J Mol Evol (2004) 58:203–211
DOI: 10.1007/s00239-003-2544-0

JOURNAL OF MOLECULAR EVOLUTION

Molecular Evolution in Large Genetic Networks: Does Connectivity Equal Constraint?

Matthew W. Hahn,[1] Gavin C. Conant,[2] Andreas Wagner[2]

[1] Department of Biology, Box 90338, Duke University, Durham, NC 27708, USA
[2] Department of Biology, 167 Castetter Hall, University of New Mexico, Albuquerque, NM 87131, USA

Received: 20 December 2002 / Accepted: 6 September 2003

Abstract. Genetic networks show a broad-tailed distribution of the number of interaction partners per protein, which is consistent with a power-law. It has been proposed that such broad-tailed distributions are observed because they confer robustness against mutations to the network. 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, subject to more severe selective and evolutionary constraints. We find, however, that no correlation between highly connected proteins and evolutionary rate exists in the *E. coli* metabolic network and that there is only a weak correlation in the yeast protein-interaction network. Furthermore, we show that the observed correlation is function-specific within the protein-interaction network: only genes involved in the cell cycle and transcription show significant correlations. Our work sheds light on conflicting results by previous researchers by comparing data from multiple types of protein-interaction datasets and by using a closely related species as a reference taxon. The finding that highly connected proteins can tolerate just as many amino acid substitutions as other proteins leads us to conclude that power-laws in cellular networks do not reflect selection for mutational robustness.

Key words: Power-law — Mutational robustness — Selective constraint — Genetic network

Introduction

Recent advances in the mathematical theory of random graphs (Watts 1999; Watts and Strogatz 1998) have led to an explosion of research concerned with the architecture of biological networks (Albert et al. 2000; Barabasi and Albert 1999; Bhalla and lyengar 1999; Jeong et al. 2000, 2001; Wagner 2000, 2001; Wagner and Fell 2001; Watts 1999). This research has shown that the topology of many biological networks, including metabolic networks (Jeong et al. 2000; Wagner and Fell 2001) and protein-interaction networks (Jeong et al. 2001; Wagner 2001), 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 is a broad-tailed distribution consistent with a power-law, i.e., $P(D) \sim D^{-\gamma}$, with a constant γ characteristic of the network (Jeong et al. 2000; Wagner and Fell 2001).

The discovery of power-laws and small-worldness has given rise to an intriguing hypothesis relating large-scale network structure to mutational robustness (Albert et al. 2000). This hypothesis rests on the observation that random node removal leaves the diameter of networks with power-law connectivity

The first two authors contributed equally to the reported work.
Correspondence to: Matthew W. Hahn; *email:* mwhahn@ucdavis.edu

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 (Albert et al. 2000). This striking structural stability has led to the suggestion that we observe networks with power-law connectivity in cells *because* of their robustness to random node removal (Albert et al. 2000; Jeong et al. 2000). However, power-law networks are extremely dependent on the small number of highly connected nodes: removal of these can lead to huge changes in network architecture and diameter.

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 (Edwards and Palsson 2000). 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 highly connected proteins should be more important to the cell and, thus, be subject to severe selective and evolutionary constraints. One way to test this prediction is to study the rate of evolution of proteins in these networks. We examine this rate in two genetic networks: the *Saccharomyces cerevisiae* protein-interaction network (Ito et al. 2001; Uetz et al. 2000; Wagner 2001) and the *Escherichia coli* metabolic network (Wagner and Fell 2001).

Several recent papers have studied the relationship between a gene product's position in protein-interaction networks and its rate of evolution. Using 164 genes in the yeast protein-interaction network and their *Caenorhabditis elegans* orthologs, Fraser and coauthors (2002) found a weak negative correlation between protein distance and the degree of connectedness of network proteins. Two recent analyses used the more closely related yeast *S. pombe* as a reference taxon but different protein-interaction data to construct networks (Fraser et al. 2003; Jordan et al. 2003a). Fraser and coauthors again found a significant, though weak, negative correlation between network degree and rate of evolution. Jordan and coauthors, however, found no significant correlation in either their original analysis or their later corrected analysis (Jordan et al. 2003b). Below, we attempt to resolve this controversy using several reference taxa, including a very recent ancestor of *S. cerevisiae*, *S. paradoxus* (Kellis et al. 2003), as well as multiple types of protein-interaction data.

Materials and Methods

Constructing Cellular Networks

Wagner and Fell (2001) represented the *E. coli* core metabolic network as a graph in which two proteins are connected if a reaction product of one is a substrate for the other. In other words, proteins are connected if they share at least one metabolite. We define the connectedness, D, of a protein in this network as the number of proteins it is connected to. The resulting metabolic network is comprised of 237 distinct proteins with a range of connectedness, D, from 1 to 46 (Wagner and Fell 2001).

