Protein Bipartivity and Essentiality in the Yeast Protein–Protein Interaction Network

Ernesto Estrada*

Complex System Research Group, X-rays Unit, RIAIDT, University of Santiago de Compostela, Edificio CACTUS, Santiago de Compostela 15782, Spain

Received March 21, 2006

Protein–protein interaction networks (PINs) are structured by means of a few highly connected proteins linked to a large number of less-connected ones. Essential proteins have been found to be more abundant among these highly connected proteins. Here we demonstrate that the likelihood that removal of a protein in a PIN will prove lethal to yeast correlates with the lack of bipartivity of the protein. A protein is bipartite if it can be partitioned in such a way that there are two groups of proteins with intergroup, but not intragroup, interactions. The abundance of essential proteins found among the least bipartite proteins clearly exceeds that found among the most connected ones. For instance, among the top 50 proteins ranked by their lack of bipartivity 62% are essential proteins. However, this percentage is only 38% for proteins ranked according to their number of interactions. Protein bipartivity also surpasses another 5 measures of protein centrality in yeast PIN in identifying essential proteins and doubles the number of essential proteins selected at random. We propose a possible mechanism for the evolution of essential proteins in yeast PIN based on the duplication-divergence scheme. We conclude that a replica protein evolving from a nonbipartite target will also be nonbipartite with high probability. Consequently, these new replicas evolving from nonbipartite (essential) targets will with high probability be essential.

Keywords: essential proteins • protein interactions • protein networks • yeast • centrality measures • graph theory

I. Introduction

Protein–protein interaction networks (PINs) are complex systems which can be represented by means of networks, in which the nodes represent proteins and links, protein–protein interactions. The building of these networks is made possible by the availability of protein interaction data obtained from direct large-scale experiments, such as yeast two-hybrid technology and mass spectrometry. It is known that protein–protein interactions play a prominent role in most biological processes, and in recent years, there has been an intensive study of the properties of PINs, including the study of functional modules, protein function, domain–domain interactions, protein evolution, and protein essentiality as well as the investigation of global network topologies, such as “small-worldness” and “scale-freeness”. The discovery of the scale-free properties of PINs in particular has attracted much attention. This property is intimately related to the statistical distribution of the number of interactions per protein. In a scale-free PIN, there are very few proteins exhibiting a large number of interactions, the so-called hubs; most of the proteins display a low connectivity. One of the most important consequences of this topological characteristic of PINs is that a few highly connected proteins guarantee the functional and structural integrity of the network. The last can be exemplified by the fact that the removal of such highly connected nodes makes the network collapse into several isolated clusters. It has been observed that a correlation exists between the connectivity and essentiality of a given protein. A protein is considered to be essential if a knock-out results in (conditional) lethality or infertility. On the contrary, nonessential proteins are those for which knock-outs yield viable and fertile individuals.

The identification in silico of essential proteins in a PIN is important from both a theoretical and practical point of view. On one hand, the correlation observed between connectivity and essentiality has triggered a debate about the emergence of correlations between the number of interactions of a protein and its evolutionary conservation. On the other hand, proteins essential to the viability of a pathogen would make the most obvious and experimentally feasible targets for drug design. It is clear that if a protein is essential for a cellular process in a pathogenic microorganism then a drug that blocks its activity will cause cessation of growth or death of the cell. However, a biological function cannot easily be assigned to a single protein but is more likely to emerge from the interaction of many proteins in the PIN. We have previously demonstrated that, in correlating with protein essentiality, the implication of a protein in a large number of
network substructures is significantly more important than the number of interactions.36

Network substructures include single cycles, such as triangles, quadrilaterals, and pentagons, as well as fully connected clusters. The protein modules formed by these substructures can be of two different topological forms based on their

