Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks

Matthew W. Hahn[1] *and Andrew D. Kern*

Center for Population Biology, University of California, Davis

Most proteins do not evolve in isolation, but as components of complex genetic networks. Therefore, a protein's position in a network may indicate how central it is to cellular function and, hence, how constrained it is evolutionarily. To look for an effect of position on evolutionary rate, we examined the protein-protein interaction networks in three eukaryotes: yeast, worm, and fly. We find that the three networks have remarkably similar structure, such that the number of interactors per protein and the centrality of proteins in the networks have similar distributions. Proteins that have a more central position in all three networks, regardless of the number of direct interactors, evolve more slowly and are more likely to be essential for survival. Our results are thus consistent with a classic proposal of Fisher's that pleiotropy constrains evolution.

To examine the evolution of protein-interaction networks and their corresponding components, we examined a subset of three networks made up of only those proteins that have orthologs in a closely related genome. We inferred the networks from 20,252 interactions among 2,434 yeast proteins, 5,977 interactions among 1,997 worm proteins, and 16,002 interactions among 5,082 fly proteins (see Supplementary Material online). It has been shown previously that the connectivity (number of protein–protein interactions) for each protein in the *Saccharomyces cerevisiae* (Jeong et al. 2001; Wagner 2001), *Caenorhabditis elegans* (Li et al. 2004), and *Drosophila melanogaster* (Giot et al. 2003) protein interaction networks is distributed as a power law, where the frequency, $P(d)$, of proteins with d interactors follows $P(d) \approx d^{-\gamma}$. We also see power law distributions in the connectivities within the networks constructed only from proteins with orthologs, with R^2 values equal to or above 0.90 in all three species. All three networks also have similar slopes, with $\gamma = -1.57$, -1.77, and -1.85 for yeast, worm, and fly, respectively. All three networks, therefore, share a highly skewed distribution of protein interactions: a few proteins have many interactors, whereas most have only a small number of interactors. Even the small number of proteins with many interactors are more than are expected in a random graph.

In addition to similarities in the distribution of physical interactions within the network, we also see similarities in the organization of the networks. Although the connectivity of a protein is a one-dimensional metric of its importance in the network, there are two general two-dimensional measures of a protein's centrality in a network: ''betweenness'' and ''closeness'' (Wasserman and Faust 1994). Betweenness is based on the frequency with which a node lies on the shortest path between all other nodes; proteins with high betweenness control the flow of information across a network and, thus, can be important for minimizing response times within a cell (Jeong et al. 2000; Wagner and Fell 2001). Nodes that bridge gaps separating clusters in a network (Ravasz et al. 2002) are likely to be proteins with high betweenness. Closeness measures the average number of nodes connecting a protein to all other proteins; this metric, therefore, takes into account both direct and indirect interactions between proteins in the network. The distribution of betweenness in all three networks is also well explained by a power law distribution ($R^2 > 0.85$ in all three), with similar slopes in each distribution ($\gamma = -3.08$, -3.46, and -2.72 for yeast, worm, and fly, respectively). This highly skewed distribution is analogous to the distribution of metabolic flux across nodes in a metabolic network, which is also a highly skewed distribution (Almaas et al. 2004). Closeness is distributed approximately normally, and is similar among the protein networks with mean values of 0.24 (0.0008) in yeast, 0.18 (0.0012) in worm, and 0.22 (0.0006) in fly.

A protein's connectivity in the yeast protein interaction network has previously been found to correlate negatively with its evolutionary rate (Fraser et al. 2002; Krylov et al. 2003; Hahn, Conant, and Wagner, 2004). The major cause of this correlation appears to be a dependence between the proportion of a protein directly involved in interactions and the proportion of the amino acids conserved (Fraser et al. 2002). We investigated whether this relationship held both in the worm and in the fly networks and whether the other measures of a protein's centrality were any better at predicting rates of evolution in all three networks. We identified orthologs of the proteins in the yeast, worm, and fly networks in the related species *S. paradoxus*, *C. briggsae*, and *D. pseudoobscura*, respectively, and calculated either d_N, the number of nonsynonymous differences per nonsynonymous site (for worm and fly), or d_N/d_S, the nonsynonymous rate divided by the number of synonymous differences per synonymous site (for yeast). (Synonymous substitutions were saturated in the *C. elegans*–*C. briggsae* and *D.melanogaster*–*D. pseudoobscura* comparisons.) Although all three measures of centrality are correlated with one another in each network, we found that betweenness is more strongly correlated with evolutionary rate than the other measures of centrality in all three networks (table 1). Betweenness is significantly correlated with d_N in the worm (Spearman's ρ: −0.12; $P < 0.0001$), fly (Spearman's ρ: −0.07; $P < 0.0001$), and yeast (Spearman's ρ: −0.18; $P < 0.0001$). In the yeast,

[1] Present address: Department of Biology and School of Informatics, Indiana University, Bloomington, IN 47405, USA

The authors contributed equally to this work.