In our analysis of the *Saccharomyces cerevisiae* protein-interaction network, our primary dataset (hereafter Pooled Data I) consists of three independent, large-scale datasets on pairwise protein interactions, two of which were generated using the yeast two-hybrid assay (Ito et al. 2001; Uetz et al. 2000). The third dataset was generated using various non-two hybrid techniques (Mewes et al. 1999). In our graph representation of this network, nodes (proteins) are connected to other proteins if they interact *in vivo* according to the experimental data. We define the connectedness, D, of a protein as its number of protein interaction partners. The protein-interaction data we use contain enormous amounts of experimental noise (Ito et al. 2001; Uetz et al. 2000; von Mering et al. 2002). However, the relevant feature of the network, the distribution of the number of interaction partners per protein, is robust. It is identical in all three datasets, and highly connected proteins in one dataset are also highly connected in the others (Wagner 2002). We are thus confident that the broad-tailed connectivity distribution is not an artifact of experimental technique.

For comparative purposes, we also study the filtered interaction dataset of von Mering and coauthors (2002; hereafter Pooled Data II), where we have accepted only interactions annotated with "medium" or "high" confidence by these authors. This is a dataset very similar to that of Fraser and coauthors (2003) and includes information on pairwise protein interactions, as well as information on the composition of purified protein complexes (Gavin et al. 2002; Ho et al., 2002). We also pooled two datasets from the high-throughput spectrometric identification of purified protein complexes (Gavin et al. 2002; Ho et al., 2002) for a third dataset, which is a subset of Pooled Data II. Self-interactions were excluded from the data of Gavin and coauthors, and we used the filtered data presented by Ho and coauthors.

To compare the degree distributions of the protein complex data and Pooled Data II datasets to that of Pooled Data I, we used a maximum likelihood approach to fit each dataset to the continuous power-law distribution described by the probability density function:

$$p(x) \propto x^{-b}, \quad 1 \leq x < \infty, \quad b > 1 \qquad (1)$$

To determine the significance of the differences in the power-law exponent b seen between datasets, we employed a likelihood ratio test. The test is based on comparing the likelihood of Pooled Data II and protein complex data under the constraint that these two datasets have the same degree distribution as Pooled Data I, to the maximum likelihood estimate of the degree distribution for the two datasets. In terms of Eq. (1), we estimate the maximum likelihood value of b for Pooled Data I, referred to hereafter as b_{I}. We then calculate the likelihood L_{c} of observing the Pooled Data II and the protein complex data with b fixed at b_{I}. This constrained likelihood will never be greater than the likelihood L of observing either dataset when b is allowed to take on its maximum likelihood value. The ratio of the likelihoods L/L_{c} between the unconstrained (L) and the constrained (L_{c}) models is a measure of whether Pooled Data II and the protein complex data follow a different degree distribution than does Pooled Data I. To gauge the statistical significance of L/L_c, we first simulate 10,000 sample degree distributions where $b = b_I$. Each of these simulations has the same number of data points as either Pooled Data II or the protein complex data but follows the degree distribution of Pooled Data I. We then calculate L/L_c for these simulated datasets (because the simulations have finite size, the maximum likelihood estimate [MLE] for b in a simulation will almost never be exactly b_I). Performing this calculation for all 10,000 simulations yields a distribution of L/L_c, which indicates how much L may differ from L_c due only to chance. If L/L_c for our real data does not fall within the range of values seen in the simulations, then the two datasets have significantly different distributions.

We discuss results in detail for Pooled Data I. However, we note that our results are qualitatively identical when separately considering each of the datasets contained within Pooled Data I.

Estimating Evolutionary Rate

We were interested in whether the connectivity (number of interaction partners) of a gene's product affects that gene's rate of sequence evolution. We used two related approaches to study this question. First, we identified genes in the yeast network with closely related duplicates and used the ratio K_a/K_s of amino acid replacement to silent substitutions (Kimura 1977) to measure selective constraint. Specifically, we searched the genome of *S. cerevisiae* (Goffeau et al. 1996) for the nearest paralog (if any) of each gene in the yeast protein-interaction network (an intragenome search). The *E. coli* genome (Blattner et al. 1997) has too few duplicate genes for an intragenome analysis. Our second approach uses reference genomes to find orthologs of network genes in both the *E. coli* and the yeast networks. In *S. cerevisiae* we identified orthologous genes in the genomes of *Schizosaccharomyces pombe* (Wood et al. 2002) and *Saccharomyces paradoxus* (Kellis et al. 2003). For *E. coli*, we used *Haemophilus influenzae* (Fleischmann et al. 1995) as the reference species. We estimated the value of K_a, the rate of amino acid replacement substitutions for each pair of orthologs, allowing us to determine the correlation between K_a and protein degree. Because *S. paradoxus* and *S. cerevisiae* are very closely related (Kellis et al. 2003), we were also able to estimate the synonymous distance (K_s) between orthologs for this comparison. To avoid mistaking gene duplicates for truly orthologous genes in all these between-genome comparisons, we retained only gene pairs where neither species contained any duplicates.

We used our previously described tool, GenomeHistory (Conant and Wagner 2002), to identify orthologs and paralogs. GenomeHistory uses gapped BLASTP (Altschul et al. 1997) to identify related genes which are globally aligned using ClustalW (Thompson et al. 1994). K_a and K_s are calculated by maximum likelihood estimation under the model of Goldman and Yang (1994). We accepted all BLAST hits with E-values less than 1×10^{-8} for the *S. cerevisiae–S. pombe*/*S. paradoxus* comparison and 1×10^{-7} for the *E. coli–H. influenzae* comparison. We further required 40% global amino acid identity between any two putative orthologs or paralogs. For the estimation of K_a and K_s, we removed gap residues, calculated base frequencies separately at each codon position, and estimated the transition/transversion ratio from the data.

