Journal of Bioinformatics and Computational Biology
Vol. 11, No. 3 (2013) 1341008 (13 pages)

DOI: 10.1142/S0219720013410084

Imperial College Press
www.icpress.co.uk

DISRUPTION OF PROTEIN COMPLEXES

GOLNAZ TAHERI[*,†,‖], MAHNAZ HABIBI[‡,**],
LIMSOON WONG[§,††] and CHANGIZ ESLAHCHI[¶,‡‡,§§]

[*]*School of Mathematics and Computer Sciences*
College of Science
University of Tehran, Tehran, Iran

[†]*Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*

[‡]*Department of Mathematics, Islamic Azad University of Qazvin, Iran*

[§]*School of Computing, National University of Singapore, Singapore*

[¶]*Department of Computer Science*
Shahid Beheshti University, G. C., Tehran, Iran
[‖]*golnazthr@ipm.com*
[**]*mhabibi@ipm.ir*
[††]*wongls@comp.nus.edu.sg*
[‡‡]*ch-eslahchi@sbu.ac.ir*

Received 12 October 2012
Revised 9 January 2013
Accepted 9 January 2013
Published 26 March 2013

Protein complexes are a cornerstone of many biological processes and, together, they form various types of molecular machinery that perform a vast array of biological functions. Different complexes perform different functions and, the same complex can perform very different functions that depend on a variety of factors. Thus disruption of protein complexes can be lethal to an organism. It is interesting to identify a minimal set of proteins whose removal would lead to a massive disruption of protein complexes and, to understand the biological properties of these proteins. A method is presented for identifying a minimum number of proteins from a given set of complexes so that a maximum number of these complexes are disrupted when these proteins are removed. The method is based on spectral bipartitioning. This method is applied to yeast protein complexes. The identified proteins participate in a large number of biological processes and functional modules. A large proportion of them are essential proteins. Moreover, removing these identified proteins causes a large number of the yeast protein complexes to break into two fragments of nearly equal size, which minimizes the chance of either fragment being functional. The method is also superior in these aspects to alternative methods based on proteins with high connection degree, proteins whose neighbors have high average degree, and proteins that connect to lots of proteins of high connection degree. Our spectral bipartitioning method is able to efficiently identify a biologically meaningful minimal set of proteins whose removal causes a massive disruption of protein complexes in an organism.

Keywords: Protein complexes; spectral bipartitioning; disruption.

[§§]Corresponding author.

1. Introduction

An essential protein is a protein whose removal from an organism is lethal to the organism. The identification of essential proteins typically requires experimental approaches that are time consuming and laborious. However, with advances in high-throughput technologies, a large number of protein−protein interactions are available. This presents new opportunities for detecting a protein's essentiality from the protein interaction network.[1] Nevertheless, the identification of essential proteins remains a challenging task. In this work, we consider a related problem of identifying a minimal set of proteins whose collective removal is lethal to the organism.

Our starting point is protein complexes. Recent studies have highlighted the importance of inter-connectivity in a large range of complexes and human disease-related systems. Network medicine has emerged as a new paradigm to deal with complex diseases. Connections between protein complexes and key diseases have been suggested for decades.[2] For example, Vanunu *et al.*[3] recently show the relationship between disease-causing genes and protein complex associations with the diseases of interest.

A protein complex is a group of two or more associated polypeptide chains. Protein complex formation sometimes serves to activate or inhibit one or more of the complex members. Individual proteins can participate in the formation of a variety of different protein complexes. Different complexes perform different functions and, the same complex can perform very different functions depending on a variety of factors. In short, protein complexes are central to many biological processes and, together, they form various types of molecular machinery that perform a vast array of biological functions. Clearly, a massive disruption to protein complexes in an organism would effectively disrupt the survival of the organism.

Hence, it interesting to identify a minimal set of proteins whose removal disrupts many complexes in an organism. We expect such a set of proteins to contain many essential proteins and to be involved in many biological processes and functional modules. In this work, we approach this problem by first constructing a network of an organism, where each node represents a protein in a complex and, two proteins are connected whenever they are in the same complex. We then present an algorithm based on spectral bipartitioning to find a minimum set of proteins whose removal disrupts the maximum number of complexes in the network. We consider the essentiality and other biological properties of the identified proteins. We also compare our proposed method to simple methods based on proteins with high connection degree, proteins whose neighbors have high average degree, and proteins that connect to lots of proteins of high connection degree. We show that the results of our algorithm have meaningful biological properties.

