

Evolutionary Rate in the Protein Interaction Network

Hunter B. Fraser,^{1*}† Aaron E. Hirsh,^{2*} Lars M. Steinmetz,³
Curt Scharfe,³ Marcus W. Feldman²

High-throughput screens have begun to reveal the protein interaction network that underpins most cellular functions in the yeast *Saccharomyces cerevisiae*. How the organization of this network affects the evolution of the proteins that compose it is a fundamental question in molecular evolution. We show that the connectivity of well-conserved proteins in the network is negatively correlated with their rate of evolution. Proteins with more interactors evolve more slowly not because they are more important to the organism, but because a greater proportion of the protein is directly involved in its function. At sites important for interaction between proteins, evolutionary changes may occur largely by coevolution, in which substitutions in one protein result in selection pressure for reciprocal changes in interacting partners. We confirm one predicted outcome of this process—namely, that interacting proteins evolve at similar rates.

A protein's rate of evolution is thought to depend both on its dispensability to the organism and on the proportion of potential amino acid changes that are compatible with proper protein function (1). We recently analyzed functional genomic data (2) in conjunction with genomic comparisons (3) to confirm and further characterize the relation between protein dispensability and evolutionary rate (4). Here we apply a similar approach to investigate how protein function constrains evolution. Early studies of the structure and function of individual proteins suggested that, because molecular interactions require precisely specified structures, they impose constraints on sequence evolution (5, 6). Recent advances in the rapid detection of protein-protein interactions (7–9), as well as in the sequencing of complete genomes, allow us to expand the scale on which the evolutionary effects of molecular interactions are investigated and shift from a focus on individual proteins to a broad survey of the proteome and characterization of the general relation between protein interaction and evolution.

We compiled a list of 3541 interactions between 2445 different yeast proteins (10). To estimate the evolutionary rates of these proteins, we compared putatively orthologous sequences between *Saccharomyces cerevisiae* and the nematode *Caenorhabditis elegans* (11). A subclass of putative orthologs, which we called “well-conserved orthologs,”

exhibited >50% amino acid identity over aligned regions; 1531 sequence pairs met our criteria for putative orthologs, and 309 of these were in the well-conserved class. For each pair of orthologs, we estimated the evolutionary distance (K) that separates the two sequences, where K is defined as the number of substitutions per amino acid site that have taken place since the fungi-animal split (12). There were 164 yeast proteins for which we had both an estimate of the number of interactors and a well-conserved ortholog in the nematode. Among these proteins, there is a significant negative correlation between each protein's number of interactors I and protein evolutionary rate, as estimated by distance K [Fig. 1; linear regression: $K = -0.0175I + 0.8995$, Pearson's $r_{IK} = -0.24$, $P = 0.002$; Spearman's rank correlation $r_{IK} = -0.21$, $P = 0.007$ (13)]. We have corroborated this relation between protein interaction and rate of evolution with data from two recent studies (14, 15) that were not considered in our initial compilation of protein interactions [supplemental fig. 1 (16)].

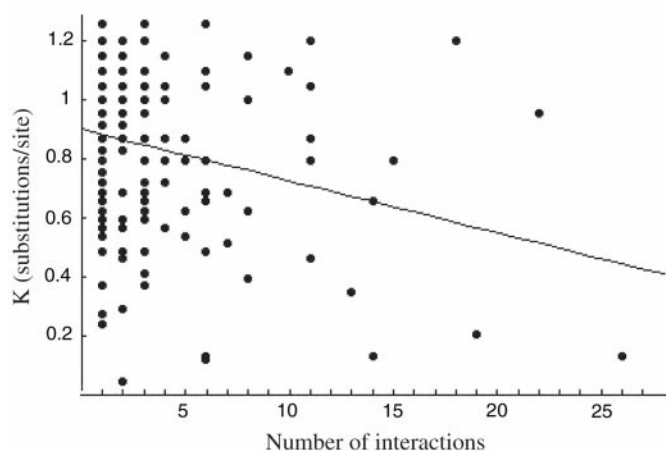


Fig. 1. The relation between the number of protein-protein interactions (I) in which a yeast protein participates and that protein's evolutionary rate, as estimated by the evolutionary distance (K) to the protein's well-conserved ortholog in the nematode *C. elegans*.