Using the identified paralogs and orthologs, we calculated the correlation between evolutionary distance and the degree of protein connectivity (D). For the *S. cerevisiae* paralogs and orthologs in *S. paradoxus*, we calculated the correlation coefficients (both Pearson and Spearman) between K_a/K_s for the closest paralog and D using only unsaturated duplicate pairs with $K_s < 3$. For the remaining two sets of orthologs (*S. pombe* and *H. influenzae*) we calculated the correlation coefficients between K_a and D (using only pairs where $K_a < 1$). We were unable to use K_a/K_s ratios for the latter two analyses, as most K_s values showed saturation (see below). Excluding pairs with high divergence values ($K_s > 3$, $K_a > 1$) not only avoids the high variances associated with estimating such values, but also serves to exclude spurious gene pairs where homology is uncertain.

To determine if a relationship between selective constraint and degree of connectivity of yeast proteins depended upon protein function, we used a previously described high-level functional classification of the yeast genome (Conant and Wagner 2002) based on the Gene Ontology database (The Gene Ontology Consortium 2000). This classification system groups annotations into 10 high-level categories, which we have further reduced into five overall groups: (1) All Metabolism (includes Energy, Metabolism, and Protein Metabolism),(2) Cell Cycle/DNA Processing, (3) Cytoskeleton, (4) All Transcription (includes Transcription and Transcription Factors), and (5) Transport. Using the *S. paradoxus* orthologs, we calculated the statistical association within these five groups between D and K_a/K_s. In addition, we repeated the above intra- and intergenome analyses excluding ribosomal proteins and found that so doing did not change our results (not shown).

Significance tests for Pearson's product–moment correlation require normally distributed data, an assumption violated by the broad-tailed distribution of the network degree, D (Wagner 2001; Wagner and Fell 2001). We thus adopted a permutation approach to test for significant correlations. Degree observations were randomly reordered with respect to evolutionary distances and the resulting Pearson and Spearman correlations calculated. The correlations from the original data were then compared to a distribution of correlations obtained from 10,000 permutations of those data.

Identifying Gene Pairs with Saturated Synonymous Substitutions

Saturation is a problem when estimating the fraction of synonymous substitutions, K_s, that occurred between two genes. Saturation occurs when every synonymous site has undergone at least one substitution, and it makes accurate estimates of K_s impossible. However, because K_s gives only the *average* number of synonymous substitutions per synonymous site, it is misleading to speak of saturation occurring at a certain value of K_s (such as $K_s = 1$). Instead, one must also account for the degree of variation among sites and the sequence length. Analytical calculation of the probability of saturation is possible for simple evolutionary models such as the Jukes–Cantor model (Li 1997) but cannot be done easily for the codon-based models needed to calculate K_s. Instead, we use a heuristic saturation test. It relies on the fact that we have obtained maximum likelihood estimates of (MLEs) K_s, such that (for unsaturated sequences) no other value of K_s will give a higher likelihood than the MLE. Saturated sequences may have undergone an arbitrary number of substitutions and thus have no unique MLE value of K_s. Put differently, under saturation there should be no *decrease* in likelihood if the divergence (K_s) of two sequences *increases* beyond the MLE estimate. This concept is illustrated in Fig. 1. In practice, two distinct cases of saturation must be distinguished. First, if two sequences are "exactly" saturated (such as genes whose sequence identity equals exactly 25%), then the likelihood, L, no longer changes past some critical value K_s. Second, if the sequences are "oversaturated," such as when a (chance) nucleotide identity of less than 25% is observed, then L will asymptotically increase as K_s approaches infinity (see Fig. 1B). We take advantage of these facts by calculating, for each sequence pair, the MLE K_s value (determined by numerical optimization and therefore always finite), as well as the likelihood of a K_s value 10 times larger. To use our numerical analysis method, we make the implicit assumption that the likelihood function is free of local maxima in the neighborhood of the MLE. If the likelihood of $(10K_s)$ is greater than or equal to the likelihood of K_s, then we call a sequence pair saturated and exclude it from further analysis.

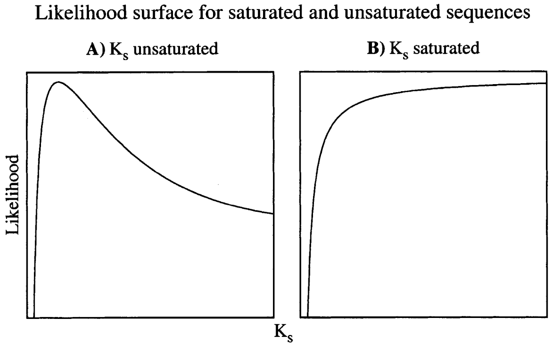

Fig. 1. Diagrammatic view of the principle upon which our heuristic test for saturation in K_s is based. **A** In cases where K_s is not saturated, there exists a distinct likelihood maximum which can be identified numerically. **B** When synonymous substitutions have saturated, the likelihood will remain constant or actually increase (the illustrated case) as K_s approaches infinity.