2. Problem Statement

The formation of a protein complex often serves to activate or inhibit one or more of the associated proteins. There are some proteins that participate in many protein

complexes; by removing them, their associated complexes and cell functions are likely to be disrupted. In this work, we are interested in identifying a minimum number of proteins whose removal potentially disrupts a maximum number of complexes in an organism.

Given a set of reference protein complexes $C = \{C_1, C_2, \ldots, C_n\}$. These protein complexes are collectively viewed as a weighted graph $G = \langle V, E, \omega \rangle$, where $V = \cup_{C_i \in C} C_i$, $E = \{uv \,|\, \exists C_i \in C : \{u, v\} \subseteq C_i\}$ and, the edge weight function $\omega : E \rightarrow Z$ is defined as $\omega(uv) = |\{C_i \in C \mid \{u, v\} \subseteq C_i\}|$ for each edge $uv \in E$. We assume that G is a connected graph. If it is not, we treat each connected component of the graph separately.

We define a subset S of V as a cut set of G. Let $E(S, \bar{S})$ denote the subset of edges with one vertex in S and another vertex in $\bar{S}$, where $\bar{S} = V - S$. The weight of a cut set S is defined as the sum of the weight of the edges in $E(S, \bar{S})$. The weighted max-cut problem in a weighted graph $G = \langle V, E, \omega \rangle$ is to find a cut set S with maximum weight. It is well known that the weighted max-cut decision problem is a NP-complete problem.[4–10]

Let the integer k be a given threshold and $\omega(uv) > k$. We refer to u and v as the source and sink of uv. The main goal of our work is to find a subset T of vertices of $G = \langle V, E, \omega \rangle$ whose removal splits G into two partitions, such that sources and sinks are in different partitions. We aim to remove the minimum number of vertices subject to the following two conditions. Firstly, we are interested in disrupting as many cell functions of the organism as possible. So we need to simultaneously maximize the number of protein complexes having at least one protein in the cut set. Secondly, if the difference between the sizes of the two partitions is high, the probability of a whole undisrupted protein complex existing in one partition is increased. Therefore, we also aim to partition the graph into two balanced partitions.

The problem to be addressed in this work is thus to pick a subset of nodes T from a weighted connected graph $G = \langle V, E, \omega \rangle$ such that (i) T is as small as possible; (ii) the removal of T partitions G into two disjoint subgraphs G_1 and G_2; (iii) the weight of the cut is maximized; (iv) the ratio $|G_1|/|G_2|$ is as close to 1 as possible, where the size of a graph is measured by the number of vertices it has.

3. Method

3.1. *Algorithm*

Our proposed problem is a version of the balanced graph partitioning problem, which is known to be NP-complete.[5] Therefore, we approximate the balanced bipartitioning with spectral bipartitioning. This spectral method recursively bisects a graph by considering the eigenvectors of the Laplacian matrix of the graph. Spectral partitioning is a very successful heuristic approach for graph partitioning.[11] We have also earlier applied it to solve the problem of finding a minimal set of proteins whose removal potentially disrupts the maximum number of pathways in an organism.[12]

We proceed as follows. Given a weighted graph $G = \langle V, E, \omega \rangle$, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of vertices in G; $e_{ij} \in E$ is the edge between vertices v_i and v_j in V; and ω_{ij} is the weight of edge e_{ij}. Let $A = [a_{ij}]$ be the adjacency matrix of G where

$$a_{ij} = \begin{cases} \omega_{ij} & \text{if } (v_i, v_j) \in E \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

We define the diagonal matrix of G as $DI = \text{diag}(d_i)$, where $d_i = \sum_{e_{ij} \in E} \omega_{ij}$. Now, the Laplacian matrix of the graph G is defined[13] by $L(G) = DI - A$.

The Laplacian matrix of a connected graph has the following desirable property. Let $u_1, u_2, \ldots, u_n$ be the normalized eigenvectors of the Laplacian matrix $L(G)$ of the graph G. Suppose $u_1, u_2, \ldots, u_n$ are sorted according to the value of their respective second coordinate. Let $\hat{u}$ be the median of $u_1, u_2, \ldots, u_n$; that is, $\hat{u}$ is the middle point in this sort order. Let S be the set of vertices of G corresponding to those u_i coming before $\hat{u}$. Let $\bar{S}$ be the set of vertices of G coming after $\hat{u}$. Then, according to Spielman,[11] this partition is a good approximation of the best max cut.