¹Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. ²Center for Computational Genetics and Biological Modeling, Department of Biological Sciences, Stanford University, Stanford, CA 94305, USA. ³Stanford Genome Technology Center, Palo Alto, CA 94306, USA.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: hunter@ocf.berkeley.edu

Interactions could reduce evolutionary rate in two distinct ways (Fig. 2). First, if different interactions depend on different sites, proteins with more interactors could evolve more slowly because a greater proportion of the protein is involved in protein functions (Fig. 2, arrow a). Alternatively, if proteins with many interactors have a greater effect on organism fitness, they could evolve more slowly, not because a greater proportion of the sequence is required for proper function, but because the entire sequence is subject to stronger selection against slightly deleterious mutations (4). Under this hypothesis, the correlation shown in Fig. 1 emerges because a protein's number of interactors is correlated with its effect on organism fitness, which in turn affects rate of evolution (Fig. 2, arrows b and c). To determine which of these two hypotheses provides a more likely explanation for the correlation between number of interactors and evolutionary rate, we analyzed our data on interactions and evolutionary rate in conjunction with results from genetic footprinting (17) and parallel analysis (2), high-throughput methods for estimating the growth rates of yeast strains in which a single gene has been disrupted or deleted. As expected in view of the recent demonstration that highly interactive proteins are more likely to be required for viability (18), we found that a protein's fitness effect F , estimated as the reduction in relative growth rate due to deleting or disrupting the gene that encodes the protein, is positively correlated with that protein's number of interactors I ; with fitness effects measured by parallel analysis for 2235 proteins for which interaction data were available, $r_{IF} = 0.15$, $P = 3.4 \times 10^{-13}$. In addition, among all putative orthologs, evolutionary rate is negatively correlated with fitness effect (4); with parallel analysis data for 1484 yeast proteins with putative orthologs, $r_{FK} = -0.13$, $P = 4.3 \times 10^{-7}$ (19). Thus, among all putative orthologs, both correlations required by our second hypothesis are present: Number of interactors is corre-

lated with fitness effect (Fig. 2, arrow b), which is correlated with evolutionary rate (Fig. 2, arrow c).

However, when we consider only well-conserved orthologs, for which the correlation between protein interaction and evolutionary rate is strongest (Fig. 1), no relation between fitness effect and evolutionary rate (Fig. 2, arrow c) is detected. Therefore, protein fitness effect is very unlikely to mediate the correlation between protein interaction and evolutionary rate. We can confirm this conclusion statistically by using parametric (20) and nonparametric (21) partial correlation to estimate the correlation between number of interactors and evolutionary rate while fitness effect is held constant. The parametric path coefficient ($p_{IK} = -0.25, P = 0.001$) and nonparametric partial measure of association (Kendall's partial $\tau_{IK} = -0.15, P = 0.002$) indicate a significant correlation between number of interactions and evolutionary rate that does not depend on overall protein fitness effect.

