Molecular Evolution in Large Genetic Networks: Connectivity Does not Equal Importance

Matthew W. Hahn
Gavin C. Conant
Andreas Wagner

SFI WORKING PAPER: 2002-08-039

SANTA FE INSTITUTE

Molecular evolution in large genetic networks: connectivity does not equal importance

Matthew W. Hahn,[†1] Gavin Conant, [†2] and Andreas Wagner[2,3]

[†] These authors contributed equally to the reported work. [1]Department of Biology, Duke University, Durham, North Carolina, [2]Department of Biology, University of New Mexico, and [3]The Santa Fe Institute, Santa Fe, New Mexico

Corresponding author: M. W. Hahn, Department of Biology, Duke University, Box 90338, Durham, NC 27708, USA. Tel: (919)684-2295; Fax: (919)660-7293; Email: mwh3@duke.edu

Biological networks are extremely robust to many different perturbations[1-6]. Both drastic environmental changes and loss-of-function mutations in a variety of genes often have no detectable effect on the phenotype of an organism, even if the lost gene is considered *a priori* to be important for cellular activity[7-9]. This robustness may be a function of a network's architecture: power-law distributions of the number of interactors observed in genetic networks have been hypothesized to confer robustness against mutations[3]. We evaluate this hypothesis for two genetic networks, that of the *E. coli* core intermediary metabolism and that of the yeast protein interaction network. Specifically, we test the hypothesis through one of its key predictions: highly connected proteins should be more important to the cell, and thus be subject to severe selective and evolutionary constraints. We find that highly connected proteins can tolerate just as many amino acid substitutions as other proteins, and thus conclude that power-laws in cellular networks do not reflect selection for mutational robustness.

Recent advances in the mathematical theory of random graphs[2,10] have led to an explosion of research concerned with the architecture of biological networks[2,3,5,6,11-15]. This research has shown that the topology of many biological networks, including metabolic networks[13,15], and protein-interaction networks[6,14] share two important features. First, they have a small diameter, *L*, defined as the shortest path between network nodes, averaged over all nodes. Second, the frequency, *P(d)*, of nodes with *d* immediate

neighbors follows a power law, i.e., $P(d) \sim d^{-\gamma}$ with a constant γ characteristic of the network[13,15].

The discovery of power-laws and small-worldness has given rise to an intriguing hypothesis relating large-scale network structure to mutational robustness[3]. This hypothesis rests on the observation that random node removal leaves the diameter of networks with power-law connectivity largely unchanged. In a biological network, node removal corresponds to synthetic-null or gene-knockout mutations. In contrast to networks with power-law connectivity, more homogeneous random networks respond to node removal with a rapidly increasing diameter[3].

The striking structural stability has lead to the suggestion that we observe networks with power-law connectivity in cells *because* of their robustness to random node removal[3,13].

Do small network diameters matter to cells? A possible advantage of small mean path lengths in metabolic networks stems from the importance of minimizing transition times between metabolic states in response to environmental changes[4]. Metabolic networks with small diameters thus might adjust more rapidly to environmental change. Answering this question by direct experimentation, however, is currently impossible, for doing so would require comparing biological networks of different large-scale structure *in vivo*. Absent direct experimental tests showing whether genetic network diameter matters to organisms, one can still test key predictions of this hypothesis. One such prediction is that removing highly connected genes in a network is more detrimental than removing less connected genes. Evidence for this prediction was provided using knockout data and the protein-interaction network from yeast, where Jeong et al [6] found

that highly-connected proteins are much more likely to be essential for survival than less-connected proteins.

Using gene-knockout data to assess the importance of highly connected genes or proteins, however, has disadvantages. First, even apparently neutral knockouts may have subtle but undetectable fitness defects. In the huge populations characteristic of microbes like yeast, growth rate differences of 10^{-6} between mutant and wild-type may be evolutionarily important, but chemostat experiments have difficulty resolving differences smaller than 10^{-3} (ref. 16). Second, laboratory experiments assess fitness differences only in a small number of environments and not over the entire environmental spectrum important for life in the wild. A neutral knock-out mutation in one environment may be lethal in other environments.

