RESEARCH ARTICLE

# Path lengths in protein–protein interaction networks and biological complexity

*Ke Xu[1], Ivona Bezakova[2], Leonid Bunimovich[3] and Soojin V. Yi[1]*

[1] School of Biology, Georgia Institute of Technology, Atlanta, GA, USA
[2] Department of Computer Science, Rochester Institute of Technology, Rochester, NY, USA
[3] School of Mathematics, Georgia Institute of Technology, Atlanta, GA, USA

Received: October 26, 2010
Revised: January 16, 2011
Accepted: January 24, 2011

We investigated the biological significance of path lengths in 12 protein–protein interaction (PPI) networks. We put forward three predictions, based on the idea that biological complexity influences path lengths. First, at the network level, path lengths are generally longer in PPIs than in random networks. Second, this pattern is more pronounced in more complex organisms. Third, within a PPI network, path lengths of individual proteins are biologically significant. We found that in 11 of the 12 species, average path lengths in PPI networks are significantly longer than those in randomly rewired networks. The PPI network of the malaria parasite *Plasmodium falciparum*, however, does not exhibit deviation from rewired networks. Furthermore, eukaryotic PPIs exhibit significantly greater deviation from randomly rewired networks than prokaryotic PPIs. Thus our study highlights the potentially meaningful variation in path lengths of PPI networks. Moreover, node eccentricity, defined as the longest path from a protein to others, is significantly correlated with the levels of gene expression and dispensability in the yeast PPI network. We conclude that biological complexity influences both global and local properties of path lengths in PPI networks. Investigating variation of path lengths may provide new tools to analyze the evolution of functional modules in biological systems.

**Keywords:**
Bioinformatics / Biological complexity / Node eccentricity / Path length / Protein–protein interactions

## 1 Introduction

One of the most extensively analyzed biological networks is protein–protein interaction (PPI) networks, where each protein is treated as a ''node'' in the network and the interaction between two proteins as an ''edge.'' Analyses of PPI networks led to several major, if sometimes controversial, findings. The PPI network of the budding yeast *Saccharomyces cerevisiae*, similar to many other networks (such as metabolic networks and internet), exhibits the so-called ''power-law'' distribution [1] where most proteins have few neighbors and only a few proteins have a large number of neighbors. Proteins with large number of neighbors are thus referred to as ''hubs.''

It is proposed that hubs in PPI networks are more essential than non-hub proteins [2] and evolve slower [3]. These proposals inspired numerous in-depth studies to evaluate the functional and evolutionary importance of the position of a protein in PPI networks [4–6]. Often, these studies examine the correlations between a network characteristic (such as the connectivity) and a biological variable (such as the expression level). It is now widely appreciated that it is critical to confirm that a correlation between a network characteristic and the biological measures is not

**Correspondence:** Professor Soojin V. Yi, School of Biology, School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA
**E-mail:** soojin.yi@biology.gatech.edu; soojinyi@gatech.edu
**Fax:** +1-404-894-2295

**Abbreviation: APL,** average path length; **CAI,** codon adaptation index; **NCC,** network clustering coefficient; **PPI,** protein–protein interaction

**Colour Online:** See the article online to view Figs. 1 and 3 in colour.

 

due to confounding effects of other variables. For example, the negative relationship between the connectivity and the evolutionary rates [3] may have been caused by the fact that in yeast, highly expressed genes tend to evolve slower than those that are lowly expressed, and those genes sometimes are recorded as protein hubs due to biases in experimental procedures. When data obtained from the experiments that are relatively free from such a bias were used, the observed trend of a negative correlation between network centrality and evolutionary rates decreased or disappeared altogether [7, 8]. Thus, network approach, combined with a careful statistical treatment of biological variables, can yield important insights into the properties of biological networks.

In this study, we investigated a relatively little studied aspect of PPI networks, namely the path length between nodes, defined as the shortest path between two nodes. Biological networks such as PPI typically exhibit the ''small-world property,'' meaning that path lengths between nodes are generally short [1, 2]. Interestingly, despite the widespread small world property, path lengths in real networks are generally longer than those from random networks (where connections between two nodes are made entirely random, known as Erdös -Rényi network: [9]) [10]. It is proposed that longer path lengths may be beneficial [10].

Here, we investigated biological importance of path lengths between nodes in detail. Specifically, we hypothesized that the average path length (APL) may increase as biological complexity increases. As complex biological systems evolve, specific sets of biological interactions need to be under tight regulation. This hypothesis is in line with the idea that longer path lengths may be favored due to the advantage related to functionality [10]. Furthermore, we note that within a network, different proteins harbor different path lengths. We hypothesized that path length traits of individual proteins also harbor biological importance.

To test these hypotheses, here we analyzed the PPI networks of the following 12 organisms: *Homo sapiens* [11], *Drosophila melanogaster* [11], *Caenorhabditis elegans* [11], *Arabidopsis thaliana* [11], *S. cerevisiae* [12], *Schizosaccharomyces pombe* [11], *Plasmodium falciparum* [13], *Helicobacter pylori* [14], *Treponema pallidum* [15], *Campylobacter jejuni* [16], *Escherichia coli* [17], and *Synechocystis* sp. *PCC6803* [18]. These networks represent the most comprehensive PPI networks currently available (Section 2, Table 1). Path length traits from these real PPIs are compared with those from rewired networks where connections are made randomly following Erdös-Rényi networks (Section 2).

We observed that PPI networks generally exhibit longer path lengths than simulated random networks with identical degree distribution. As we proposed, the deviation from the expected path lengths is significantly greater in eukaryotes compared with prokaryotes.

To investigate the importance of path lengths of individual proteins, we analyzed the significance of the ''eccentricity'' (shortest path length of each protein) in relation to other biological traits in a multivariate statistical setting. For this analysis, we used the yeast PPI, where extensive functional data are available. We uncovered that the path lengths of individual proteins are indeed significantly correlated with essentiality of the proteins themselves. These results provide supports to the idea that path lengths of PPIs are related to biological complexity.