Results and Discussion

The Yeast Protein-Interaction Network

In the *S. cerevisiae* protein-interaction network, two proteins are neighbors if they physically interact *in vivo*. Our primary dataset consists of data (Pooled Data I; see also Materials and Methods) on such pairwise interactions from large-scale two-hybrid experiments (Ito et al. 2001; Uetz et al. 2000) and from experiments not using the two-hybrid approach (Mewes et al. 1999). We first calculated K_a/K_s (the ratio of amino acid replacement to silent substitutions per site) between each network protein and its closest paralog in the yeast genome. K_a/K_s is the most reliable indicator of selective constraint on protein evolution because it accounts for mutation rate variation among genes. For the paralog comparison, we used a total of 65 gene pairs, with $1 \leq D \leq 13$ and $0.006 \leq K_a/K_s \leq 0.74$. Figure 2A shows that there is a weakly significant Pearson's correlation between protein degree D and K_a/K_s but no significant Spearman's correlation (Pearson's r, −0.187, p = 0.047; Spearman's s, −0.151, p = 0.12). Using paralogs for this analysis has shortcomings, including the fact that paralogs may diverge at different rates and that recent duplicates (rare in our data) diverge at rates different than old duplicates (Lynch and Conery 2000). In addition, the range of D values in this dataset is quite small.

To alleviate these shortcomings, we carried out complementary analyses using not paralogous but orthologous genes from various species. We first used the recently published genome sequence of the yeast species *S. paradoxus*, which is closely related to *S. cerevisiae* (Kellis et al. 2003). Use of this species allows us to compare protein connectivity, D, to the ratio K_a/K_s for many more orthologous genes than the 65 paralogs above. Specifically, we identified 1393 orthologous gene pairs with protein-interaction information from Pooled Data I. In these data, D ranges between 1 and 76, while K_a/K_s ranges between 0.003 and 0.64. There was a weak but highly significant correlation between D and the selective constraint (K_a/K_s) experienced by a gene (r, −0.10; s, −0.11; p < 0.0001 for both) (Fig. 2B and Table 1). (Comparing just K_a to D in *S. paradoxus* using Pooled Data I produces a similar Spearman correlation: s, −0.10.)

We also used a reference taxon other than *S. paradoxus* to compare our results with data recently presented by other researchers. Fraser and coauthors (2002, 2003), using either *C. elegans* or *S. pombe* as a reference taxon, found a weak but significant negative correlation between amino acid distance and D.These results are in contrast to those of Jordan and coauthors (2003a, b), who found no such correlation. We examined the relationship between D and amino acid distance (measured by K_a) in different protein-interaction datasets with *S. pombe* as a reference. Using our Pooled Data I, which includes no data from purified protein complexes, we found no relationship between D and K_a (r, −0.02 [p = 0.27]; s, 0.04 [p = 0.17]; n = 605), a result in accord with Jordan and coauthors, who used a similar dataset. In contrast to these authors, Fraser and coauthors (2003) used data from multiple experimental methods including mass spectrometry of purified protein complexes in their analyses. We analyzed a similar dataset from von Mering and collaborators (2002; Pooled Data II), which contains both data on pairwise protein-interactions and data from protein complexes. Analyzing 1175 gene pairs from *S. cerevisiae* and *S. pombe* with this dataset, we indeed found a correlation similar in magnitude to that obtained by Fraser and coauthors. (r, −0.21; s, −0.22; p < 0.0001 for both).

We are thus able to explain the discrepancy in results between Jordan and coauthors (2003a, b) and Fraser and coauthors (2002, 2003) by using slightly different protein-interaction datasets. The reference taxon used by both of these groups, *S. pombe*, is less than ideal because it is only a distant relative of *S. cerevisiae*, with a most recent common ancestor 0.3–1.3 billion years ago (Wood et al. 2002). We thus repeated the analysis with the much more closely related *S. paradoxus* as the outgroup. In addition to the identification of many more orthologs, use of *S. paradoxus* allows us to control for mutation rate biases among genes by using the ratio K_a/K_s instead of simply K_a. Pooled Data II (*cf.* Fraser et al. 2003) again yields significant correlations (r, −0.14; s, −0.24; p < 0.0001 for both) (Fig. 3A and Table 1) that are stronger than that using our Pooled Data I (r, −0.10; s, −0.11; p < 0.0001 for both; see also Table 1). The major difference between Pooled Data I and Pooled Data II is the inclusion of protein inter-