The procedure above splits the vertices of the graph G into two partitions S and $\bar{S}$ of roughly equal size. This induces a set $E(S, \bar{S})$ of cut edges that cross the two partitions. Next, using $E(S, \bar{S})$, we find the set T of cut vertices such that (i) T is as small as possible, (ii) the removal of T partitions G into two disjoint subgraphs G_1 and G_2, (iii) the weight of the cut is maximized, and (iv) the ratio $|G_1|/|G_2|$ is as close to 1 as possible. To do this, we form a bipartite graph H with $V(H) = V(G)$ and $E(H) = E(S, \bar{S})$.

Now, we sort the vertices in S and $\bar{S}$ according to weighted degree (d_i) of these vertices in the graph G. Then we select the vertex with the highest degree in S, say s_i, and remove s_i and all edges connected to it. We choose another vertex with the highest degree in $\overline{S}$, say $\overline{s_i}$, and remove it and all edges connected to it. This procedure is iterated until all edges in $E(S, \bar{S})$ are removed.

By this algorithm we try to find a minimum vertex cut set T with maximum weight of edges between G_1 and G_2 where $G - T$ splits into two balanced partitions G_1 and G_2. It holds because this Laplacian procedure is an approximation to separate vertices, with highest edge weight, into two partitions.[11] In this way, we try to disrupt maximum number of complexes with removing minimum number of genes.

3.2. *Data*

Many protein complexes and protein−protein interactions have been collected, especially for the model organism *Saccharomyces cerevisiae* (bakers yeast). So we test our ideas on yeast data.

Our yeast protein complex data set was obtained from the Munich Information Center for Protein Sequences (MIPS)[14] in September 2009. This data set contains 1,142 complexes. There are 2,752 proteins in these complexes. The number of complexes of size 1 is 79 (these are complexes containing multiple instances of the same

protein); the number of complexes of size 2 is 229; the number of complexes of size 3 is 177; the number of complexes of size 4 is 139; and the remaining 518 complexes are size greater than 4. In this work, we only consider complexes of size greater than 2.

We also use the collection of protein−protein interaction obtained from BioGRID.[15] This data collection includes interactions obtained by several techniques. We only consider interactions derived from mass spectrometry and two-hybrid experiments as these represent physical interactions and co-complexed proteins.

4. Results

4.1. *Performance evaluation measures*

To validate our algorithm, we define two measures. The first measure is the number of complexes which have an intersection with the cut vertices. We say that a complex, x, is disrupted if it has a vertex in the cut set. Another measure is c_x, defined as follows:

$$c_x = \frac{\max\{|x \cap G_1|, |x \cap G_2|\}}{|x|}. \tag{2}$$

Here, G_1 and G_2 are the two partitions resulting from our proposed algorithm, and $|x|$ is the size of a complex x. c_x is a number between 0 and 1. If the difference between the sizes of the complex partitions is high, the probability of the existence of an undistrupted or functioning complex in one partition is increased. Therefore, we try to partition each complex into two halves of roughly equal size, such that each half resides in one of the two partitions G_1 and G_2. For each complex x, the best situation occurs when

$$0 < c_x \leq \frac{\lceil \frac{|x|}{2} \rceil}{|x|}.$$

We should notice that, for complexes of small size, if we were to remove just one vertex and put this node in the other partition or in the cut set, the complex would be disrupted. So we consider complexes of small size separately. For a complex $x, c_x \leq 0.5$ is the ideal value. Now, for a complex of size 3, if we separate just one vertex from two other vertices, we already achieve $c_x = 0.6$, which is the ideal value. For complexes of size 4, 0.25, 0.5, 0.75, and 1 are possible values for c_x. For a complex of size 4, the next value after the ideal value is 0.75, and this number occurs when just one vertices is separated from the other 3 vertices. In this case, we consider the complex to be disrupted. For a complex of size greater than 4, we define a confidence interval:

$$I_\epsilon = \left[0, \frac{\lceil \frac{|x|}{2} \rceil}{|x|} + \epsilon\right]. \tag{3}$$

We say that a complex x, is disrupted if $c_x \in I_\epsilon$. By increasing ϵ, our confidence of disrupting the function of x is decreased. So we call this complex is ϵ-departed.