## 2 Materials and methods

### 2.1 PPI data set

PPI data from different species are obtained by mining all major databases of PPI data. We used the following two criteria for selection: first, the methods of PPI

**Table 1.** Network statistics of the 12 PPI networks analyzed

| | Initial network | | Largest connected components | | | |
|---|---|---|---|---|---|---|
| | Node | Edge | Node | Edge | Density[a)] | Source |
| *H. sapiens* | 8448 | 30 296 | 8240 | 30 178 | 3.66 | BioGRID [11] |
| *D. melanogaster* | 7373 | 23 923 | 7272 | 23 869 | 3.28 | BioGRID [11] |
| *C. elegans* | 2794 | 4457 | 2575 | 4328 | 1.68 | BioGRID [11] |
| *A. thaliana* | 1678 | 3006 | 1241 | 2624 | 2.11 | BioGRID [11] |
| *S. cerevisiae* | 2708 | 7123 | 2559 | 7031 | 2.75 | [12] |
| *Sc. pombe* | 1432 | 2560 | 1342 | 2501 | 1.86 | BioGRID [11] |
| *P. falciparum* | 1260 | 2597 | 1221 | 2577 | 2.11 | MINT [13] |
| *H. pylori* | 724 | 1403 | 710 | 1396 | 1.97 | [14] |
| *T. pallidum* | 576 | 978 | 561 | 969 | 1.73 | [15] |
| *Ca. jejuni* | 1091 | 2966 | 1081 | 2961 | 2.74 | [16] |
| *E. coli* | 1755 | 6168 | 1687 | 6133 | 3.64 | DIP [17] |
| *Sy. sp. PCC6803* | 1849 | 2986 | 1775 | 2944 | 1.66 | MPIDB [18] |

a) Network density is calculated by the number of edges divided by the number of nodes as in [52].

 

characterization should be experimental. Second, the number of interactions should be at least 1000 (except the case of *T. pallidum*, which has 978 interactions). Following these criteria, we selected 12 PPI networks. Main references, methods of obtaining these data, and the sizes of the initial networks as well as those of the maximal components (see below) are listed in Table 1.

### 2.2 Network analysis

To analyze the distribution of path lengths between all pairs of proteins, we analyzed the ''largest connected component'' of each PPI network, defined as the maximum *connected* subgraph (the subgraph cannot be made larger by adding additional vertices to it) within each PPI network. We used the breadth first search algorithm [19] to obtain these components.

We identified the shortest path (path length) between any pair of proteins using the ''breadth first search'' algorithm. Node eccentricity is the largest shortest distance for any protein. Node clustering coefficient is defined as $C_v = e_v/k_v(k_v-1)/2$ where each node v has $k_v$ neighbors and the actual number of edges existing among the neighbors is $e_v$. Network clustering coefficient (NCC) is the average of $C_v$ over all the nodes.

### 2.3 Simulation

The main difference between our work and other analyses of PPI networks lies in our focus on the largest connected component. Our goal is to understand the distribution of path lengths in this component and compare it with the expected distribution, i.e., distribution likely to be seen in a random connected network with similar characteristics. Typically, the degree sequence, i.e., the sequence of numbers of neighbors of all nodes in the network, is considered to be a fixed characteristic. Thus, we needed to generate (a large number of) random connected networks with the same degree sequence as the input data.

We used a Markov chain simulation approach proposed by Gkantsidis et al. [20]. The simulation starts with an arbitrary connected network that satisfies the target degree sequence; we generate this initial network using the Havel–Hakimi algorithm [21, 22] and then we iteratively merge individual connected components (this needs to be done carefully so that the degrees of the nodes do not change, see [20] for details) to get a single connected network.

In each step of the Markov chain, we pick two edges with distinct end points uniformly at random, let these edges be $(u,v)$ and $(x,y)$. We replace these edges by edges $(u,x)$ and $(v,y)$, if both of these conditions hold: (i) edges $(u,x)$ and $(v,y)$ are not present in the network, and (ii) this change keeps the network connected. Since connectivity testing is computationally much more expensive than the edge swap, we implemented the ''window-based'' speed-up of the Markov chain described in [20]. The idea is to remember the last generated network that positively tested for connectivity, then perform a sequence of random edge swaps and check for connectivity again; if the resulting network is connected, we keep it, otherwise, we ''rollback'' to the memorized connected network. We used 3 000 000 swaps to generate each connected graph (in comparison, [20] showed that for power-law graphs with up to 14 000 nodes, this Markov chain will reach convergence within 2 000 000 steps). Using this simulation method, we generated 1000 random connected graphs with the given degree distribution for all the PPI networks analyzed.

### 2.4 Biological traits

We analyzed the relationship between network traits and biological traits more deeply using functional data from yeast. The yeast protein abundance data are obtained from a *S. cerevisiae* fusion library, where the absolute abundance of each open reading frame in its natural chromosomal location is visualized by immunodetection [23]. The yeast codon adaptation index (CAI) values and evolutionary constraints ($dN/dS$) are obtained from [24]. Data on yeast gene expression are obtained from [25]. Fitness effects of gene deletion in yeast are obtained from [26].

### 2.5 Statistical analyses

In order to investigate the relationship between two variables while controlling other confounding effects, partial correlation analysis is conducted using R (version 2.6.1, http://www.r-project.org/, partial correlation package is from [27]. Deviation of observed value from simulated values is normalized through calculating *Z*-score, which is defined as $Z = X-m/s(x)$ where $X$ is the observed value, $m$ is the mean of the simulated values, and $s(x)$ is the standard deviation of the simulated values.

## 3 Results

### 3.1 Most, but not all, PPI networks have longer path lengths than randomly rewired connected networks with identical degree distribution

PPI networks typically characterize the interactions between proteins in an ''undirected'' manner (e.g. [28]). Here, we are interested in comparing path lengths between proteins in the PPI networks with those of simulated networks where each connection is rewired randomly (as in [9]). To preserve the characteristic power-law distribution of connectivities, these networks keep identical power-law distribution of

 

degrees to the original networks. Furthermore, to avoid having nodes with no connections to other nodes, we kept the simulated networks connected.