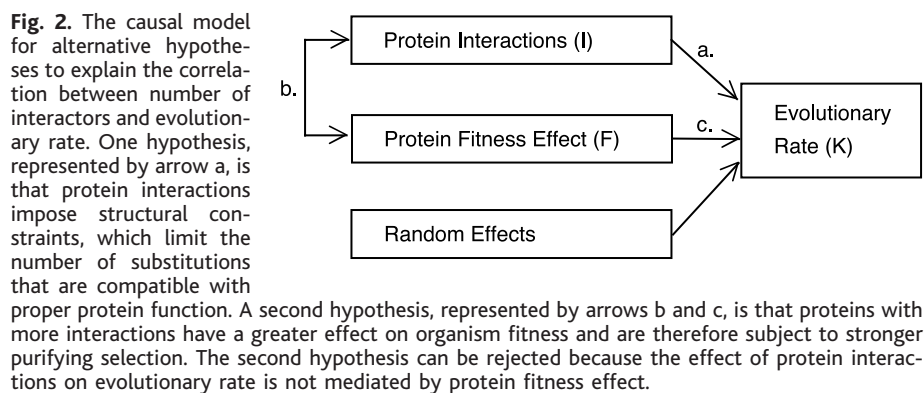


Fig. 3. Interacting proteins have similar fitness effects, but this cannot explain the similarity in their rates of evolution. (A) The distribution of mean difference in evolutionary rate ($\Delta\bar{K}$) between yeast proteins randomly chosen from the list of all 411 interacting protein pairs in which both members had an ortholog in *C. elegans*. The mean difference in evolutionary rate between proteins that interact ($\Delta\bar{K}^* = 1.3$ substitutions per site) is indicated by an arrow. (B) The distribution of mean difference in fitness effect ($\Delta\bar{F}$) between yeast proteins randomly chosen from the list of all 2821 interactions in which the effect on growth rate of deleting each protein was estimated by parallel analysis (2). The mean difference in fitness effect between proteins that interact ($\Delta\bar{F}^* = 0.41$) is indicated by an arrow. (C) The causal model for path analysis to determine whether similarity in fitness effects between interacting proteins explains the similarity in their evolutionary rates. The correlation between evolutionary rates of interacting proteins that is expected to result from observed correlations between fitness effects (r_{F_1, F_2}), and between fitness effect and evolutionary rate (p_{FK}), can be estimated as $\hat{r}_{K_1, K_2} \sim (p_{FK})^2 r_{F_1, F_2}$. The observed correlation between evolutionary rates is much larger than that expected to result from fitness effects ($r_{K_1, K_2} \gg \hat{r}_{K_1, K_2}$), indicating that one or more additional factors must contribute to the similarity of evolutionary rates of interacting proteins. (Observed correlation coefficients, including essential proteins: $r_{F_1, F_2} = 0.16, P < 10^{-15}; p_{FK} = -0.06, P = 0.02; r_{K_1, K_2} = 0.11, P = 0.03$. Excluding essential proteins: $r_{F_1, F_2} = 0.07, P = 0.01; p_{FK} = -0.14, P = 2 \times 10^{-5}$.)

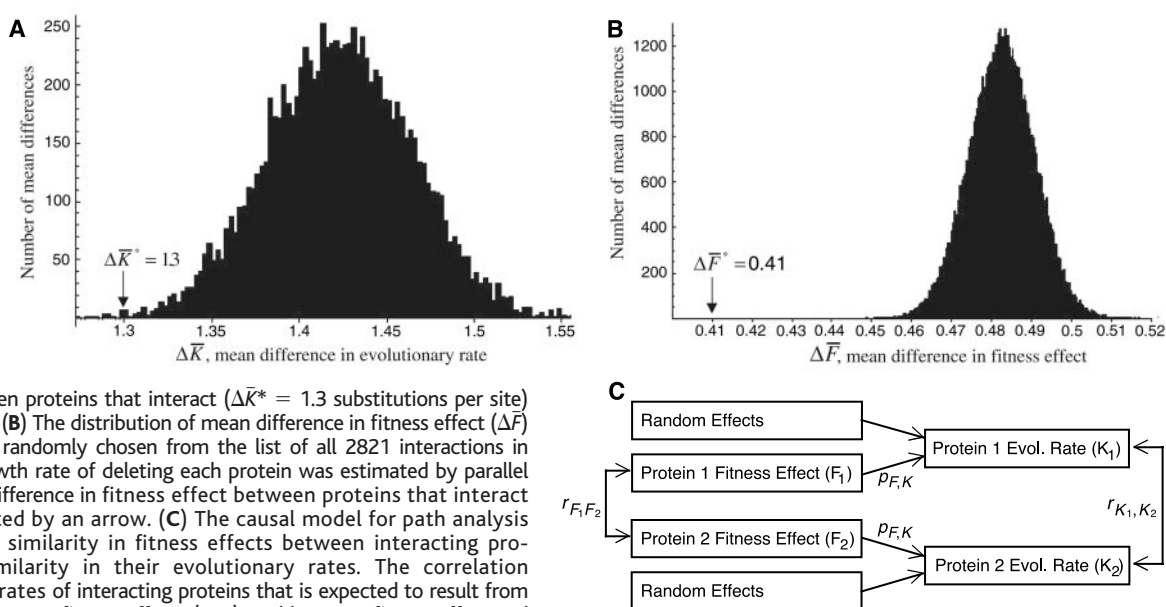
Protein sites may be involved in interactions directly, through participation in intermolecular contacts; or indirectly, through effects on overall protein conformation. In either category of sites, substitutions would be likely to perturb proper interaction and would often be removed by selection. However, removal might not occur if a substitution in one protein were followed by a complementary change in its interacting partner. In this case, the pair of substitutions might be fixed by drift or positive selection (22). If such coevolution is indeed an important mode of change in proteins constrained by interactions, then interacting proteins should evolve at similar rates. We tested this prediction by examining all 411 protein interactions in which each protein had a putative ortholog in *C. elegans* and showed no significant sequence similarity with its interacting partner. For each interaction, we calculated $\Delta\bar{K}$, the difference between the evolutionary distances separating the yeast proteins from their respective orthologs in the nematode. We