II. Methods

Protein Bipartivity. A protein in a PIN is bipartite if it does not take part in any closed-walk (CW) of odd length.36 A walk of length $k$ in a PIN is a sequence of (not necessarily different) proteins $v_1, v_2, \ldots, v_{k+1}$ such that for each $i = 1, 2, \ldots, k$ there is an interaction (link) between $v_i$ and $v_{i+1}$, A closed walk (CW) is a walk in which $v_{k+1} = v_1$ and protein bipartivity is measured as the proportion of even to total number of CW in the PIN. Let $G = (V,E)$ be a PIN with $N$ proteins and having eigenvalues of the adjacency matrix $\alpha = (\lambda_1, \ldots, \lambda_N)$. It is known that the total number of CWs in which protein $i$ takes part is given by the following expression:39

$$\text{SC}(i) = \sum_{j=1}^{N} |v_j(i)|^2 e^{\lambda_j}$$

where $(v_1, v_2, \ldots, v_N)$ is an orthonormal basis of $R^N$ composed by eigenvectors of the adjacency matrix associated with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$, and $v_j(i)$ is the $ith$ component of $v_j$. The total number of even CWs in which protein $i$ participates is given by the following expression:

$$\text{SC}_{\text{even}}(i) = \sum_{j=1}^{N} |v_j(i)|^2 \cosh(\lambda_j)$$

Then the protein $i$ bipartivity, $\beta(i)$, is given by39

$$\beta(i) = \frac{\sum_{j=1}^{N} |v_j(i)|^2 \cosh(\lambda_j)}{\sum_{j=1}^{N} |v_j(i)|^2 e^{\lambda_j}}$$

When a protein does not take part in any odd-length CW, i.e., $\text{SC}_\text{odd}(i) = 0$, and $\beta(i) = 1$, this means that this protein is bipartite. Protein bipartivity can take values between 1/2 and 1, $1/2 < \beta(i) \leq 1$, where the lower bound is reached when all proteins interacting with $i$ interact with each other, a fully connected motif or clique. The exact value of 1/2 is reached in the limit when the number of proteins in such cluster tends to infinity, i.e., $\infty$-clique. In general, the effect of “adding” a new interaction into a bipartite cluster of interacting proteins affects protein bipartivity in different degrees.38

Protein Centrality Measures. The other protein centrality measures used in this work include degree (DC), closeness (CC), betweenness (BC), eigenvector (EC), Information (IC), and odd subgraph (SC) centralities. DC$(i)$ is the number of proteins that are interacting with protein $i$. CC$(i)$ is the sum of graph-theoretic distances from all other proteins, with respect to protein $i$, in the PIN, where the distance $d(i,j)$ from one protein $i$ to another $j$ is defined as the number of links in the shortest path from one to the other. The closeness centrality of protein $i$ in a PIN is given by the following expression:40

$$\text{CC}(i) = \frac{N - 1}{\sum_j d(i,j)}$$

BC characterizes the degree of influence a protein has in “communicating” between protein pairs and is defined as the fraction of shortest paths going through a given node. If $\rho(i,j)$ is the number of shortest paths from protein $i$ to protein $j$, and $\rho(i,k,j)$ is the number of these shortest paths that pass through protein $k$ in the PIN, then the betweenness centrality of node $k$ is given by41

$$\text{BC}(k) = \sum_j \frac{\rho(i,k,j)}{\rho(i,j)} i \neq j \neq k$$

EC is defined as the principal eigenvector of the adjacency matrix $A$ defining the network. It simulates a mechanism in which each node affects all of its neighbors simultaneously. The defining equation of an eigenvector is $\lambda e = Ae$, where $A$ is the adjacency matrix of the graph, $\lambda$ is an eigenvalue, and $e$ is the eigenvector. Thus, $A$ of protein $i$ is defined as the $ith$ component of the eigenvector $e_i$, $e_i(i)$, that corresponds to the largest eigenvalue of $A$, $\lambda_1$ (principal eigenvalue or index)42,43