Key words: networks, pleiotropy, constraint, natural selection.

E-mail: mwh@indiana.edu.

Mol. Biol. Evol. 22(4):803–806. 2005
doi:10.1093/molbev/msi072
Advance Access publication December 22, 2004

Table 1
Network Centrality and Evolutionary Rate

	Yeast	Worm	Fly
Connect-Between	**0.21**	**0.96**	**0.94**
Connect-Close	**0.24**	**0.55**	**0.84**
Between-Close	**0.69**	**0.54**	**0.78**
d_N-Betweenness[a]	**−0.174**	**−0.118**	**−0.071**
d_N-Connectivity	**−0.085**	**−0.114**	**−0.064**
d_N-Closeness	**−0.161**	−0.027	***−0.053***

NOTE.—All values are Spearman's nonparametric correlation, ρ. Numbers in bold are significant at $P < 0.0001$, and bold and italicized are significant at $P < 0.001$.

[a] In the yeast comparison the correlations listed are with d_N/d_S.

Table 2
Essentiality and Centrality in Protein Networks

	Yeast	Worm	Fly
Betweenness			
Essential	**0.0009(0.00013)**	**0.0017(0.00031)**	**0.0007(0.00006)**
Nonessential	0.0007(0.00007)	0.0009(0.00008)	0.0004(0.00002)
Connectivity			
Essential	**19.3(1.11)**	**8.2(0.73)**	**9.8(0.43)**
Nonessential	15.8(0.70)	5.6(0.25)	5.7(0.14)
Closeness			
Essential	***0.244(0.0015)***	***0.183(0.004)***	**0.238(0.0012)**
Nonessential	0.239(0.0010)	0.175(0.001)	0.221(0.0006)
d_N			
Essential	**0.031(0.0011)**	**0.102(0.008)**	**0.096(0.003)**
Nonessential	0.044(0.0008)	0.143(0.003)	0.137(0.002)

NOTE.—Mean values for essential and nonessential genes with standard error in parentheses. Mean values for essential genes in bold are significantly different from nonessential genes at $P < 0.0001$ (Wilcoxon two-sample test), and bold and italicized are significant at $P < 0.001$.

we are also able to estimate selective constraint on each protein using d_N/d_S and we also find a significant correlation with betweenness (Spearman's ρ: −0.17; $P < 0.0001$). The negative correlations found here indicate that more central nodes evolve more slowly in all three networks. In the yeast and fly networks, we are able to ask what the independent effects of all three measures of centrality were; these measures are too highly correlated in the worm network to do this analysis. We performed a multiple regression with betweenness, closeness, and connectivity values as effects (betweenness log-transformed and connectivity Box-Cox transformed to satisfy the assumptions of normality) and found that there were still significant, independent effects of both betweenness and connectivity on selective constraint in both the yeast network (betweenness: $F = 11.5$; $P = 0.0007$; connectivity: $F = 8.7$; $P = 0.0033$; closeness: $F = 1.6$; $P = 0.20$) and the fly network (betweenness: $F = 5.9$; $P = 0.014$; connectivity: $F = 4.4$; $P = 0.036$; closeness: $F = 0.1$; $P = 0.77$). For the yeast network, we also estimated the level of expression of individual genes using the codon adaptation index and found that the effects of betweenness are still significant after taking this into account ($F = 27.6$; $P < 0.0001$).

Although proteins with a large number of interactors can be found at the edge of a network, those at the center—regardless of the number of interactors—appear to be more conserved evolutionarily. The more modular a network's structure (e.g., Ravasz et al. 2002), the more important those proteins that lie between modules and control the flow of cellular information become. Because betweenness has effects independent of connectivity on evolutionary rate, the correlation with rate cannot solely be caused by a greater proportion of each of these proteins being involved in direct interactions (cf. Fraser et al. 2002). This independence from connectivity also indicates that the correlation is not caused by any bias toward counting more protein interactions for more abundant proteins (Bloom and Adami 2003). Our results suggest that previous studies that failed to find a correlation between number of interactors and evolutionary rate in a metabolic network (Hahn et al. 2004), in which there are no direct protein–protein interactions, may have missed significant correlations with other measures of centrality; alternatively, it may be that there are simply different evolutionary dynamics in metabolic networks compared with protein-interaction networks.