To achieve this, we first extracted the largest ''connected'' subnetwork in which every node can eventually reach every other node in the network. This step allows us to investigate the meaning of path lengths while still preserving the interactions among all the nodes in the network. These rewired networks are sometimes referred to as ''random networks'' henceforth.

The sizes of the PPI network of each species and the largest connected subnetworks are summarized in Table 1. We identified the shortest path (path length) between any pair of proteins (Section 2). To compare the characteristics of PPI networks with those of random networks, we generated 1000 random connected networks with identical degree distribution as the PPI network of interest. We achieved this by randomly rewiring the ''edges'' in the network while keeping the whole network connected, using the algorithm developed in Gkantsidis et al. [20].

We first compared the APL between any pair of proteins in PPI networks with those in random networks. We found that APLs of PPI network are greater than those from the simulated random networks for all the species with the notable exception of the malaria parasite *P. falciparum* (Fig. 1, Table 2). APL of the PPI of the *P. falciparum* is found well within those of the randomized networks (Fig. 1, Table 2). In all other species examined, APLs of the real PPIs are far greater than those from rewired random networks. For example, the APLs of the simulated network starting with the yeast PPI network ranged from 4.089 to 4.182. The observed APL of the yeast PPI (4.745) is much greater than even the maximal APL of the randomized networks.

The disparity between the random networks and the PPI network was the most dramatic in *A. thaliana* PPI network. The APLs of randomized networks range from 4.19 to 4.33, whereas the observed APL of *A. thaliana* PPI network is 8.46 (Figs. 1 and 2, Table 2).

### 3.2 Eukaryotic PPIs exhibit significantly greater deviation from random networks compared with prokaryotic PPIs

We summarized the deviation from the observed variance of the simulated random networks and the PPI network by computing the *Z*-score, which is defined as $Z = X - m/s(x)$ where $X$ corresponds to the observed APL in a PPI network, $m$ corresponds to the mean path lengths of the random networks, and $s(x)$ is the standard deviation of the path lengths from the 1000 simulated random networks. It is important to note that the *Z*-scores are independent of the size and the density of the networks analyzed: *Z*-score is not correlated with network characteristics including the numbers of nodes and edges, and densities ($p$>0.1 for all correlations).

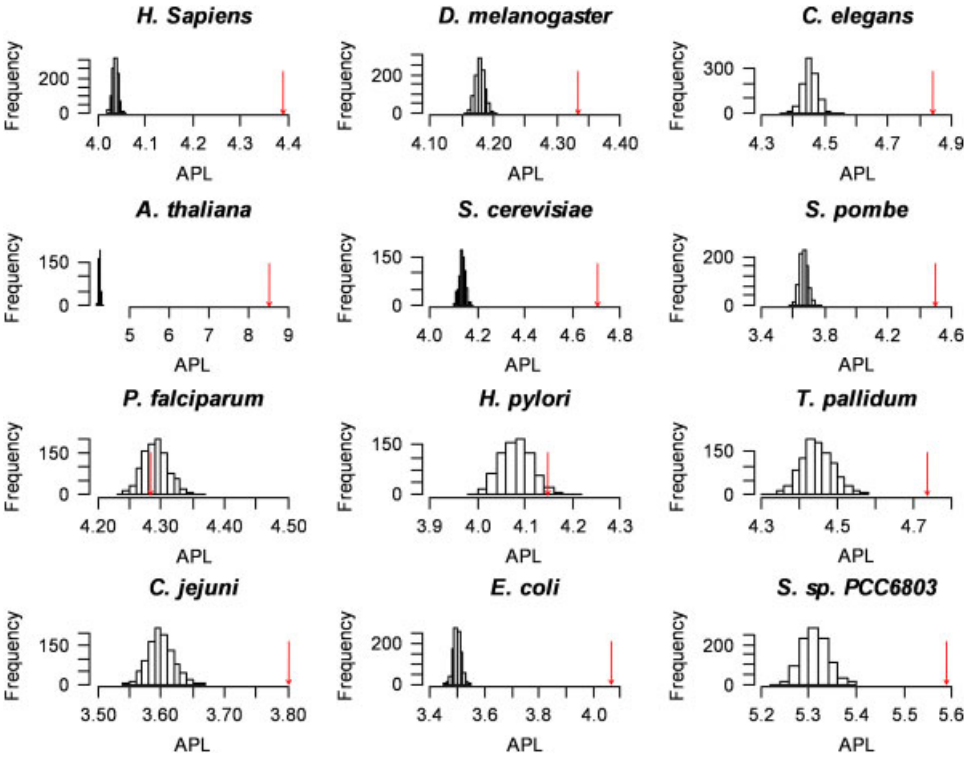


**Figure 1.** Distributions of APLs of the randomized networks with identical degree distributions as the PPI networks of the 12 species analyzed in this study. The observed APL for each species is indicated by red arrows.

**Table 2.** APL and NCC (defined as average of all clustering coefficients of all nodes within the network) of biological and random networks

| | APL | | | NCC | | |
|---|---|---|---|---|---|---|
| | Observed | Random | *Z*-score | Observed | Random | *Z*-score |
| **Eukaryotes** | | | | | | |
| *H. sapiens* | 4.39 | 4.04 | 61.05** | 0.16 | 0.01 | 183.07** |
| *D. melanogaster* | 4.33 | 4.18 | 21.78** | 0.02 | 0.01 | 11.25** |
| *C. elegans* | 4.84 | 4.45 | 19.50** | 0.05 | 0.03 | 5.63** |
| *A. thaliana* | 8.46 | 4.25 | 210.50** | 0.28 | 0.02 | 68.34** |
| *S. cerevisiae* | 4.75 | 4.14 | 50.75** | 0.28 | 0.02 | 147.46** |
| *Sc. pombe* | 4.55 | 3.67 | 34.46** | 0.36 | 0.12 | 31.50** |
| *P. falciparum* | 4.28 | 4.29 | −0.46 | 0.02 | 0.02 | 1.23 |
| | | Mean | 56.80 | | Mean | 64.07 |
| **Prokaryotes** | | | | | | |
| *H. pylori* | 4.15 | 4.08 | 2.09* | 0.03 | 0.04 | −2.29* |
| *T. pallidum* | 4.73 | 4.45 | 6.74** | 0.07 | 0.03 | 6.73** |
| *Ca. jejuni* | 3.80 | 3.60 | 10.39** | 0.06 | 0.07 | −2.94** |
| *E. coli* | 4.07 | 3.50 | 42.61** | 0.13 | 0.08 | 9.88** |
| *Sy. sp. PCC6803* | 5.59 | 5.31 | 10.33** | 0.01 | 0.007 | 1.91* |
| | | Mean | 14.43 | | Mean | 2.66 |

*$p<0.05$, **$p<0.001$.