$$\text{EC}(i) = e_i(i)$$

Information centrality (IC) is based on the information that can be transmitted between any two points in a connected network. In the case of PINs, the transmission of information can be understood as the process in which a protein interacts with another and changes it, by a conformational or chemical transformation.44 If $A$ is the adjacency matrix of a network, $D$ a diagonal matrix of the degree of each node, and $J$ a matrix with all its elements equal to one, then IC is defined by inverting the matrix $B$ defined as $B = D - A + J$ in order to obtain the matrix $C = (c_{ij}) = B^{-1}$ from which the information matrix is obtained as follows:45

$$I_{ij} = (c_{ij} + c_{ji} - c_{ij})^{-1}$$

The information centrality of the protein $i$ is then defined by using the harmonic average

$$\text{IC}(i) = \frac{1}{N_i} \sum_{j} \frac{1}{I_{ij}}$$

Stephenson and Zelen proposed defining $I_k$ as infinite for computational purposes, which makes $1/I_k = 0$.45 The odd subgraph centrality, SC$_\text{odd}(i)$, is defined by counting the number
of odd-length CW in the PIN as was previously carried out for the CWs of even-length\(^{39}\)

\[
SC_{\text{odd}}(i) = \sum_{j=1}^{N} |v_j(i)|^2 \sinh(\lambda_j)
\]

\text{(9)}

**PIN Dataset.** The protein—protein interaction network (PIN) of yeast *Saccharomyces cerevisiae* was compiled by Bu et al.\(^{46}\) The protein interaction network is viewed as a set of nodes representing proteins and a set of links, which join together pairs of proteins that physically interact as determined experimentally by any of the technologies used with this end, mainly the yeast two-hybrid system and mass spectrometry identification of proteins that co-affinity purify with a bait protein.\(^{3}\) The original data was obtained by von Mering et al. by assessing a total of 80,000 previously reported interactions among 5400 proteins and assigning each interaction a confidence level.\(^{47}\) Bu et al. focused on 11,855 interactions with high and medium confidence between 2617 proteins in order to reduce the interference of false positives.\(^{45}\) They reported a network of 2361 nodes and 6646 links (http://vlado.fmf.uni-lj.si/pub/networks/data/bio/Yeast/Yeast.htm). This interaction map may be considered to be a network in which proteins are represented as the nodes and two nodes are linked by an edge if the two proteins that are connected can be expected with high or medium confidence to be interacting.

The indispensability of a protein defines the functional significance of a gene at its most basic level. Essential genes are those upon which the cell depends for its viability. Using the GENECENSUS database (http://bioinfo.mbb.yale.edu/genome/), we checked for all proteins in the main cluster of the yeast PIN for essentiality. The main cluster, i.e., the larger cluster containing no isolated protein, of 2224 proteins sharing 6608 interactions which is used here for the analysis is illustrated in Figure 1. Protein bipartivity was calculated by an in-house program using Matlab.

**III. Results and Discussion**

According to our calculations, most of the proteins in the yeast PIN show very low bipartivity. As illustrated in Figure 2A, 86% of all proteins in the yeast PIN display bipartivity values lower than 0.6, i.e., \(0.5 < \beta(i) \leq 0.6\) and only 4% of proteins can be considered to be highly bipartite, i.e., \(0.9 \leq \beta(i) \leq 1\). These results indicate that most yeast proteins are implicated in a large number of odd-length cycles in the PIN. In Figure 2B, we illustrate the percentage of essential proteins which are found among the least bipartite proteins in the PIN, e.g., those proteins for which \(0.5 < \beta(i) \leq 0.6\). As can be seen, there is a
The interactions a protein has.18 This very intuitive idea has of a protein will prove lethal correlates with the number of interactions the protein has. It was previously concluded that “the likelihood that removal of a protein will prove lethal correlates with its lack of bipartivity”. Unfortunately, the combination of several centrality criteria in a linear discriminant analysis does not improve the number of essential proteins identified. Thus, we think that the best possible choice for combining the information obtained from different centrality measures would be the use of partial order ranking techniques in order to improve the predictability of this method of essential proteins identification.