4.2. *Evaluation with respect to disruption of complexes*

Let T define the cut set which is obtained from the algorithm for complexes in MIPS. As described earlier, spectral bipartitioning splits the graph G into two partitions S and $\bar{S}$ of roughly equal size, inducing a set of cut edges $E(S, \bar{S})$. We need to pick a minimum number of vertices T to remove so that all the cut edges in $E(S, \bar{S})$ are destroyed. This is equivalent to the set cover problem, which is NP complete.[16] Our algorithm does this by removing a minimum number of vertices in descending order of their degree, alternating between S and $\bar{S}$. This greedy heuristic is used in our algorithm because such a greedy heuristic has been shown[17] to be essentially the best possible polynomial time approximation algorithm for the set cover problem. The sum of weight of the edges of the corresponding graph G is 33,636. The size of the vertex cut is 50 with $|E(S, \overline{S})| = 2{,}689$ and $\sum_{e \in E(S,\overline{S})} w(e) = 9{,}496$. By removing the cut set from the graph, we obtain two parts G_1 and G_2, each containing 1,278 vertices.

We first evaluate our cut set with respect to its ability to disrupt the complexes. Note that we do not consider complexes of size 1 and 2. The number of complexes in MIPS of size at least 3 is 834. The number of such complexes having at least one vertex in the cut set is 343. We have 177 complexes of size 3 and, only 25 of these complexes have some intersection with the cut set. We find that 8 of these complexes have no intersection with other complexes and 73 of them have only one intersection with other complexes. It is not surprising that these 81 complexes have no intersection with the cut set. If we want disrupt these 81 complexes we must select at least one of their vertices and, this increases the size of the cut set. In this work, we disrupt only 25 complexes of size 3. Therefore, the number of complexes of size 3 which have no intersection with the cut set is 152. We also find that 318 complexes of size at least 4 have some intersection with the cut set.

There are many complexes of size 3 which are not disrupted yet. So we construct a new graph on complexes of size 3 which are not visited by any element of T. We run our algorithm on this new graph, H. This gives us a new cut, T^*, of size 30. $H - T^*$ has two parts A_1 and B_1, and each of them has 185 vertices.

Now we define a new cut set, $\tilde{T}$, as the union of T and T^*. We find that by removing $\tilde{T}$ from the graph G, the graph is separated into two nearly equal partitions A_2 and B_2 with 1,381 and 1,384 proteins respectively. The number of complexes disrupted by $\tilde{T}$ is 461. Only 65 of them is of size 3, 49 of size 4, 24 of them of size 5, 24 of them size 6 and 299 of them of size greater than 6.

The cut set $\tilde{T}$ has size 80. How good is this cut set in terms of maximizing the number of disrupted complexes? Let us first compare it to the obvious alternative of choosing 80 vertices with the highest connection degree in the graph. We denote the list of the 80 vertices with the highest connection degrees by $\underline{T}$. The degree distribution of these vertices are shown in Fig. 1. The maximum degree in this list is 584 and the minimum degree is 175. The number of complexes visited by $\underline{T}$ is 311, which is much fewer than the number (461) of complexes visited by $\tilde{T}$. We also find that only 21 proteins from $\underline{T}$ are in $\tilde{T}$. Figure 2 shows the distribution degree of vertices

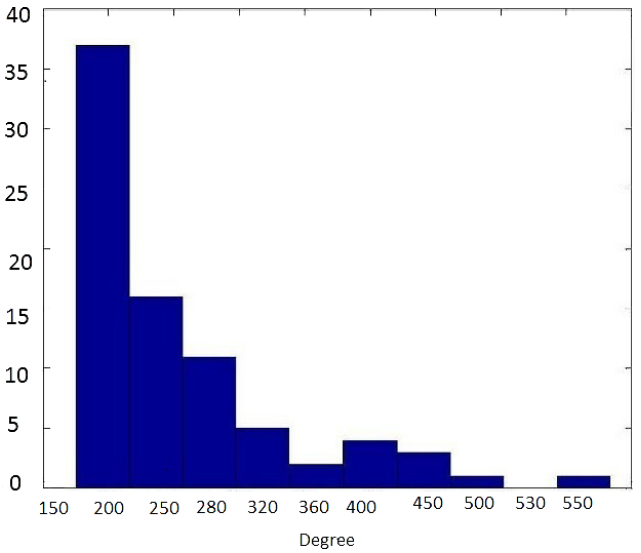

Fig. 1. Degree distribution of first 80 vertices with maximum degree.

in $\tilde{T}$. These figures show that choosing vertices with the highest connection degrees is not a good approach for disturbing the maximum number of complexes. Our algorithm makes a better selection.