A large *Z*-value indicates that the PPI network deviates greatly from the simulated random networks. PPI networks have greater APLs than simulated random networks, except the case of *P. falciparum* (Table 2).

We then asked if the deviation increases in eukaryotes compared with prokaryotes. We found that the mean *Z*-score is significantly greater in eukaryotes than in prokaryotes (Wilcoxon rank sum test, $p = 0.015$, Fig. 2A). Since *A. thaliana* appears as an extreme outlier in terms of its high deviation from randomized networks (Table 2), we removed *A. thaliana*. The difference between prokaryotes and eukaryotes is still significant ($p = 0.028$).

### 3.3 Analysis of clustering coefficients

In addition to path lengths, we analyzed the distributions of clustering coefficients. Clustering coefficient of a node can inform us how well connected the neighborhood of the node is [29]. If a node resides in a fully connected neighborhood, its clustering coefficient is 1. A value close to 0 means connections are very sparse around the specific node. We present the average clustering coefficient of all nodes in each PPI, the "NCC" (Fig. 3, Table 2). NCC is an indication of network's potential modularity [30, 31]. Networks with large NCCs tend to be more modular in their structures than those with smaller NCCs.

Interestingly, NCCs of two prokaryotes, *H. pylori* and *T. pallidum*, were smaller than those of the random networks (Fig. 3, Table 2), indicating the lack of intrinsic modularity in these two species' networks.

In other analyzed species, the observed NCCs were greater than those in random networks (Fig. 3, Table 2).

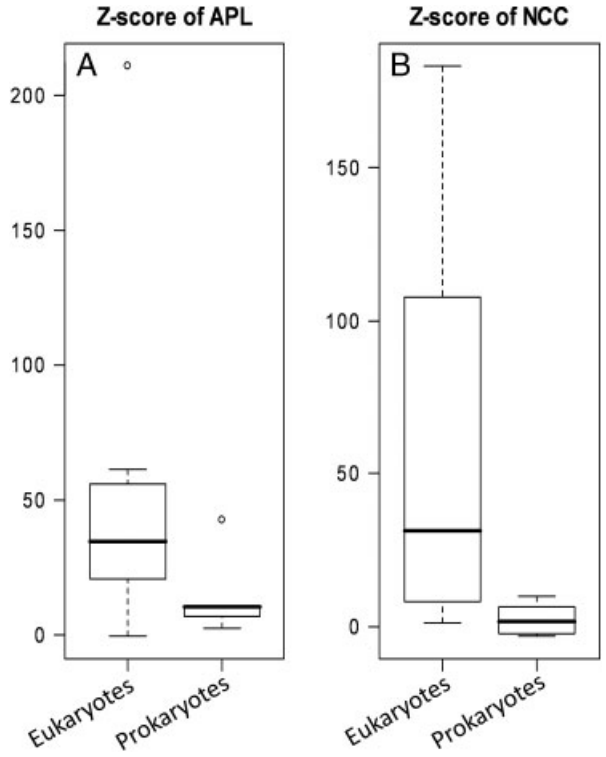


**Figure 2.** Eukaryotic PPI networks exhibit significantly greater deviation from random networks than prokaryotic PPI networks, as measured by the *Z*-scores. (A) *Z*-scores of APLs of eukaryotic PPI networks are significantly higher than those of prokaryotes ($p = 0.0015$). (B) *Z*-scores of NCCs of eukaryotic PPI networks are significantly higher than those of prokaryotes ($p = 0.009$).

Again, eukaryotes generally exhibited larger NCCs than prokaryotes (Wilcoxon rank sum test, $p = 0.009$, Fig. 2B. $p = 0.016$ when *A. thaliana* is removed).

 

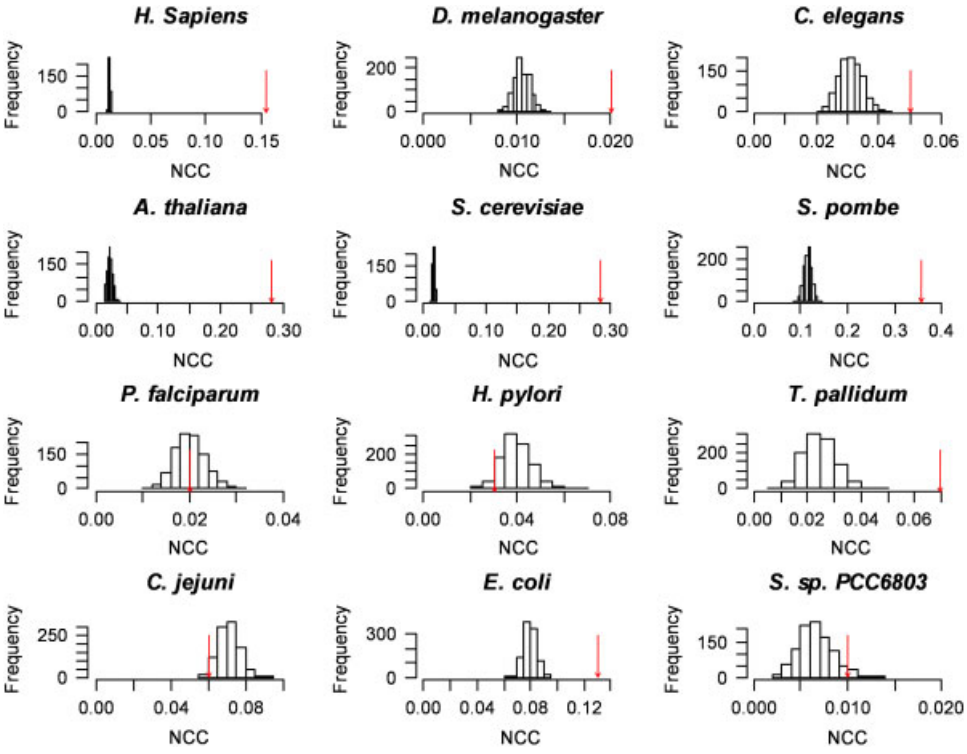


**Figure 3.** Distributions of the NCCs of the connected random networks with identical degree distributions as the real PPI networks of the 12 species analyzed in this study. The observed NCCs for each species are indicated by red arrows.