Better alternatives in testing this prediction would use an evolutionary record of mutations experienced over millions of years, mutations whose effects manifested themselves in the environments the organism experienced in its evolution. Such an alternative is available in the ratio of nonsynonymous to synonymous substitutions per nucleotide site (K_a/K_s)[17]. It indicates how likely it is that an amino acid change has deleterious effects. Generally, $K_a/K_s<1$, and the smaller it is, the fewer amino acid substitutions a protein tolerates. Here we use this ratio to test the hypothesis that more highly connected genes can tolerate fewer mutations in two genetic networks: the *Escherichia coli* metabolic network[15] and the *S. cerevisiae* protein-interaction network[14,18].

The *E. coli* core metabolic network encompasses the catabolic and biosynthetic metabolism central to the cell's function. Wagner and Fell[15] constructed a graph

representation of this network from stoichiometric equations in which two genes are connected by a vertex if the chemical reactions their products catalyze share at least one substrate. In order to measure K_a/K_s ratios for the genes found in the *E. coli* network, we identified similar genes in the closely related *Haemophilus influenzae* genome, as outlined in the Methods. Of the 133 genes in the *E. coli* metabolic network with related genes in *H. influenzae*, 106 have no paralogs (matches with amino acid identity over 35%) in either species. To avoid ambiguities due to mistaking gene duplicates for truly orthologous genes, we used only these bona fide orthologous genes in the following analysis.

While the selective constraint, as measured by K_a/K_s, displayed a broad range across enzymes ($0.014< K_a/K_s <0.159$), it showed no statistical association with the degree of connectedness using Pearson's r (Fig. 1a; r=-0.18, **P**=0.18; n=57), Spearman's rank correlation coefficient ρ indicated a weak such association (ρ=-0.35, **P**=0.008; n=57). However, we note that this association becomes marginal when two outlier nodes having the highest values of K_a/K_s but low connectivity are removed from the analysis. (Spearman's ρ= -0.29, **P**=0.031; n=55). Amino acid distance (K_a) alone shows the same qualitative association with the number of interactors.

In the *S. cerevisiae* protein-interaction network, two proteins are neighbors if they physically interact *in vivo*. We here use available two-hybrid protein interaction data[18] to indicate neighboring proteins. Unfortunately, the fully sequenced eukaryotic genomes of *Schizosaccharomyces pombe* and *Neurospora crassa* are too distantly related to budding yeast to carry out our analysis. We thus estimated selective constraints on yeast genes through their closest paralogues (duplicates) within the yeast genome. Note that using

K_a/K_s (and not just amino acid divergence K_a) in the analysis compensates for duplicates of different age. Figure 1b shows that no statistical association exists between protein connectivity and evolutionary constraint in yeast (Pearson's r=0.0075, **P**=0.95, Spearman's ρ=0.083, **P**=0.57; $0.006<K_a/K_s<0.438$).

There are a few caveats to these results. First, K_a/K_s may differ among proteins simply because of differences in structure. Other candidate factors determining selective constraints include physical position in the genome[19], expression level[20], and "essentialness"[21]. This makes it difficult to compare evolutionary constraints between two proteins, but in our statistical analysis of many proteins such effects should merely be noise distorting the expected pattern. Second, the protein-interaction data we use contains much experimental noise[18,22]. However, the power-law connectedness of protein interaction networks is identical in different high-throughput data sets generated with the same method, the yeast two-hybrid assay, and is also observed for protein interaction data obtained with techniques other than this assay[23]. Moreover, highly connected proteins in one data set are also highly connected in another. We are thus confident that the connectivity distribution we use is not an artifact of one experimental technique.

As a third caveat, we cannot completely exclude the possibility that many genes in the *E.coli* metabolic network have acquired their functions (and numbers of network neighbors) very recently, long after the divergence of *E. coli* from *H. influenzae*. In this case, their K_a/K_s ratio may reflect past rather than present function. However, the *H. influenzae* core metabolic network is probably very similar to that of *E. coli*, not only because of their close evolutionary relationship, but also because core metabolism is nearly universal among free-living non-extremophiles[24-26]. In addition, 56% (133) of the

genes in the *E.coli* network have a similar gene in *H. influenzae*. This percentage is much larger than the approximately 26% of genes shared overall between the two species[27].