We also consider three other scenarios to further evaluate our algorithm. In the first scenario, we try to select proteins with high connection degree (hub) in G as a

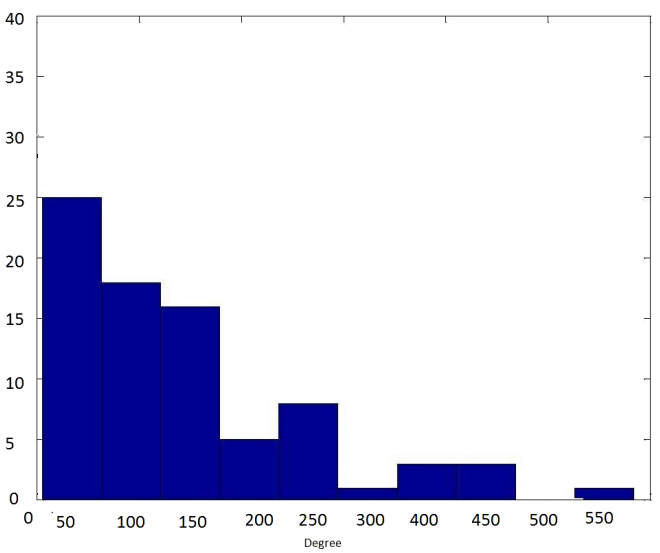

Fig. 2. Degree distribution of vertices in $\tilde{T}$.

cut set ($\underline{T}$). So we select hubs as a cut vertex. In this way, we sort hubs decreasingly, then we start to delete these vertices until our network separate into two parts. In this way, we find $\underline{T}$ with 13 proteins and two partitions with 11 and 2,582 proteins. We find that $\underline{T}$ is smaller than our cut set T, but the new partitions have a high difference in size. In this work, we try to separate our network into two balanced partitions to disrupt the maximum number of complexes. So we continue removing hub proteins until we have a cut set with size 50. In this case, the number of components of the resulting graph is 12. One of size 2,549, six of size 6, two of size 4, and the sizes of the remaining components are 2, 5, 11, respectively. Therefore, removing hub proteins is not a good approach.

In the second scenario, we select proteins whose neighbors have high average connection degree as a cut set $CUT2$. In this way, we find a cut set of size 49 and, two partitions with 2 and 2,555 proteins respectively. $CUT2$ has approximately the same size as T, but the partitions have a high difference in size.

In the third scenario, we select proteins that connect to lots of proteins of high connection degrees as a cut set $CUT3$. In this way, we find a cut set having 409 proteins and two partitions with 1 and 2,196 proteins. This cut set is much bigger than T and, the partitions have a great difference in size.

It is obvious that our cut set T is better than other cut sets.

We also compute a p-value test for the number of essential proteins of our cut set T. Let m be the cardinality of the union of the set of proteins in all complexes in our graph and the set of essential proteins (3,728), let a be the cardinality of our cut set (80), let b be the cardinality of the set of essential proteins (1,168) and h_0 be the cardinality of set of proteins in T that are essential (46). Now the exact probablity of getting an intersection of size greater than h_0 between cut set and the set of essential proteins due to chance;

$$\sum_{h \geq h_0} \frac{\binom{m}{h}\binom{m-h}{a-h}\binom{m-a}{b-h}}{\binom{m}{a}\binom{m}{b}}.$$

This p-value for essential genes in our cut set is $9.06 * 10^{-7}$, which shows that our cut set cannot be explained by chance alone.

4.3. *Evaluation with respect to biological properties*

To obtain the biological properties of the cut set proteins, first we consider essentiality. Essential genes are indispensable for the survival of an organism. Therefore, the essential genes are potentially corelated with massive disruption of protein complexes. Table 1 shows that 30 out of 50 proteins of the set T are essential. The essentiality information of each gene was retrieved from these two sites:

- www-sequence.stanford.edu/group/yeast_deletion_project/Essential_ORFs.txt
- http://bioinfo.mbb.yale.edu/genome/yeast

Table 1. Cut genes and their essentiality in network.