Using all three networks, we can also ask whether a protein's centrality is informative with respect to its effect on phenotype. Jeong et al. (2001) found that yeast proteins with a larger number of interaction partners were more likely to be lethal when knocked out. We looked for differences in connectivity, betweenness, and closeness between essential and nonessential genes in all three networks. For yeast, we used the knock-out data from Giaever et al. (2002) to ask whether a gene was essential; for worm we used the RNAi knock-down data from multiple studies (Maeda et al. 2001; Kamath et al. 2003) to assess embryonic lethality; and for fly we found all the genes known to have lethal mutants that were cataloged in FlyBase (http://flybase.bio.indiana.edu). For all three organisms, essential genes were more likely to be central in the protein interaction network by any measure of centrality (table 2). Essential genes had higher betweenness, higher connectivity, and higher closeness than nonessential genes (Wilcoxon two-sample test, all $P < 0.001$[table 2]). Again, we can use a multiple regression in the yeast and fly networks to show that betweenness has an effect on the probability of being essential independent of the number of direct protein interactions (yeast: likelihood ratio test $\chi^2 = 20.0$; $P < 0.0001$; fly: likelihood ratio test $\chi^2 = 32.3$; $P < 0.0001$).

It has previously been shown in bacteria (Jordan et al. 2002), yeast (Hirsh and Fraser 2001; Yang, Gu, and Li 2003), and worm (Stein et al. 2003) that essential proteins evolve more slowly than nonessential proteins. If essential genes are more likely to be centrally located in a network, then it is possible that the correlation we observe between evolutionary rate and centrality is a result of this related phenomenon. To test for this effect in our data, we compared the average d_N in network proteins from worm and fly and the average d_N /d_S in proteins from yeast between essential and nonessential genes. We find that essential genes evolve more slowly in all three genomes (all $P < 0.0001$ [table 2]). Our results, thus, agree with previous studies of larger numbers of proteins in the yeast and worm genomes (Yang, Gu, and Li 2003; Stein et al. 2003), and show that this same pattern holds within the protein interaction network of *Drosophila*. If we look within only nonessential genes, however, we find that there is still a correlation between evolutionary rate and betweenness centrality in yeast (Spearman's ρ: -0.12; $P < 0.0001$), worm (Spearman's ρ: -0.10; $P < 0.0001$), and fly (Spearman's ρ: -0.04; $P = 0.022$). These results indicate that the overrepresentation of essential genes in the center of protein interaction networks is not responsible for the correlation between centrality and rate of evolution in any of the networks. Those proteins that are more central to the network, regardless of whether they are essential to the organism, appear to be more constrained by natural selection.

Interestingly, we find a consistent reduction in evolutionary rate for essential proteins in all three species: essential genes in the protein interaction network evolved at 70% the rate of nonessential genes (yeast: 70.5%; worm: 71.4%; fly: 70.1%). The consistent reduction is especially surprising because genes were determined as being essential by knock-out in yeast, knock-down in worm, and an assortment of methods and mutations in fly (including gain-of-function lethals). Although the fly data is necessarily a nonrandom set of genes because researchers have studied them in depth, we believe it is representative of the genome as a whole. Previous data from all of the *C. elegans* proteins subjected to RNAi knock-downs that have orthologs in *C. briggsae*—whether or not they are in the network studied here—also revealed an approximately 30% reduction (29% [Stein et al. 2003]), whereas a much smaller data set from mouse ($n = 141$) shows a 23% reduction in genes that result in high levels of inviability or infertility when knocked out (Hurst and Smith 1999). Data from the bacteria *E. coli* show a much larger reduction in evolutionary rate in essential genes (71% [Jordan et al. 2002]). Although bacteria are haploid organisms and, therefore, may not be able to withstand as many deleterious mutations in essential genes, we know of no biological explanation for there being a consistent 30% reduction among the eukaryotes. The ''70% rule'' may simply be the result of consistent differences in the level of functional constraint between essential and nonessential genes across eukaryotic organisms.

Although there is clearly a difference in the distribution of selection coefficients among mutations between essential and nonessential genes, our results show that the dependence between evolutionary rate and measures of centrality is not driven solely by whether or not a gene is essential. As we showed above, there is a significant relationship between evolutionary rate and betweenness for all three networks when considering only the nonessential genes. We also find this relationship to be significant within only essential genes for yeast (ρ: -0.25; $P < 0.0001$) and fly (ρ: -0.13; $P = 0.0002$); it is nonsignificant in the worm network (ρ: -0.09; $P = 0.14$). Because not every mutation in an ''essential'' gene is lethal, these results demonstrate that there is also a distribution of selective effects among these essential genes that is affected by the centrality of a protein. Therefore, for both essential and nonessential genes, the position of a protein in the interaction network affects the fitness consequences of mutations and, as a result, the rate of evolution.