### 3.4 Eccentricity varies in biological networks compared with random networks

In the previous section, we contrasted two "global" characteristics of PPI networks, namely APLs and NCCs, with those of random networks. In this section, we focus on the path lengths of individual proteins to other proteins in a PPI network, by investigating the distribution and potential significance of "node eccentricity," which is defined as the greatest distance among a node's (shortest) distance to any other node in the network [32]. Intuitively, the eccentricity is the number of steps a protein needs to influence all other proteins in a PPI network.

Table 3 summarizes the distributions of eccentricities in 1000 random networks compared with the distribution of eccentricities in a PPI network. In a connected graph, the minimum eccentricity of all nodes is called the "graph radius." The maximum eccentricity, equivalent to the maximum distance between any two nodes, is referred to as the "graph diameter." We found that the PPI networks of *H. sapiens*, *A. thaliana*, *S. cerevisiae*, *Sc. pombe*, and *E. coli* have significantly greater radii and diameters than those of the random networks.

### 3.5 Analysis of the yeast network and functional data suggest that eccentricity is biologically significant, independent of connectivity and closeness

In this section, we investigate whether node eccentricity is likely to have biological significance. To achieve this, we investigated the significance of node eccentricity in a multivariate statistics setting, using data from *S. cerevisiae*. We chose to investigate *S. cerevisiae* because there exist a large amount of excellent functional data from this species. Importantly, the relationship between network variables and other functional measures in yeast has been analyzed in depth [24, 27, 33, 34].

We evaluated the significance of node eccentricity with two other network traits, namely the connectivity and the closeness. Since node eccentricity is defined as the greatest distance between the node and any other nodes in the network, it is likely to be negatively correlated with the connectivity (degree), because a well-connected protein can reach other proteins in fewer steps. Indeed, eccentricity is strongly negatively correlated with connectivity in all species analyzed (for example, in yeast, Kendall's $\tau = -0.51$, $p < 10^{-15}$; results from other species are not shown). Closeness is two-dimensional measures of a protein's centrality in a network, as opposed to the connectivity, which is a one-dimensional metric of importance in the network [35, 36]. As expected, closeness is positively correlated with degree and negatively correlated with eccentricity (significant in all species: for example, in yeast, Kendall's $\tau = -0.71$, $p < 10^{-15}$ for eccentricity and closeness; results from other species are not shown).

We sought to determine whether the these network traits (connectivity, closeness, and eccentricity) have independent influence on the biological traits, including CAI, protein abundance, gene expression level, fitness effects of heterozygous and homozygous deletion (these can be considered

 

**Table 3.** Radii and diameters of biological networks and random networks derived from the biological networks

| | Radius | | | Diameter | | |
|---|---|---|---|---|---|---|
| | Observed | Random | *Z*-score | Observed | Random | *Z*-score |
| *H. sapiens* | 7 | 5.54 | 2.93* | 13 | 9.98 | 4.79* |
| *D. melanogaster* | 6 | 5.92 | 0.26 | 11 | 10.41 | 0.88 |
| *C. elegans* | 7 | 6.25 | 1.64 | 13 | 11.75 | 1.39 |
| *A. thaliana* | 16 | 5.69 | 21.48** | 31 | 10.23 | 26.32** |
| *S. cerevisiae* | 7 | 5.56 | 2.87* | 12 | 9.99 | 2.87* |
| *Sc. pombe* | 7 | 5.25 | 3.43* | 13 | 10.01 | 3.29* |
| *P. falciparum* | 6 | 5.78 | 0.51 | 10 | 10.23 | −0.28 |
| *H. pylori* | 5 | 5.44 | −0.84 | 9 | 9.99 | −1.16 |
| *T. pallidum* | 7 | 5.97 | 2.19 | 11 | 10.84 | 0.16 |
| *Ca. jejuni* | 5 | 4.91 | 0.25 | 9 | 8.81 | 0.25 |
| *E. coli* | 6 | 4.65 | 2.79* | 12 | 8.28 | 6.30** |
| *Sy. sp. PCC6803* | 8 | 7.34 | 1.24 | 16 | 13.48 | 2.36* |

*$p<0.05$, ** $p<0.001$.

as indicators of gene essentiality [26]), and protein evolutionary rates (d*N*/d*S*). We investigated these by performing partial correlation analyses.

The results from multivariate statistical analyses are summarized in Table 4. As discussed previously, the degree of a node in the yeast PPI network is significantly correlated with the abundance of the protein, the level of expression, gene dispensability, and also with evolutionary rates (Table 4). Similarly, eccentricity is significantly correlated with the CAI, protein abundance, gene expression, and dispensability (Table 4). Closeness is also significantly correlated with several biological variables.

After controlling eccentricity, closeness and all the other biological variables, the correlations between connectivity and CAI/abundance/expression disappear (Table 4). Table 4 also summarizes that when all other variables are considered, connectivity is significantly correlated with the fitness effect of homozygous deletion as well as evolutionary rates. Closeness shows a weak and significant correlation with the fitness effects of heterozygous deletion.

When controlling connectivity, closeness and all the other biological variables, eccentricity was still significantly negatively correlated with the levels of gene expression level (Table 4). In addition, eccentricity was also significantly correlated with the fitness effect of heterozygous deletion. In other words, eccentricity appears to have weak yet significant effects on two highly important biological traits. This novel finding supports the idea that path lengths in PPI are biologically important.