Our results, as well as recent results of others[28], show that the outcomes of gene-knockout experiments have to be taken with extreme caution when inferring the "importance" of a gene. They also demonstrate that a gene's position in a network alone may have no bearing on its importance as defined through evolutionary constraints. And they refute the claim that power-law connectivity in cellular networks reflects selection for robust network diameters. What, then, can we learn from purely qualitative, topological analysis of genetic networks? The work of Rausher et al[29] on genes responsible for anthocyanin biosynthesis raises the possibility that a gene's role in controlling flux through a metabolic pathway may determine its rate of evolution. This suggests that a gene's position in a network, although uninformative on its own, may become informative when supplemented by additional biological information.

Methods

In order to estimate K_a/K_s ratios for the genes found in the *E. coli* network, we searched the *Haemophilus influenzae* genome for similar genes via BLASTP[30] and retained pairs matching at $E<1x10^{-7}$. We then aligned the retained pairs with ClustalW[31] and used the 106 pairs of genes in the metabolic network with over 35% amino acid identity and no paralogues to calculate K_a and K_s by maximum likelihood estimation under the models of Muse and Gaut and Goldman and Yang[32-34] ($0.111<K_a<0.731$, $2.58<Ks<6$).

For the yeast data gene pairs with a BLASTP E-value of $<10^{-8}$ and percent amino acid identity of 40% were used. K_a/K_s was calculated in the same manner as in the bacterial dataset. For the analysis shown, we used a total of 48 genes in the network that met the following criteria: $10^{-4}<K_s<4$, $10^{-4}<K_a<1$, $0.006<K_a/K_s<0.438$. Use of data points with large values of K_s is often viewed as problematic because the variance in estimates of K_s scales with the K_s value itself[35]. However, results of simulated sequence evolution (not shown) based on the 48 yeast duplicates used indicate that, within this range of K_s values, variance in K_a/K_s is not significantly correlated with K_s (Pearson's *r*: 0.085, **P**=0.57, Spearman's *s*: 0.249, **P**=0.095). This indicates that reporting K_a/K_s ratios even for large values of K_s does not lead to unacceptably large errors.

Acknowledgements

MWH thanks M. Rausher, M. Rockman, M. Rutter, and R. Zufall for comments and suggestions; an NIH training grant administered by the Duke University Program in Genetics provided support. GC is supported by the Department of Energy's Computational Sciences Graduate Fellowship program, administered by the Krell Institute. AW acknowledges financial support through NIH grant GM63882 and the Santa Fe Institute.

References

1. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47-C52 (1999).
2. Watts, D.J. *Small worlds*, (Princeton University Press, Princeton, New Jersey, 1999).
3. Albert, R., Jeong, H. & Barabasi, A.L. Error and attack tolerance of complex networks. *Nature* **406**, 378-382 (2000).
4. Edwards, J.S. & Palsson, B.O. The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 5528-5533 (2000).
5. Wagner, A. Mutational robustness in genetic networks of yeast. *Nature Genetics* **24**, 355-361 (2000).
6. Jeong, H., Mason, S.P., Barabasi, A.-L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41-42 (2001).
7. Smith, V., Chou, K.N., Lashkari, D., Botstein, D. & Brown, P.O. Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274**, 2069-2074 (1996).
8. Winzeler, E.A. et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901-906 (1999).
9. Ross-Macdonald, P. et al. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413-418 (1999).

10. Watts, D.J. & Strogatz, S.H. Collective dynamics of small-world networks. *Nature* **393**, 440-442 (1998).

11. Barabasi, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509-512 (1999).

12. Bhalla, U.S. & Iyengar, R. Emergent properties of networks of biological signaling pathways. *Science* **283**, 381-387 (1999).

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

14. Wagner, A. The yeast protein interaction network evolves rapidly and contains few duplicate genes. *Molecular Biology and Evolution.* **18**, 1283-1292 (2001).

15. Wagner, A. & Fell, D. The small world inside large metabolic networks. *Proceedings of the Royal Society London Series B* **280**, 1803-1810 (2001).