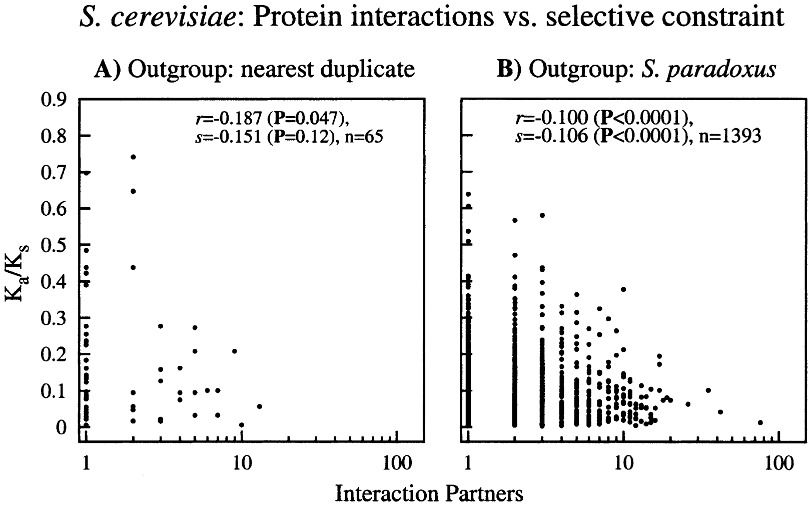

Fig. 2. Connectivity (D) in the yeast protein-interaction network (Pooled Data I) vs. selective constraint (K_a/K_s) of a gene and (**A**) its nearest paralog or (**B**) its ortholog in the *S. paradoxus* genome.

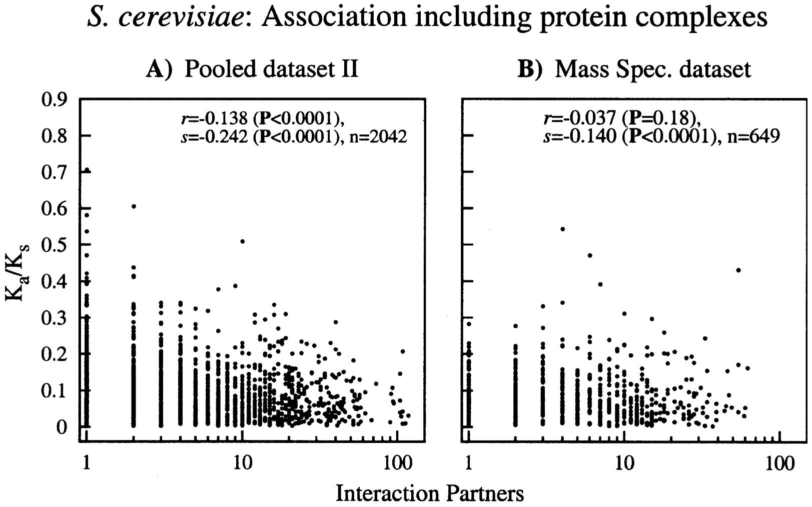

Fig. 3. Effects of dataset selection on the magnitude of statistical association between D and K_a/K_s (measured with *S. paradoxus* orthologs). **A** Protein interaction data taken from the medium- and high-confidence interactions of von Mering and coauthors (2002; Pooled Data II). **B** Protein interaction data taken from mass spectrometry experiments on purified protein complexes (Gavin et al. 2002; Ho et al. 2002) included in Pooled Data II.

actions from purified protein complexes in the latter. We therefore estimated the statistical association of K_a/K_s and connectivity, D, when D was computed strictly from the protein complex data of Gavin and coauthors (2002) and Ho and coauthors (2002) included in Pooled Data II. These data suggest a stronger (Spearman's) association than does our Pooled Data I (r, −0.04 [p = 0.18]; s, −0.14 [$p <$ 0.0001] (Table 1). Figure 3 illustrates the effect of using either Pooled Data II or just the protein complex data (compare to Fig. 2B). We conclude that the strong correlations seen by Fraser and coauthors (2002, 2003) and our own analyses of Pooled Data II data are due at least in part to the inclusion of data from purified protein complexes.

Mass spectrometry assays such as those of Gavin and coauthors (2002) and Ho and coauthors (2002) identify all members of a protein complex, rather than direct interactions between two proteins. Such data are less appropriate for our analysis because many member proteins of a large complex need not interact directly and because any one protein need not interact with a large number of other proteins to be part of a large complex. We also note that distributions of protein-interaction connectedness have significantly different statistical properties if data from purified complexes are included. We calculated maximum likelihood estimates (MLEs) for the power-law exponent b for both Pooled Data I and Pooled data II, as well as for data stemming only from purified complexes (see Materials and Methods and Eq. [1]). The purified complex data and the Pooled Data II have b = 1.64 and b = 1.71, respectively, values significantly different from the MLE of our Pooled Data I (b = 2.74, p < 0.0001) (Table 1). Qualitatively, this implies that there are many more genes with a large number of interactions in the purified complex data, exactly as one would expect from such data. We suspect that analyses using protein-complex data may therefore overstate the magnitude of the association between evolutionary constraint and network connectivity.

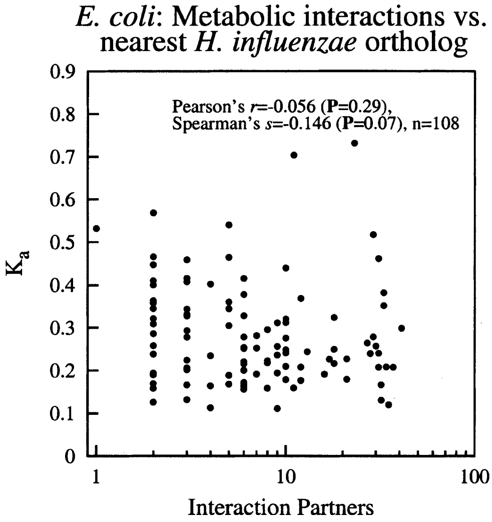

Fig. 4. Amino acid sequence divergence (K_a) from *H. influenzae* orthologs vs. connectivity (*D*) in the core intermediary metabolic network of *E. coli* (Wagner and Fell 2001).