Our results provide evidence that a protein's position in the interaction network of three eukaryotes has an effect on both its rate of evolution and probability of being essential. Although these effects are relatively small, they appear to be independent of simple relationships with either protein abundance or the amount of a protein's surface needed for protein–protein contacts.

Multiple models assume that adaptation becomes more difficult when a trait is highly constrained (Fisher 1930; Orr 2000; Barton and Keightley 2002). Proteins more central to protein-interaction networks may have more pleiotropic effects on cellular functions (Promislow 2004) and, thus, may be more constrained during evolution. The evidence that these proteins are more likely to be essential and to evolve more slowly—because there are either fewer adaptive mutations or fewer neutral mutations available (Waxman and Peck 1998; Rausher, Miller, and Tiffin 1999)—therefore, supports the assumptions of Fisher's original model of adaptation. Further work comparing protein interaction networks from multiple species will reveal whether the position of homologous proteins is highly conserved in evolution or whether new proteins can be readily co-opted into the center of networks.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online (www.molbiolevol.org).

Acknowledgments

We thank D. Begun, C. Jones, J. Mezey, and L. Moyle for helpful discussion and comments. We also thank J. Stajich and K. Thornton for sharing their data. M.W.H. is supported by an Integrative Informatics postdoctoral fellowship from the National Science Foundation. A.D.K. is supported by a predoctoral fellowship from the Howard Hughes Medical Institute.

Literature Cited

Almaas, E., B. Kovacs, T. Vicsek, Z. N. Oltvai, and A.-L. Barabasi. 2004. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. Nature **427**:839–843.

Barton, N. H., and P. D. Keightley. 2002. Understanding quantitative genetic variation. Nat. Rev. Genet. **3**:11–21.

Bloom, J. D., and C. Adami. 2003. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. BMC Evol. Biol. **3**:21.

Fisher, R. A. 1930. The genetical theory of natural selection. The Clarendon Press, Oxford.

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

Giaever, G., A. M. Chu, L. Ni et al. (63 co-authors). 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature **418**:387–391.

Giot, L., J. S. Bader, C. Brouwer et al. (39 co-authors). 2003. A protein interaction map of *Drosophila melanogaster*. Science **302**:1727–1736.

Hahn, M. W., G. C. Conant, and A. Wagner. 2004. Molecular evolution in large genetic networks: Does connectivity equal constraint? J. Mol. Evol. **58**:203–211.

Hirsh, A. E., and H. B. Fraser. 2001. Protein dispensability and rate of evolution. Nature **411**:1046–1049.

Hurst, L. D., and N. G. C. Smith. 1999. Do essential genes evolve slowly? Curr. Biol. **9**:747–750.

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

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

Jordan, I. K., I. B. Rogozin, Y. I. Wolf, and E. V. Koonin. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. **12**:962–968.

Kamath, R. S., A. G. Fraser, Y. Dong et al (13 co-authors). 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature **421**:231–237.

Krylov, D. M., Y. I. Wolf, I. B. Rogozin, and E. V. Koonin. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. **13**:2229–2235.

Li, S. M., C. M. Armstrong, N. Bertin et al. (38 co-authors). 2004. A map of the interactome network of the metazoan *C. elegans*. Science **303**:540–543.

Maeda, I., Y. Kohara, M. Yamamoto, and A. Sugimoto. 2001. Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. Curr. Biol. **11**:171–176.

Orr, H. A. 2000. Adaptation and the cost of complexity. Evolution **54**:13–20.

Promislow, D. E. L. 2004. Protein networks, pleiotropy and the evolution of senescence. Proc. R. Soc. Lond. B Biol. Sci. **271**:1225–1234.

Rausher, M. D., R. E. Miller, and P. Tiffin. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol. Biol. Evol. **16**:266–274.

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

Stein, L. D., Z. R. Bao, D. Blasiar et al. (26 co-authors). 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. PLoS Biol. **1**: 166–192.

Wagner, A. 2001. The yeast protein interaction network evolves rapidly and contains few duplicate genes. Mol. Biol. Evol. **18**:1283–1292.

Wagner, A., and D. Fell. 2001. The small world inside large metabolic networks. Proc. R. Soc. Lond. B Biol. Sci. **280**:1803–1810.

Wasserman, S., and K. Faust. 1994. Social network analysis. Cambridge University Press, Cambridge.

Waxman, D., and J. R. Peck. 1998. Pleiotropy and the preservation of perfection. Science **279**:1210–1213.

Yang, J., Z. L. Gu, and W. H. Li. 2003. Rate of protein evolution versus fitness effect of gene deletion. Mol. Biol. Evol. **20**: 772–774.

Michelle Vendruscolo, Associate Editor

Accepted December 20, 2004