16. Hartl, D.L. & Clark, A.G. *Principles of Population Genetics*, (Sinauer associates, Sunderland Mass., 1988).

17. Hughes, A.L. *Adaptive Evolution of Genes and Genomes*, (Oxford University Press, Oxford, 1999).

18. Uetz, P. et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623-627 (2000).

19. Williams, E.J.B. & Hurst, L.D. The proteins of linked genes evolve at similar rates. *Nature* **407**, 900-903 (2000).

20. Akashi, H. Gene expression and molecular evolution. *Current Opinion in Genetics & Development* **11**, 660-666 (2001).

21. Hurst, L.D. & Smith, N.G.C. Do essential genes evolve slowly? *Current Biology* **9**, 747-750 (1999).

22. Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America ;* **98**, 4569-4574 (2001).

23. Mewes, H.W. et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Research* **27**, 44-48 (1999).

24. Tatusov, R.L. et al. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Current Biology* **6**, 279-291 (1996).

25. Edwards, J.S. & Palsson, B.O. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *Journal of Biological Chemistry* **274**, 17410-17416 (1999).

26. Morowitz, H.J. *Beginnings of cellular life*, (Yale University Press, New Haven, 1992).

27. Blattner, F.R. et al. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1462 (1997).

28. Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. & Feldman, M.W. Evolutionary rate in the protein interaction network. *Science* **296**, 750-752 (2002).

29. Rausher, M.D., Miller, R.E. & Tiffin, P. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Molecular Biology and Evolution* **16**, 266-74. (1999).

30. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410 (1990).

31. Thompson, J.D., Higgins, D.G. & Gibson, T.J. Clustal-W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673-4680 (1994).

32. Muse, S.V. & Gaut, B.S. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**, 715-724 (1994).

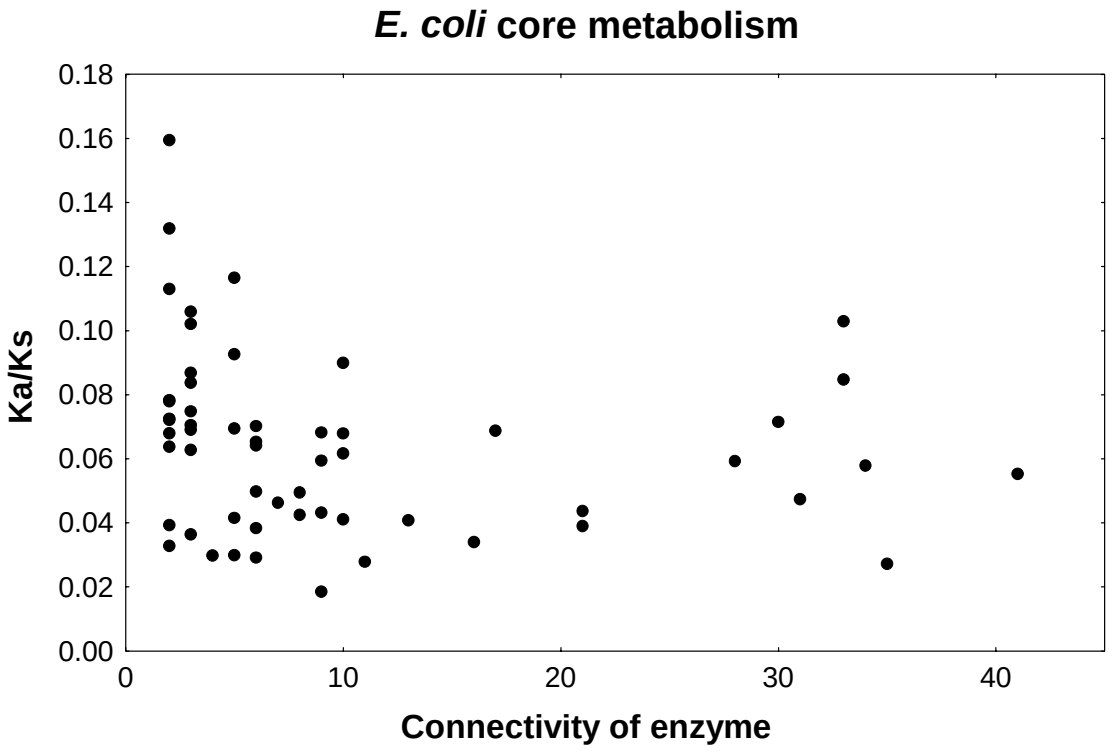
33. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**, 725-736 (1994).