The E. coli *Metabolic Network*

The *E. coli* core metabolic network encompasses the catabolic and biosynthetic metabolism central to the cell's function. Wagner and Fell (2001) constructed a graph representation of this network in which genes are connected by an edge if the chemical reactions their products catalyze share at least one substrate. We measured K_a by comparing genes in the network to their orthologs in the closely related *Haemophilus influenzae* genome (see Materials and Methods). Of the 133 genes in the *E. coli* metabolic network with related genes in *H. influenzae*, we used only gene pairs that have no paralogs in either species. Figure 4 shows the relationship between connectedness and K_a for the *E. coli* genes and their orthologs (Pearson's *r*, −0.056; Spearman's *s*, 0.146; $n = 108$; $1 \leq D \leq 41$; $0.111 \leq K_a \leq 0.731$). Permutation analyses indicate that neither correlation is significant (*r*, $p = 0.29$; *s*, $p = 0.067$).

This dataset includes many fewer genes than the protein-interaction comparisons and may, therefore, suffer from reduced statistical power. However, the lack of a correlation between selective constraint and degree of connectedness in the *E. coli* metabolic network is consistent with the results of Fraser and coauthors (2002, 2003) for the protein-interaction network. These researchers found that the weak negative correlation between evolutionary rate and protein connectivity was not due to the fact that highly connected genes are more important to the organism but, instead, due to the greater proportion of a highly connected protein's surface that is in contact with other proteins. In a metabolic network, where genes are connected not through physical contact but by shared metabolites, there should be no such effect. This is exactly what we observe.

How Strong Is the Effect of Protein Connectivity on Selective Constraint?

Many factors may contribute to differences in evolutionary rates among genes. Rates may differ simply because of differences in structure or function; other candidate factors determining rates of evolution include physical position in the genome (Williams and Hurst 2000), expression level (Akashi 2001), and "essentialness" (Hurst and Smith 1999), as well as the estimation errors introduced by gene length and GC-content (Kumar and Subramanian 2002). This makes it difficult to measure the effects of a single factor on selective constraint. In order to determine the influence of several factors other than connectivity, we carried out two further analyses. First, we used an analysis of variance (ANOVA) for our primary yeast dataset, Pooled Data I, and for the *E. coli* data. Each ANOVA included degree of connectedness (*D*), codon adaptation index (CAI; a correlate of expression level [Akashi 2001]), gene length, and GC-content as separate effects in the model (all analyses carried out in JMP, SAS Institute, Inc.). Second, we broke the yeast Pooled Data I protein-interaction dataset into five major functional categories to see whether there were function-specific correlations (see Materials and Methods for details).

As in the above analysis, only the yeast ortholog comparison showed a significant effect of protein connectedness on selective constraint in the ANOVA for Pooled Data I (*S. cerevisiae–S. paradoxus* [$F = 16.2, p < 0.0001$], *S. cerevisiae* paralogs [$F = 1.7, p = 0.19$], *E. coli–H. influenzae* orthologs [$F = 0.05, p = 0.82$]). CAI and GC-content also had significant effects in the Pooled Data I ortholog comparison (CAI: $F = 59.5$, $p < 0.0001$; GC-content: $F = 7.1$, $p = 0.008$), but only CAI was a significant effect in the other two comparisons (yeast paralogs: $F = 4.56, p = 0.037$; *E. coli–H. influenzae* orthologs: $F = 18.1, p < 0.001$). The fact that CAI is associated with evolutionary constraint is not surprising, since it is known from a number of organisms that expression level and evolutionary rate are negatively correlated (Akashi 2001). Degree of connectedness and CAI were significantly correlated with each other in the *E. coli* metabolic network ($r = 0.168, p = 0.08$; $s = 0.265, p = 0.006$) (highly connected genes were expressed at higher levels), but they were not significantly correlated in the yeast protein-interaction network. The full ANOVAs have controlled for many effects in attempting to detect the correlation between connectivity and selective constraint; nonetheless, they explain very little of the variation in K_a/K_s. The R^2

Table 1. Correlations between connectivity and selective constraint in yeast

Dataset	Source	Experimental technique	Number of genes	Spearmans's s^a	Power-law exponent (MLE)
Pooled Data I	Mewes et al. (1999), Ito et al. (2001), Uetz et al. (2000)	Two-hybrid and non-two-hybrid	1874	-0.11 ($p < 10^{-4}$)	2.74
MIPS	Mewes et al. (1999)	Various, no two-hybrid	674	-0.13 ($p = 0.002$)	2.60
Ito/Uetz	Ito et al. (2001), Uetz et al. (2000)	Two-hybrid	779/501	$-0.07/-0.01$ ($p = 0.04/0.45$)	3.68/2.37
Pooled Data II	von Mering et al. (2002)	Two-hybrid and protein complexes	2609	-0.24 ($p < 10^{-4}$)	1.71
Protein Complexes	Gavin et al. (2002), Ho et al. (2002)	Protein complexes identified by mass spectroscopy	820 (combined)	-0.14 ($p < 10^{-4}$)	1.64

[a] Spearman's correlation of D and K_a/K_s, using orthologs from *S. paradoxus*.

for the full Pooled Data I *S. cerevisiae—S. paradoxus* ANOVA is 0.061; using only connectivity as an effect in the model gives an $R^2 = 0.01$.