Gene Name	E.	Gene Name	E.	Gene Name	E.	Gene Name	E.	Gene Name	E.
YGL120c	E	YKR081c	E	YKR026c	N	YGL049c	E	YGR159c	N
YGR090w	E	YBR154c	E	YCR002c	N	YGL195w	N	YLR357w	N
YKL104c	E	YPL204w	E	YOR116c	E	YLR438w	N	YDR188w	E
YOL094c	E	YNR035c	E	YNL330c	N	YNL250w	N	YKL014c	E
YJL138c	N	YGR240c	N	YGR155w	N	YGL011c	E	YPR175w	E
YHR027c	N	YOL038w	E	YGR135w	N	YMR229c	E	YNL139c	N
YPL043w	E	YIL033c	N	YER025w	E	YDL014w	E	YPR110c	E
YOL041c	N	YMR146c	E	YLR216c	E	YNL189w	E	YDR429c	E
YPL004c	N	YOR341w	E	YOR151c	E	YHR099w	N	YPL237w	E
YJL074c	E	YHL030w	N	YBL004w	E	YOL139c	E	YKL085w	N

In this table, the essentiality is denoted by E and non-essentiality is denoted by N.

Table 2 shows that 16 out of the 30 proteins of the set T^* are essential gens. Thus, 46 out of the 80 proteins in the cut set $\tilde{T}$ are essential. Thus the algorithm selected biologically meaningful proteins.

Next, we consider biological processes and molecular functions of these 50 and 80 cut set proteins and the associated complexes. We see that the 50 proteins participate in 117 different biological process and 75 different functional modules. The numbers for the cut set with size 80 are 195 and 121 respectively.

In contrast, the corresponding numbers for the 50 vertices with the highest connection degree are 69 biological processes and 53 different functional modules. Similarly, for the 80 vertices with the highest connection degree, these numbers are 90 and 55 respectively. This is another advantage of our cut set.

Now we pick 50 random proteins. Let x be the number of different biological processes that they participate it. We repeat this 10,000 times and calculate the number of x's that are greater than or equal to 117. The p-value is 0.07. We do this for functional modules and this p-value for functional modules is 0.062.

We also do this for cut set of size 80 and the p-value for biological process and functional modules are 0.061 and 0.058 respectively.

Table 2. Cut genes and their essentiality in constructed graph on complexes of size 3 (H).

Gene Name	E.	Gene Name	E.	Gene Name	E.	Gene Name	E.	Gene Name	E.
YBR009c	N	YDR473c	E	YJR093c	E	YMR049c	E	YOR310c	E
YBR084w	N	YEL060c	N	YKL139w	N	YMR246w	N	YOR370c	E
YBR247c	E	YER133w	E	YLL039c	E	YNL085w	N	YPL129w	N
YDL192w	N	YGR103w	E	YLR347c	E	YNL103w	E	YPL235w	E
YDR224c	E	YGR186w	E	YML057w	N	YOL004w	N	YPR010c	E
YDR343c	N	YHR052w	N	YMR012w	N	YOL086c	N	YPR178w	E

In this table, the essentiality is denoted by E and non-essentiality is denoted by N.

Table 3. Percentage of complexes with size of at least 5 which departed with respect to the given ϵ.

ϵ	0.05	0.1	0.15	0.2	0.25
Presence of complexes in I_ϵ	36%	46%	63%	73%	81%

4.4. *Evaluation with respect to protein interactions*

Now we validate our algorithm using the BioGrid dataset. The average degree of vertices in BioGrid is 10.93, but the average of degree of vertices in $\tilde{T}$ is 45.76 and the average of degree of vertices in T is 44.42. BioGrid has 5,040 vertices and 27,557 edges. If the degree distribution of vertices in BioGrid is a normal distribution, we can expect that two random sets of 50 and 80 vertices in BioGrid to visit nearly 275 and 440 edges respectively. But the number of edges visited by the vertices in our cut set of size 80 is 3,499 and in our cut set of size 50 is 2,091 respectively. This is much higher than the two simple averages above. To test whether this is significant, we select 10,000 random subsets of size 80 (and 50) from the BioGrid protein interaction network and check the number of edges visited. None of these randomly chosen subsets visit 3,499 (and 2,091) edges or more. Also, the average number of edges in visited by the subsets of size 80 and 50 are 861 and 540 respectively, which are also significantly lower than the number of edges visited by our two cut sets.

4.5. *Evaluation with respect to tolerence I_ϵ*

In Table 3, the number of the departed complexes with different ϵ is shown. It is obvious that by increasing ϵ the number of departed complexes increased. With a cut set of size 80 the number of complexes of size 3 and 4 which departed are 65 and 124, respectively. With size at least 5 for $\epsilon = 0.05$, the percentage of complexes disrupted is 36%. In fact, for $\epsilon = 0.05$, only 347 complexes from 518 complexes are undisrputed. But we know by increasing ϵ, the number of undisrupted complexes decreases. For $\epsilon = 0.25$, we have only 97 from 518 complexes of size at least 5 not departed.