Our previous results raise several questions concerning the importance of protein bipartivity and its role in the structure, function, and evolution of the yeast PIN. The first question is related to the differences found between bipartivity and connectivity to identify essential proteins. Simple reasoning suggests that the most connected proteins in the PIN are the least bipartite ones. This is based on a simple transitivity principle, if protein A interacts with proteins B and C there is a high probability that proteins B and C interact. Despite the fact that this principle is in general obeyed, in the yeast PIN there are many cases in which proteins with the same number of interactions have different bipartivities. One simple example is provided in Figure 1, panels B and C, where two proteins, YGR200C and YHR052W, having interactions with 6 other proteins show very different bipartivities, very high for YGR200C and very low for YHR052W. This situation is even more critical for proteins with a low number of interactions which can display bipartivities ranging from 0.5 to 1.0 (data not shown). A simple example is provided by proteins forming a triangle or a square in the PIN. Both proteins have interactions with only two other proteins, but although the one forming the square is bipartite, the other is not. A similar lack of correlation exists between the bipartivity and the transitivity measured through the so-called “clustering coefficient”,20 which is the number of triangles in which a protein takes part divided by the number of triples centered on such a protein (data not shown). In this case, we can find pairs of proteins having zero clustering and different bipartivities. This is the case of proteins forming a square or a pentagon; both have zero clustering but the one forming the square is bipartite and that forming the pentagon is not. There are also proteins with the same bipartivity and different clustering, such as proteins in a triangle and a pentagon; both are bipartite but have clustering of one and zero, respectively. Thus, protein bipartivity captures topological information which is not represented by other centrality measures of a protein in a PIN and more importantly, this information is relevant for protein essentiality.

It is known that biological function generally emerges from the interaction of many separate components, such as genes, proteins, or metabolites, which form distinct functional modules.33–35 These modules are organized in the form of structural motifs, topologically distinct interaction patterns within a network.49,50 In contrast, a biological function can only rarely be assigned to a single component and some of these motifs, in particular the fully connected ones, tend to identify protein complexes responsible for distinct biological functions.50 From the point of view of our current study, this means that those proteins forming part of highly connected motifs representing functional modules have the highest probability of being essential in a PIN. However, not all proteins in a large...
protein complex interact with each other, and in many cases, highly interconnected motifs are formed among proteins of different complexes. An illustrative example is provided by the subnetwork formed by 11 proteins of the 26S proteasome...
regulatory subunit and 10 proteins of the proteasome component (Figure 4). The first acts as a regulatory subunit of the 26S proteasome which is involved in the ATP-dependent degradation of ubiquitinated proteins and the second is a multicatalytic proteinase complex. Most proteins in this subnetwork show very low bipartivity, and consequently, 81% of them are essential. Close examination of this subnetwork permits identification of several fully connected motifs, such as those illustrated in Figure 4. The largest of these motifs is formed by 6 proteins of the 26S proteasome complex, which includes the protein YFR052W necessary for the activation of the CDC28 kinase. There are also fully connected motifs formed by combinations of the proteins of both complexes as illustrated at the bottom of this figure.