Using *S. cerevisiae–S. paradoxus* orthologs taken from Pooled Data I, we separated genes by function and calculated individual correlations for five major categories: cytoskeleton, metabolism, transport, cell cycle/DNA processing, and transcription. Interestingly, the relationship between network degree and evolutionary constraint is highly dependent on a gene's function. Genes involved in metabolism, transport, and the cytoskeleton show no significant relationship between D and K_a/K_s (p always > 0.05) (see Table 2). However, genes involved in the cell cycle and transcriptional processes show a significant, though weak, effect (s: -0.20 and -0.25) (Table 2). The overall correlation that we observe for these data, therefore, is an average of these two cases, as the cell cycle and transcription correlations are higher than the overall correlation (s: -0.11) (Table 1).

Is there any reason to expect certain functional categories to show stronger correlation than others? The connectivity distributions do not consistently differ between the categories (Table 2), so there do not appear to be systematic differences in the numbers of interactions the genes in each category take part in. One possibility is that the divergent roles proteins take on in the different categories are responsible for this disparity. For instance, the functions of metabolic enzymes are defined by their specific interactions with small molecules, whereas the function of transcription factors and cell-cycle regulators often involves mediation of protein interactions in larger protein complexes. In this context, it is surprising that cytoskeletal proteins do not show a strong effect, but we also note that these proteins constitute our smallest sample, rendering statistical inference more difficult.

Table 2. Function-specific correlations between connectivity and selective constraint in the yeast protein-interaction network

Functional classification	Number of genes	Spearman's s^a	Power-law exponent (MLE)
Cytoskeleton	49	-0.11 ($p = 0.23$)	2.76
Metabolism	231	0.04 ($p = 0.27$)	2.92
Transport	75	0.11 ($p = 0.17$)	2.08
Cell cycle/DNA processing	138	-0.20 ($p = 0.008$)	2.14
Transcription	134	-0.25 ($p = 0.002$)	2.42

[a] Spearman's correlation of D and K_a/K_s, using orthologs from *S. paradoxus*.

Are Highly Connected Proteins Especially Important to the Cell?

Biological networks are extremely robust to perturbation (Albert et al. 2000; Edwards and Palsson 2000; Hartwell et al. 1999; Jeong et al. 2001; Wagner 2000; Watts 1999). 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 (Ross-Macdonald et al. 1999; Smith et al. 1996; Winzeler et al. 1999). This robustness may be a function of a network's architecture: broad-tailed distributions (such as power laws) of the number of interactors observed in genetic networks have been hypothesized to confer robustness against mutations (Albert et al. 2000). However, these networks are vulnerable to loss of highly connected nodes. Here we have tested the robustness hypothesis through one of its key predictions: highly connected proteins should be subject to greater selective constraint. The *S. cerevisiae* protein-interaction network shows evidence of a significant association, but its magnitude is small. The *E. coli* metabolic network shows no significant association.

We cannot completely exclude the possibility that genes in these two networks have acquired their functions (and numbers of network interactors) very recently, after the divergence of *E. coli* from *H. influenzae* or *S. cerevisiae* from *S. paradoxus*. In that case, estimates of selective constraint might reflect past rather than present function. However, the *H. influenzae* core metabolic network is likely 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 (Edwards and Palsson 1999; Morowitz 1992; Tatusov et al. 1996). In addition, 56% (133) of the genes in the *E. coli* metabolic 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 (Blattner et al. 1997). *S. cerevisiae* is closely related *to S. paradoxus* and shares an even larger set of genes and interactions (Kellis et al. 2003).

We have used K_a and K_a/K_s to test the prediction that highly connected genes are more important to an organism. That is, we have used these measures of evolutionary constraint (the number of mutations tolerable in a coding region) as indicators of the importance of a gene. Of course, alternative indicators of importance are conceivable. For example, removing highly connected genes from a network may be more detrimental than removing less-connected genes, which could indicate the importance of highly connected genes. Evidence for this was obtained from gene knockout data and the protein-interaction network from yeast, where Jeong and coauthors (2001) 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 effects. 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} (Dykhuizen and Hartl 1983). Second, laboratory experiments cannot assess fitness differences over the entire environmental spectrum important for life in the wild. An apparently neutral knockout mutation in one environment may have severe fitness effects in environments not easily re-created in the laboratory. For these reasons, we think that our results are a more accurate reflection of the importance of genes in a genetic network. Using a protein's evolutionary rate takes advantage of an evolutionary record of mutations experienced over millions of years, mutations whose effects manifested themselves in the environments the organism experienced during its evolution.