## 4 Discussion

Complex biological functions are often achieved by a series of molecular interactions (e.g. metabolic pathways). Genome-wide functional data generated via high-throughput techniques have propelled a shift from studying the molecular interactions between a few biological molecules (e.g. [37]) to a system-level interrogation of organismal complexity [12, 23, 25, 38–41]. While the efforts to elucidate molecular interactions in biological systems are actively ongoing, the availability of genome-wide interaction data from several species has bolstered even more exciting possibilities, namely to explore the nature of evolutionary principles underlying organization of biological systems, by comparing molecular interactions from several species.

For example, several studies have demonstrated that topological structures of biological networks exert significant influence on molecular evolution of gene sequences and other aspects of genome evolution [35, 42, 43]. Thus, investigating the characteristics of biological networks upon comparative context has a potential to yield significant clues toward evolution of complex biological systems.

In this study, we analyzed path lengths between proteins in PPI networks of 12 species, including APLs between nodes, radii, and diameters, and compared them with those of randomly rewired networks with identical degree distributions. In particular, we have sought to investigate the pattern that real networks tend to have longer path lengths than random networks. For example, Albert and Barabasi [44] noted that the diameter of several real networks was greater than theoretical expectations. Zhang and Zhang [10] investigated 13 diverse real networks and demonstrated that most real networks exhibit significantly longer APLs than expected in random networks.

We investigated 12 biological PPI networks, representing an exhaustive list of experimentally defined large PPI networks up to date (i) to further characterize the pattern of path lengths in biological PPI networks, (ii) to contrast path length characteristics of prokaryotic and eukaryotic PPI networks, (iii) to evaluate the relationship between path lengths and clustering coefficients, and (iv) to evaluate biological significance of path length characteristics of individual proteins using a specific example of *S. cerevisiae*.

Indeed, we observe that PPI networks generally have longer path lengths than observed in random networks.

 

**Table 4.** Results of normal and partial Spearman correlation analyses of yeast functional data

| | Connectivity | | | Eccentricity | | | Closeness | | |
|---|---|---|---|---|---|---|---|---|---|
| | Normal | Partial I[a)] | Partial II[b)] | Normal | Partial I[c)] | Partial II[d)] | Normal | Partial I[e)] | Partial II[f)] |
| CAI | 0.064 | −0.010 | −0.065 | −0.102* | −0.026 | 0.029 | 0.106* | 0.039 | 0.066 |
| Abundance | 0.120* | 0.041 | 0.021 | −0.132* | −0.038 | −0.001 | 0.132* | 0.020 | 0.025 |
| Expression | 0.122* | 0.041 | 0.012 | −0.163* | −0.091* | −0.086* | 0.135* | −0.018 | −0.038 |
| Essentiality_het[g)] | −0.033 | −0.017 | −0.015 | 0.079* | 0.101* | 0.102* | −0.027 | 0.070 | 0.072* |
| Essentiality_hom[h)] | −0.192* | −0.167* | −0.153* | 0.106* | 0.035 | 0.019 | −0.101* | 0.051 | 0.037 |
| d*N*/d*S* | −0.010* | −0.093* | −0.086* | 0.055 | 0.028 | 0.0001 | −0.045 | 0.040 | 0.065 |

*$p < 0.05$.
a) Partial correlation given eccentricity and closeness.
b) Partial correlation given eccentricity, closeness, and all other biological traits.
c) Partial correlation given connectivity and closeness.
d) Partial correlation given connectivity, closeness, and all other biological traits.
e) Partial correlation given connectivity and eccentricity.
f) Partial correlation given connectivity, eccentricity, and all other biological traits.
g) Average fitness value for heterozygous deletion.
h) Average fitness value for homozygous deletion.

However, not all biological PPI networks exhibit longer path lengths than rewired random networks. The PPI network of a common malaria parasite *P. falciparum* does not deviate from simulated networks (Table 2). This suggests that there exists a wide range of path lengths traits in biological PPI networks. In particular, it is of great interest that the PPI network of *P. falciparum* has been shown to deviate strongly from other eukaryotic networks in terms of conserved modules [45].

Aside from the outstanding pattern observed from *P. falciparum*, we uncovered that the degree of deviation from random networks varies greatly between species. We note that this variation is not caused by the difference in the densities (a measure of how many connections individual node has on average) of different networks. In fact, the human PPI network, which has the largest number of connections per node, shows the second greatest deviation from the random network in terms of APL, and the largest deviation in terms of NCC (Tables 1 and 2). Furthermore, we found that increase of path lengths in PPI networks is common to most nodes, not due to a few specific nodes (results not shown).

Zhang and Zhang [10] posited that APLs of real networks might increase to facilitate modularity. As networks evolve to execute complex functions, some nodes may reside in modules for a specialized function. As such, modularity can promote evolvability, multifunctionality, and robustness [10]. Through extensive simulation, Zhang and Zhang [10] have demonstrated that even a modest increase in APLs can significantly favor modularity of the network, especially for networks with small path lengths. Indeed, we show that most PPI networks exhibit significantly greater clustering coefficients, a measure of potential network modularity (Table 2). Thus, longer path lengths in biological networks are likely selectively favored due to the advantage conferred by increased modularity. However, we note that as seen from the pattern in *P. falciparum*, different path length traits may exist in biological systems, which may manifest in different functionalities.

Intriguingly, we found that the deviation of APLs compared with those of random networks, measured by the *Z*-scores, is markedly different between the PPI networks of prokaryotes and the eukaryotes. Eukaryotic PPI networks generally deviated much more from the random expectation than prokaryotic PPI networks. Furthermore, potential modularities of eukaryotic PPI networks, indicated by clustering coefficients, are over an order of magnitude greater than those of prokaryotes (Table 2). As the evolution of eukaryotic species represents one of the major evolutionary leaps in biological complexity [46, 47], such observations again are in accordance with the causal selective mechanism. Namely, path lengths in biological networks may be selectively elongated to increase or favor biological complexity.