5. Discussion

The identification of a minimal set of proteins whose removal would likely cause a wide-spread disruption of protein complexes in an organism, leading to lethality of the organism, is studied here. This is a NP-complete problem, which we approximate using an approach based on spectral bipartitioning. We have applied the proposed approach on yeast protein complexes. This has resulted in a set $\tilde{T}$ of 80 proteins. These 80 proteins are found in 461 out of the 834 protein complexes considered. Thus 55% (= 461/834) of the protein complexes are potentially disrupted when these 80 proteins are removed or silenced. This compares extremely well with several alternative approaches for choosing proteins to remove, such as removing proteins with high connection degree, proteins whose neighbors have high average connection

degree, and proteins having many neighbors of high connection degree. Removing 80 proteins using these alternative strategies hits no more than 37% (= 311/834) of the protein complexes.

However, a deeper analysis is needed than this simple counting of how many complexes are hit by the removal of some proteins. Protein complexes actually come in families of isoforms.[18] Each family contains protein complexes (the isoforms) having a common set of core proteins (which are unique to the family) but with different attachment proteins (which may appear in multiple families). While the removal of some attachment proteins may disrupt a subset of the isoforms, other undisturbed isoforms having similar functions may still be formed. This may permit the organism to continue surviving in some situations. In order to ensure the disruption of an entire family of protein complexes, it is necessary to break the complexes up at the level of their core proteins. It is easy to see that this is more likely to happen when a protein complex is broken into two parts of roughly equal size than when only a small bit is broken off a protein complex.

As shown in Table 3, 36% (and up to 81%) of large protein complexes are fragmented by the removal of the 80 proteins identified by our approach into two halves, where each half is no larger than half (and up to three quarters) of the original protein complex it comes from. Moreover, the network is partitioned by the removal of these 80 proteins into two halves of nearly equal size (1,381 and 1,384 proteins). In contrast, all three alternative approaches split the network into highly unbalanced partitions, with fewer than a dozen proteins in one partition and more than two thousand five hundred proteins in the other partition. Clearly, they are breaking a protein complex by splitting off only a small number of high-connectivity proteins. High-connectivity proteins are more likely to be attachment proteins than core proteins.[19,20] This suggests removal of proteins identified using the alternative approaches are less likely to break the protein complexes at the core level and, thus, less likely to disrupt entire families of complexes. It is interesting that selecting high-connectivity proteins to remove does not seem to fragment protein complexes in a significant way.

We have also analyzed the biological properties of the 80 proteins identified by our spectral bipartitioning approach. These 80 proteins are involved in 195 different biological processes and 121 functional modules. This is more than two times higher than those proteins identified by the alternative approaches (90 biological processes and 55 functional modules). So, the removal of our 80 proteins are likely to cause much more wide-ranging disruption. This is also supported by that fact that 58% (= 46/80) of our 80 proteins are essential proteins. However, only 13 of these essential proteins are among the 49 essential proteins in the 80 proteins of the highest connection degree. So 33 (= 46 − 13) essential proteins in our cut set are not in 80 proteins with highest connection degree.

Overall, it seems our spectral bipartitioning approach has selected biologically meaningful proteins for removal, to massively disrupt protein complexes in an organism.

Acknowledgments

This work was supported in part by a Singapore Ministry of Education Tier-2 grant MOE2009-T2-2-004 (for Wong) and by Shahid Beheshti University (for Eslahchi).