Finally, we would like to note that our current findings shed some light on the nature of the evolution of essential proteins in yeast. We have shown here that essential proteins are the least bipartite with respect to their interactions in the yeast PIN. Assuming a duplication-divergence model of protein evolution\(^1\), we arrive at the hypothesis that essential proteins evolve from others which have very low bipartivity and, with high probability, are also essential. Our reasoning is as follows. Consider a bipartite protein \(i\) such as the one illustrated in black in Figure 5A. After the duplication of \(i\), a new protein \(i'\), a replica, (in gray) with an interaction to all the neighbors of \(i\), is created. At this point, the new protein forms cycles of even but not of odd length, i.e., it is also bipartite. If we allow an interaction between \(i\) and \(i'\) (self-interacting proteins), the bipartivity of both proteins decreases but their values of \(\beta(i)\) are only slightly different from one, as they are taking part only in a pair of triangles. For the sake of simplicity, we will not consider here interaction between \(i\) and \(i'\). The next step is the divergence, where some common interactions will be lost, and new interactions may be gained by either protein. During this process, to generate a protein displaying very low bipartivity, it is necessary that all previous interactions are retained and new ones created to complete links between all pairs of proteins (see the top option in Figure 5A). However, according to a simple parsimony principle this process is not very probable and the most probable one evolves in proteins with bipartivity close to one as illustrated at the bottom of Figure 5A. Consequently, a replica protein evolving from a bipartite target will also be bipartite with high probability. In Figure 5B, we illustrate a duplication-divergence process starting from a target with very low bipartivity, e.g., \(\beta(i) \approx 0.5\). After the duplication of this protein, the replica protein will display a very low bipartivity, similar to its target. During the divergence stage, it is necessary to remove all interactions between pairs of proteins which are connected to \(i\) and \(i'\) in order to produce a bipartite replica. This process appears to have very low probability as it will remove almost all interactions existing previously in the cluster. It appears to be more probable that only some of these interactions are removed, retaining some others and creating new ones, which result in a replica protein with low bipartivity, similar to its target. Our conclusion is that a replica protein evolving from a nonbipartite target will be also nonbipartite with high probability. As we shown here that nonbipartite proteins display high probability of being essential then these new replicas evolving from nonbipartite (essential) targets will also display high probability of being essential.

Our hypothesis is supported by several observations. First, there is a large number of fully connected motifs (cliques) in yeast PIN. Here we have calculated the existence of 3530 triangles (3 cliques), 2576 fully connected squares (4 cliques), and 1711 fully connected pentagons (5 cliques). This abundance of very low bipartivity clusters has been also observed by others\(^{46,50,55}\) and has been explained by evolutionary models based on duplication-divergence network growth.\(^{52}\) Even more interesting is the fact that nonbipartite motifs display significantly larger evolutionary conservation rates than are seen in bipartite motifs.\(^{50}\) For instance, the 5 clique shows the highest conservation rate among all motifs (47.24%), followed by the 4 cliques (32.53%) and the 3 clique (20.51%).\(^{50}\) This trend is also observed in nonfully connected motifs, such as squares and pentagons. Although the bipartite motif (square) shows a conservation rate of only 6.71%, the nonbipartite motif (pentagon) doubles this rate to 14.77%. Bipartite motifs never exceed 15% of the evolutionary conservation rate.\(^{50}\)

### IV. Conclusions

Our current study provides a thorough assessment of the relevance of protein bipartivity for understanding essentiality in yeast PIN. Protein bipartivity not only clearly surpasses other protein topological measures in a PIN but also gives insights about the possible mechanisms associated with the evolution of essential genes in yeast. It is important to remark that the abundance of essential proteins found among the least bipartite proteins clearly exceeds that found among the most connected ones. Based on our findings, we have proposed a
possible mechanism for the evolution of essential proteins in yeast PIN based on the duplication-divergence scheme. Using such a model, we can conclude that a replica protein evolving from a nonbipartite target will also be nonbipartite with high

Figure 5. Possible mechanisms of protein evolution starting from a bipartite protein (A) and from a nonbipartite one (B). Target is in black, replica, in gray, and the rest of proteins, in white (see text for explanation).
probability. Consequently, these new replicas evolving from nonbipartite (essential) targets will with high probability be essential.

Acknowledgment. The author thanks the "Ramón y Cajal" program, Spain for partial financial support.

References


(21) van Noort, V.; Snel, B.; Huynen, M. A. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple mode. EMBO Rep. 2004, 5, 1–5.

(22) Vázquez, A.; Flammini, A.; Maritan, A. Modeling of protein interaction networks. ComPlexUs 2003, 1, 38–44.