Previous analyses of metabolic networks from the three domains of life suggest that a similar phenomenon may exist in other biological networks as well. Ma and Zeng [41] showed that metabolic networks of eukaryotes and archea tend to harbor longer path lengths and larger diameters than those of prokaryotes. Moreover, Wang et al. [48] compared the metabolic networks of photosynthetic cyanobacteria and chloroplasts, and reported that the chloroplast metabolic networks exhibited longer APLs and larger diameters than those of cyanobacteria. Metabolic networks of chloroplasts are believed to have been derived from those of cynobacteria [49, 50], lost many genes by transfers to nuclei and became specialized to carry out photosynthetic functions. Thus, the observation that the eukaryotic (chloroplast) metabolic networks exhibit longer path lengths than prokaryotic (cyanobacteria) networks with the same

 

evolutionary origin again supports the hypothesis that a selective mechanism, associated with evolution of complex and modular biological functions, exists to favor longer path lengths in biological networks.

We note that the Z-scores are by no means in a perfect accord with the perceived levels of organismal complexity: it is unlikely that yeasts are more complex than, say, flies. Nor we are suggesting that *E. coli* is more complex than some eukaryotes (the *Z*-score of *E. coli* PPI networks is higher than those of several eukaryotic species, including *D. melanogaster* and *C. elegans*, Table 2). The lack of a linear relationship between the *Z*-scores and the complexity may be due to the fact that PPI data are largely incomplete from the studied species. It is possible that the PPI networks of two commonly used model organisms *E. coli* and *S. cerevsiae* represent more complete characterizations of the real networks, and most other networks are underrepresented in the current data sets. We expect to observe a more concordant relationship between organismal complexities and the path length characteristics once more complete and accurate data on protein interactomes become available from diverse species.

Another novel finding of our study is the reported parallel between the *global* characteristics of network (path length measures) and a *local* characteristic, the so-called ''eccentricity.'' Eccentricity of a node is defined as the maximal distance of a node to any other nodes in a network. In a connected PPI network, a protein with large eccentricity will need many interactions to influence all other proteins. In other words, a node with a large eccentricity is less ''efficient'' in terms of reaching other nodes in the same network. However, in biological networks, such characteristic may be favored for regulatory purposes. As regulation becomes more specialized, some proteins need to reside in isolated modules to be activated in specific developmental time or following specific external signals.

We investigated this hypothesis using available functional data from yeast. We chose the yeast, because there are outstanding high-quality functional data available from this species. Importantly, relationship between network characteristics and functional biological variables in yeast has been extensively characterized.

It is crucial to evaluate the role of a biological variable while accounting for the influence of other variables [27, 33]. We have used the partial correlation analyses of multiple variables to evaluate the role of eccentricity independent of other variables, in particular of the connectivity and closeness, which are strongly correlated to eccentricity. We show that eccentricity is significantly negatively correlated with measures of expression and gene dispensability, independent of connectivity and closeness, in yeast proteins (Table 4). These observations fit the idea that path lengths represent an important network characteristic that promoted evolution of biological complexities of PPI networks, at both global and local levels.

Our study demonstrates that by investigating properties of biological networks and evaluating their relationships to other biological variables, we can gain new insights into the evolutionary principles underlying evolution of complex biological systems. Although our understanding on evolution of biological networks is still very much in its infancy, observations such as ours may guide future directions to such endeavors. Moreover, our study provides an interesting prediction that can be applied to other real networks: the relationship between path lengths and functional specificity of nodes may hold in other nontrivial networks as well.

Our study also presents several intriguing observations that deserve future investigations. For example, it is of great interest that the clustering coefficients of two prokaryotes, *H. pylori* and *Ca. jejuni*, are significantly smaller than those of the random networks. Both these species are gram-negative bacteria that cause gastrointestinal diseases. In light of a recent finding that the PPI network of the herpesvirus resembles a single module on its own (possesses little modularity) yet changes its topology into interacting submodules once the viral network connects to the host networks [51], whether similar mechanisms exist in some parasitic prokaryotic PPI networks (such as those of *H. pylori* and *Ca. jejuni*) is an interesting question.

*This study is supported by the Applied & Biological Contemporary Mathematics program at the College of Science at Georgia Institute of Technology, Alfred P. Sloan Research Fellowship to S. V. Y., and other funds from the Georgia Institute of Technology to S. V. Y. The authors thank Milena Mihail and Eric Vigoda for discussions, Alba Ferrer for computational help in calculating clustering coefficients and suggestions, and two anonymous reviewers for comments on our previous version of the manuscript.*

*The authors have declared no conflict of interest.*

## 5 References

[1] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., Barabasi, A. L., The large-scale organization of metabolic networks. *Nature* 2000, *407*, 651–654.

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

[3] Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., Feldman, M. W., Evolutionary rate in the protein interaction network. *Science* 2002, *296*, 750–752.

[4] He, X. L., Zhang, J. Z., Why do hubs tend to be essential in protein networks? *PLOS Genet.* 2006, *2*, 826–834.

[5] Prachumwat, A., Li, W. H., Protein function, connectivity, and duplicability in yeast. *Mol. Biol. Evol.* 2006, *23*, 30–39.

[6] Wuchty, S., Evolution and topology in the yeast protein interaction network. *Genome Res.* 2004, *14*, 1310–1314.

 

[7] Hahn, M. W., Conant, G. C., Wagner, A., Molecular evolution in lage genetic networks: does connectivity equal constraint? *J. Mol. Evol.* 2004, *58*, 203–211.

[8] Jordan, I. K., Wolf, Y. I., Koonin, E. V., No simple dependence between protein evolution and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *Biomed. Chromatogr. Evol. Biol.* 2003, *3*, 1.

[9] Erdös, P., Rényi, A., On random graphs. *Publ. Math. Debrecen.* 1959, *6*, 290–297.

[10] Zhang, Z., Zhang, J., A big world inside small-world networks. *PloS One* 2009, *4*, 1–6.

[11] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L. et al., BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 2006, *34*, D535–D539.

[12] Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X. et al., Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 2006, *440*, 637–643.