References

1. Li M, Zhang H, Wang JX, Pan Y, A new essential protein discovery method based on the integration of protein−protein interaction and gene expression data, *BMC Syst Biol* **6**:15, 2012.
2. Nacher JC, Schwartz JM, Modularity in protein complex and drug interactions reveals new polypharmacological properties, *PLoS One* **7**(1):e30028, 2012.
3. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R, Associating genes and protein complexes with disease via network propagation, *PLoS Comput Biol* **6**(1):e1000641, 2010.
4. Karp RM, Reducibility among combinatorial problems, in Miller RE, Thatcher JW (eds.), *Complexity of Computer Computations*, Plenum, pp. 85−103, 1972.
5. Garey MR, Johnson DS, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, CA, 1979.
6. Donath WE, Hoffman AJ, Algorithms for partitioning of graphs and computer logic based on eigenvectors of connection matrices, *IBM Technical Disclosure Bulletin* **15**:938−944, 1972.
7. Donath WE, Hoffman AJ, Lower bounds for the partitioning of graphs, *J Res Development* **17**:420−425, 1973.
8. Boppana R, Eigenvalues and graph bisection: An average case analysis, *Proc 28th IEEE Symp Foundations of Computer Science*, pp. 280−285, 1987.
9. Powers D, Graph partitioning by eigenvectors, *Linear Algebra and its Applications* **101**:121−133, 1988.
10. Parrish JR, Yu J, Liu G *et al.*, A proteome-wide protein interaction map for *Campylobacter jejuni*. *Genome Biol* **8**:R130, 2007.
11. Spielman DA, Spectral graph theory and its applications, *Proc 48th Annual IEEE Symp Foundations of Computer Science*, pp. 29−38, 2007.
12. Ayati M, Taheri G, Arab S, Wong L, Eslahchi C, Overcoming drug resistance by co-targeting, *Proc 4th IEEE Int Conf Bioinformatics and Biomedicine*, pp. 198−201, 2010.
13. Mohar B, The Laplacian spectrum of graphs, in Alavi Y, Chartrand G, Oellermann OR, Schwenk AJ (eds.), *Graph Theory, Combinatorics, and Applications*, Wiley, Vol. 2, pp. 871−898, 1991.
14. Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S, Frishman D, MIPS, a database for genomes and protein sequences, *Nucleic Acids Res* **27**(1):44−48, 1999.
15. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M, BioGRID: A general repository for interaction datasets, *Nucleic Acids Res* **34**:D535−D539, 2006.
16. Richard M Karp, Reducibility among combinatorial problems, in Miller RE, Thatcher JW (eds.), *Complexity of Computer Computations*, Plenum, New York, pp. 85103, 1972.
17. Carsten L, Mihalis Y, On the hardness of approximating minimization problems, *J ACM* **41**(5):960981, 1994.
18. Gavin AC, Aloy P, Grandi P *et al.*, Proteome survey reveals modularity of the yeast cell machinery, *Nature* **440**(7084):631−636, 2006.
19. Liu G, Yong CH, Chua HN, Wong L, Decomposing PPI networks for complex discovery, *Proteome Sci* **9**(Suppl 1):S15, 2011.
20. Lee SH, Kim PJ, Jeong H, Global organization of protein complexome in the yeast *Saccharomyces cerevisiae*. *BMC Systems Biol* **5**:126, 2011.

Golnaz Taheri received her Master degree in Computer Science at University of Tehran. She received her Bachelor degree in Computer Science at Sharif University of Technology. She works as a researcher in Institute for Studies in Theoretical Physics and Mathematics (IPM). Her research interests include issues related to Bioinformatics especially protein−protein interaction network.

Mahnaz Habibi studied mathematics at the Shahid Beheshti University in Iran. She was awarded MSc degree in December 2005. Four years later, she received her PhD in Mathematics from the Shahid Beheshti University in the supervision of Dr. Ch. Eslahchi. During her education, she worked as a researcher in Institute for Studies in Theoretical Physics and Mathematics (IPM). Her main research interest is Bioinformatics and she works on protein−protein interaction network, protein structure, and phylogenetic networks. She works as the team leader of the bioinformatics research group at Gazvin Azad University in Iran from 2011.

Limsoon Wong is a provost's chair Professor of computer science and a professor of pathology at the National University of Singapore. He currently works mostly on knowledge discovery technologies and their application to biomedicine. He serves/d on the editorial boards of *Information Systems, Journal of Bioinformatics and Computational Biology, Bioinformatics, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Drug Discovery Today, and Journal of Biomedical Semantics* (*BMC*). He co-founded and is chairman of Molecular Connections, a provider of data curation services employing over 700 curators, analysts, and engineers. He received his BSc (Eng) in 1988 from Imperial College London and his PhD in 1994 from University of Pennsylvania.

Changiz Eslahchi graduated from the faculty of mathematical sciences of Teacher Training University of Tehran with BSc degree in 1987. He received his MSc degree in mathematics in 1989 from the University of Shiraz and PhD degree in 1998 from Sharif University of Technology. He is presently working as an Associate Professor at Shahid Beheshti University (Iran). His research interests are focused on Bioinformatics especially protein−protein interaction network and phylogenetic tree.