34. Yang, A. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**, 32-43 (2000).

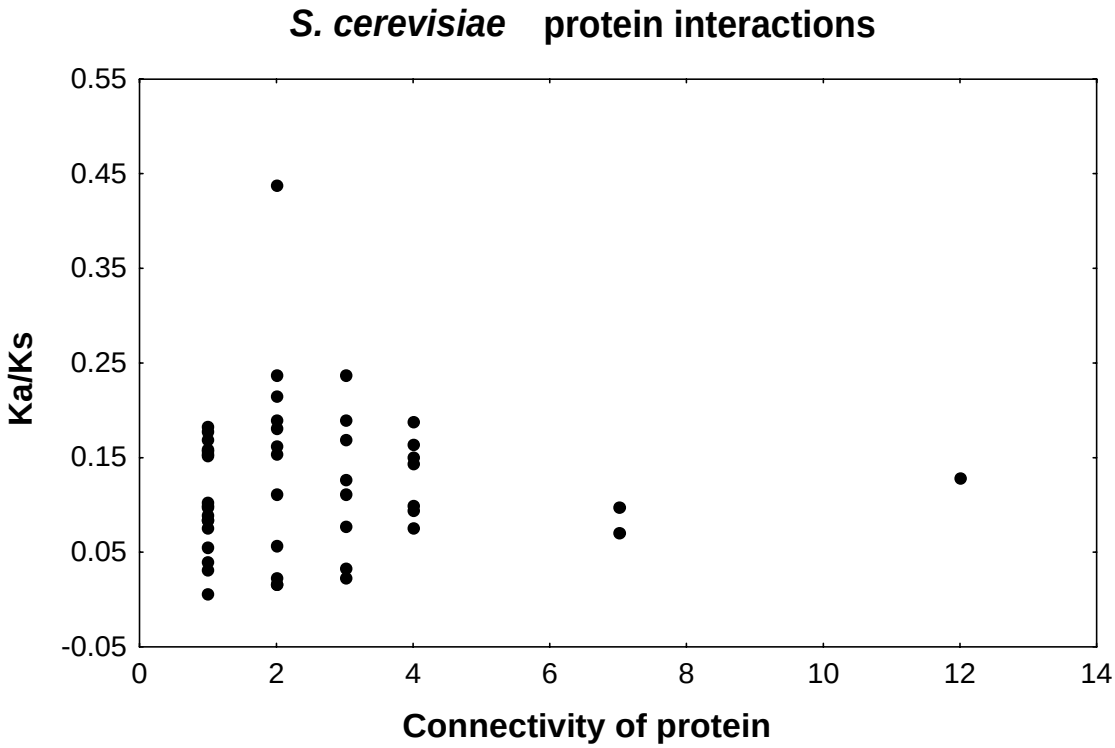
35. Li, W.-H. *Molecular evolution*, (Sinauer Associates, Sunderland, Mass., 1997).

Fig. 1 a) Statistical association between selective constraint, K_a/K_s, and connectivity (number of neighbors) of 57 enzymes in the core intermediary metabolic network of *E. coli* (Pearson's r=-0.18, **P**=0.18; Spearman's ρ=-0.35, **P**=0.008) (see text) **b)** Statistical association between K_a/K_s, and connectivity (number of neighbors) of 48 proteins in the yeast protein interaction network[18] (Pearson's r=0.0075, **P**=0.95, Spearman's s=0.083, **P**=0.57).

a)


E. coli core metabolism
Ka/Ks
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
Connectivity of enzyme

b)


S. cerevisiae protein interactions
Ka/Ks
0.55
0.45
0.35
0.25
0.15
0.05
-0.05
0
2
4
6
8
10
12
14
Connectivity of protein