[13] Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D. et al., MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.* 2010, *38*, D532–D539.

[14] Rain, J. C., Selig, L., De Reuse, H., Battaglia, V. et al., The protein-protein interaction map of *Helicobacter pylori*. *Nature* 2001, *409*, 211–215.

[15] Titz, The binary protein interactome of *Treponema pallidum* – The syphilis spirochete. *PloS One* 2008, *3*, e2292.

[16] Parrish, J. R., Yu, J. K., Liu, G. Z., Hines, J. A. et al., A proteome-wide protein interaction map for *Campylobacter jejuni*. *Genome Biol.* 2007, *8*, R130.

[17] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K. et al., The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 2004, *32*, D449–D451.

[18] Goll, J., Rajagopala, S. V., Shiau, S. C., Wu, H. et al., MPIDB: the microbial protein interaction database. *Bioinformatics* 2008, *24*, 1743–1744.

[19] Cormen, T. H., Leiserson, C. E., Rivest, R. L., Stein, C., *Introduction to Algorithms*, 2nd Edn, The MIT Press, Cambridge, 2001.

[20] Gkantsidis, C., Mihail, M., Zegura, E., *Proceedings of 5th Workshop on Algorithm Engineering and Experiments (ALENEX). SIAM* 2003, pp. 16–25.

[21] Havel, V., A remark on the existence of finite graphs. *Caposis Pest. Mat.* 1955, *80*, 477–480.

[22] Hakimi, S. L., On the realizability of a set of integers as degrees of the vertices of a graph. *SIAM J Appl. Math.* 1962, *10*, 496–506.

[23] Ghaemmaghami, S., Huh, W., Bower, K., Howson, R. W., Belle, A. et al., Global analysis of protein expression in yeast. *Nature* 2003, *425*, 737–741.

[24] Wall, D. P., Hirsh, A. E., Fraser, H. B., Kumm, J. et al., Functional genomic analysis of the rates of protein evolution. *Proc. Natl. Acad. Sci. USA* 2005, *102*, 5483–5488.

[25] Holstege, F. C. P., Jennings, E. G., Wyrick, J. J., Lee, T. I. et al., Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 1998, *95*, 717–728.

[26] Deutschbauer, A. M., Jaramillo, D. F., Proctor, M., Kumm, J. et al., Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* 2005, *169*, 1915–1925.

[27] Kim, S. H., Yi, S. V., Understanding relationship between sequence and functional evolution in yeast proteins. *Genetica* 2007, *131*, 151–156.

[28] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A. et al., A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000, *403*, 623–627.

[29] Watts, D. J., Strogatz, S. H., Collective dynamics of 'small-world' networks. *Nature* 1998, *393*, 440–442.

[30] Tanay, A., Sharan, R., Kupiec, M., Shamir, R., Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. USA* 2004, *101*, 2981–2986.

[31] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., Barabasi, A. L., Hierarchical organization of modularity in metabolic networks. *Science* 2002, *297*, 1551–1555.

[32] Harary, F., *Graph Theory*, Addison-Wesley, Reading, MA 1994.

[33] Drummond, D. A., Raval, A., Wilke, C. O., A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* 2006, *23*, 327–337.

[34] Kim, S. H., Yi, S. V., Correlated asymmetry of sequence and functional divergence between duplicate proteins of *Saccharomyces cerevisiae*. *Mol. Biol. Evol.* 2006, *23*, 1068–1075.

[35] Hahn, M. W., Kern, A. D., Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* 2005, *22*, 803–806.

[36] Wassermann, W. W., Sandelin, A., Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* 2004, *5*, 267–287.

[37] Yi, S., Bernat, B., Pal, G., Kossiakoff, A., Li, W.-H., Functional promiscuity of squirrel monkey growth hormone receptor toward both primate and nonprimate growth hormones. *Mol. Biol. Evol.* 2002, *19*, 1083–1092.

[38] Bossi, A., Lehner, B., Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.* 2009, *5*, 260.

[39] Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A. et al., A protein interaction map of *Drosophila melanogaster*. *Science* 2003, *302*, 1727–1736.

[40] Li, S. M., Armstrong, C. M., Bertin, N., Ge, H. et al., A map of the interactome network of the metazoan C-elegans. *Science* 2004, *303*, 540–543.

[41] Ma, H. W., Zeng, A. P., Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* 2003, *19*, 270–277.

[42] Jovelin, R., Phillips, P. C., Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biol.* 2009, *10*, R35.

 

[43] Vitkup, D., Kharchenko, P., Wagner, A., Influence of metabolic network structure and function on enzyme evolution. *Genome Biol.* 2006, *7*, R39.

[44] Albert, R., Barabasi, A.-L., Statistical mechanic of complex networks.. *Rev. Mod. Phys.* 2002, *74*, 47.

[45] Suthram, S., Sittler, T., Ideker, T., The Plasmodium protein network diverges from those of other eukaryotes. *Nature* 2005, *438*, 108–112.

[46] Lynch, M., Conery, J. S., The origins of genome complexity. *Science* 2003, *302*, 1401–1404.

[47] Yi, S., Non-adaptive evolution of genome complexity. *BioEssays* 2006, *28*, 979–982.

[48] Wang, Z., Zhu, X. G., Chen, Y. Z., Li, Y. Y. et al., Exploring photosynthesis evolution by comparative analysis of metabolic networks between chloroplasts and photosynthetic bacteria. *Biomed. Chromatogr. Genomics* 2006, *7*, Article 100.

[49] Kishino, H., Miyata, T., Hasegawa, M., Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* 1990, *31*, 151–160.

[50] Martin, W., Stoebe, B., Goremykin, V., Hapsmann, S. et al., Gene transfer to the nucleus and the evolution of chloroplast. *Nature* 1998, *393*, 162–165.

[51] Uetz, P., Dong, Y.-A., Zeretzke, C., Atzler, C. et al., Herpesviral protein networks and their interaction with the human proteome. *Science* 2006, *311*, 239–242.

[52] Scott, J. P., *Social Network Analysis: A Handbook*, Sage Publications Ltd, CA, USA 2000.

 