Our results, as well those of others (Fraser et al. 2002, 2003), show that a gene's position in a network has only a very limited impact on its importance as defined by evolutionary constraints. There is some controversy as to the exact level of association, with Fraser and coauthors arguing for a stronger relationship in yeast than Jordan et al (2002a, b). We have shown here that the choice of protein-interaction datasets matters when looking for correlations with distant reference taxa but that the choice of datasets has less of an effect when using more closely related organisms. Our analyses of multiple types of protein-interaction data, using a more closely related reference taxon unavailable to earlier authors, come down somewhere in the middle of previous studies, with a very weak but statistically significant association (*r* or $s \approx -0.1$). Our results also show that the correlation between connectivity and evolutionary rate is function-specific: only certain functional classes of proteins show a significant correlation. And they fail to support the claim that power-law connectivity in cellular networks reflects selection for robust network diameters: analysis of a metabolic network, where genes do not physically interact, failed to show any correlation. What, then, can we learn from a purely qualitative, topological analysis of genetic networks? The work of Rausher and coauthors (1999) on anthocyanin biosynthesis genes 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. Future genetic network analyses that include details on enzyme kinetics and rates of flux across nodes may provide the information needed to better understand how the position, function, and interactions of proteins are likely to influence which genes are targeted by natural selection.

Acknowledgments. M.W.H. thanks M. Rausher, M. Rockman, M. Rutter, A. Sweigart, M. Uyenoyama, and R. Zufall for comments and suggestions; an NSF Doctoral Dissertation Improvement Grant provided support. G.C.C. is supported by the Department of Energy's Computational Sciences Graduate Fellowship program, administered by the Krell Institute. A.W. acknowledges financial support through NIH Grant GM63882 and the Santa Fe Institute.

References

Akashi H (2001) Gene expression and molecular evolution. Curr Opin Genet Dev 11:660–666

Albert R, Jeong H, Barabasi A-L (2000) Error and attack tolerance of complex networks. Nature 406:378–382

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

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

Bhalla US, lyengar R (1999) Emergent properties of networks of biological signaling pathways. Science 283:381–387

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462

Conant GC, Wagner A (2002) GenomeHistory: A software tool and its application to fully sequenced genomes. Nucleic Acids Res 30:3378–3386

Dykhuizen DE, Hartl DL (1983) Functional effects of PGI allozymes in *Escherichia coli*. Genetics 105:1–18

Edwards JS, Palsson BO (1999) Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. J Biol Chem 274:17410–17416

Edwards JS, Palsson BO (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. Proc Natl Acad Sci USA 97:5528–5533

Fleischmann RD, Adams MD, White O, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296:750–752

Fraser HB, Wall DP, Hirsh AE (2003) A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol 3:11

Gavin AC, Bosche M, Krause R, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147

The Gene Ontology Consortium (2000) Gene Ontology: Tool for the unification of biology. Nature Genet 25:25–29

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. Science 274:563–567

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402:C47–C52

Ho Y, Gruhler A, Heilbut A, Bader GD, et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415:180–183

Hurst LD, Smith NGC (1999) Do essential genes evolve slowly? Curr Biol 9:747–750

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98:4569–4574

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi A-L (2000) The large-scale organization of metabolic networks. Nature 407:651–654

Jeong H, Mason SP, Barabasi A-L, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411:41–42

Jordan IK, Wolf YI, Koonin EV (2003a) No simple dependence between protein evolution rate and the number of protein-protein interactions: Only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3:1

Jordan IK, Wolf YI, Koonin EV (2003b) Correction: No simple dependence between protein evolution rate and the number of protein-protein interactions: Only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3:5

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254

Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267:275–276

Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. Proc Natl Acad Sci USA 99:803–808

Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S, Frishman D (1999) MIPS: A database for genomes and protein sequences. Nucleic Acids Res 27:44–48

Morowitz HJ (1992) Beginnings of cellular life. Yale University Press, New Haven, CT

Rausher MD, Miller RE, Tiffin P (1999) Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol Biol Evol 16:266–274

Ross-Macdonald P, Coelho PSR, Roemer T, Agarwal S, Kumar A, Jansen R, Cheung KH, Sheehan A, Symoniatis D, Umansky L, Heldtman M, Nelson FK, Iwasaki H, Hager K, Gerstein M, Miller P, Roeder GS, Snyder M (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. Nature 402:413–418

Smith V, Chou KN, Lashkari D, Botstein D, Brown PO (1996) Functional analysis of the genes of yeast chromosome V by genetic footprinting. Science 274:2069–2074

Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. Curr Biol 6:279–291

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal-W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang MJ, Johnston M, Fields S, Rothberg JM (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403:623–627

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417:399–403

Wagner A (2000) Mutational robustness in genetic networks of yeast. Nature Genet 24:355–361

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

Wagner A (2002) Estimating coarse gene network structure from large-scale gene perturbation data. Genome Res 12:309–315

Wagner A, Fell D (2001) The small world inside large metabolic networks. Proc Roy Soc Lond Ser B 280:1803–1810

Watts DJ (1999) Small worlds. Princeton University Press, Princeton, NJ

Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393:440–442

Williams EJB, Hurst LD (2000) The proteins of linked genes evolve at similar rates. Nature 407:900–903

Winzeler EA, Shoemaker DD, Astromoff A, et al. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion ad parallel analysis. Science 285:901–906

Wood V, Gwilliam R, Rajandream MA, et al. (2002) The genome sequence of *Schizosaccharomyces pombe*. Nature 415:871–880