_r overlaps with V_r on l vertices, then for $m = \rho r^2$

$$\mathbf{Cov}[X_\rho^{U_r}, X_\rho^{V_r}] = \sum_{k=\max\{0,l^2-r^2+m\}}^{\min\{l^2,m\}} \binom{l^2}{k} p^k q^{l^2-k} \left[\binom{r^2-l^2}{m-k} p^{m-k} q^{r^2-l^2-(m-k)}\right]^2.$$

Routine and crude calculations show that $g(r) \le \mathbf{E}[X_{r,\rho}]$, while $h(r) \le \alpha(r)\mathbf{E}[X_{r,\rho}]^2$ where $\alpha((1-\epsilon)\log n/\kappa(\rho)) \to 0$ as $n \to \infty$, which completes the proof. □

Proof of Theorem 2: Let $X_{r,\rho}^h$, $X_{r,\rho}^l$ be the number of ρ-dense subgraphs induced by only nodes in G_h or G_l, respectively. Let $X_{r,\rho}^b$ be the number of those induced by nodes from both sets. Clearly, $X_{r,\rho} = X_{r,\rho}^h + X_{r,\rho}^l + X_{r,\rho}^b$. The analysis for $G(n,p)$ directly applies for $\mathbf{E}[X_{r,\rho}^h]$ and $\mathbf{E}[X_{r,\rho}^l]$, hence we emphasize on $\mathbf{E}[X_{r,\rho}^b]$. Since $n_h = O(1)$, we have

$$\mathbf{E}[X_{r,\rho}^b] \le (1-\rho)r^2 \sum_{k=0}^{n_h} \binom{n_h}{k}\binom{n_l}{r-k} \sum_{l=0}^{2k(r-k)} \binom{2k(r-k)}{l}\binom{(r-k)^2}{\rho r^2 - l} p_b^l q_b^{2k(r-k)-l} p_l^{\rho r^2-l} q_l^{(r-k)^2-\rho r^2+l}, \tag{18}$$

where $q_b = 1-p_b$ and $q_l = 1-p_l$. Then,

$$\mathbf{E}[X_{r,\rho}^b] \le c(1-\rho)r^2 n_h \binom{n_l}{r} \sum_{l=0}^{2n_h r} \binom{2n_h r}{l}\binom{r^2}{\rho r^2 - l} p_b^l q_b^{2n_h r-l} p_l^{\rho r^2-l} q_l^{r^2-\rho r^2+l}, \tag{19}$$

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

$$\mathbf{E}[X_{r,\rho}^b] \le (1-\rho)r^2 \binom{n}{r}\binom{r^2}{\rho r^2} p_l^{\rho r^2} q_l^{r^2-\rho r^2} \sum_{l=0}^{2n_h r} \binom{2n_h r}{l} \left(\frac{p_b q_l}{p_l}\right)^l q_b^{2n_h r-l}. \tag{20}$$

Using $B = \frac{p_b q_l}{p_l} + q_b$ as defined in Theorem 2, we find $P(R_\rho > r) \le O(2^{f_1(r)})$, where $f_1(r) = -r(r\kappa(\rho) - \log n + \log r - \log e + 2n_h \log B)$.Hence,

$$P(R_\rho > r_1) \le O(2^{f_1(r_1)}) \le O\left(\frac{\log n}{n^{1/\kappa(p_l,\rho)}}\right) \tag{21}$$

for large n. □

5 Conclusion

In this paper, we make a first attempt to assess analytical statistical significance in PPI networks. Specifically, we emphasize on the notion of *dense* subgraphs, which is 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 reasonable means of assessing significance, models that mirror the topological characteristics of PPI networks should also be analyzed. This paper provides a stepping stone for the analysis of such complicated models.

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] 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.

[3] 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.

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

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

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

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

[8] 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.

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

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

[11] 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(Suppl. 1):i213–i221, 2005.

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

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

[14] M. Koyutürk, A. Grama, and W. Szpankowski. An efficient algorithm for detecting frequent subgraphs in biological networks. In *Bioinformatics Suppl. 12th Intl. Conf. Intel. Sys. Mol. Bio. (ISMB'04)*, pages i200–i207, 2004.

[15] M. Koyutürk, A. Grama, and W. Szpankowski. Pairwise local alignment of protein interaction networks guided by models of evolution. In *S. Miyano (Eds.): RECOMB 2005, LNCS*, volume 3500, pages 48–65, 2005.

[16] M. Koyutürk, Y. Kim, S. Subramaniam, W. Szpankowski, and A. Grama. Detecting conserved interaction patterns in biological networks. submitted.

[17] K.-P. Künkele, P. Juin, C. Pompa, F. E. Nargang, J.-P. Henry, W. Neuperr, R. Lill, , and M. Thieffry. The isolated complex of the translocase of the outer membrane of mitochondria. *J Biol Chem*, 273(47):31032–9, 1998.

[18] 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.

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

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

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

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

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

[24] 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.

[25] 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.

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

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

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

[29] 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.

[30] 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.

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

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

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

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

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

[36] 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.

[37] 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.