then averaged these differences across all 411 interactions to find the mean difference in evolutionary rate between interacting proteins, $\Delta\bar{K}^* = 1.3$ substitutions per site. To assess the significance of this difference, we repeatedly permuted our list of 411 interactions to random protein pairs and calculated the mean difference in evolutionary rate between arbitrarily paired proteins: 10,000 permutations yielded the distribution of $\Delta\bar{K}$ values shown in Fig. 3A. In all but 44 of the 10,000 permutations, our observed $\Delta\bar{K}^* < \Delta\bar{K}$, indicating that interacting proteins evolve at rates significantly closer than is expected to occur by chance ($P = 0.0044$).

Although coevolution provides an appealing explanation for the similarity in the evolutionary rates of interacting proteins, alternative hypotheses must be considered. The proteins in an interacting pair presumably act in the same functional pathway and therefore are likely to have similar effects on organism fitness. Because the dispensability of a protein influences its rate of evolution (4), the similarity in the evolutionary rates of interacting proteins could be a consequence of similarity in their fitness effects. Our test of this hypothesis involved two steps.

First, we tested whether proteins that interact do indeed have similar effects on organism fitness. A randomization test showed that the mean difference in fitness effects between interacting proteins, $\Delta\bar{F}^* = 0.41$, was significantly smaller than the mean difference between arbitrarily paired proteins $\Delta\bar{F}$ ($P < 10^{-5}$) (Fig. 3B). Thus, interacting proteins do have similar effects on organism fitness.

Second, we determined whether the ob-



served similarity in fitness effects of interacting proteins was sufficient to explain the similarity in their rates of evolution. Path analysis based on the causal model shown in Fig. 3C indicated that the correlation between the fitness effects of interacting proteins contributes only slightly to the correlation between their evolutionary rates. Thus, similarity in fitness effects is not sufficient to explain the observed similarity in the evolutionary rates of interacting proteins.

We also considered two other alternatives to the coevolutionary hypothesis. First, interacting proteins might evolve at similar rates simply because they have similar numbers of interactors, and, as shown in Fig. 1, the number of interactors influences the rate of evolution. However, we found that proteins that interact do not have similar numbers of interactors ($r_{I_1, I_2} = 0.02$, $P = 0.26$). A second possibility is that interacting proteins evolve at similar rates because they exhibit structural homology and therefore have similar distributions of constrained sites. The most likely origin of structural homology between interacting proteins is duplication of the gene that encodes a homodimeric protein, followed by evolution of one copy of the gene. This process would result in homology not only between the structures, but also between the sequences, of interacting proteins. Hence, we have ensured that none of the interactions in our data set occur between proteins that exhibit detectable sequence similarity. Thus, to account for the similarity in evolutionary rates that we observe, structural similarity would have to be independent of sequence, which would be difficult to explain evolutionarily. In sum, having considered a number of alternative hypotheses, we conclude that the coevolution of interacting proteins may be largely responsible for the observed similarity in their rates of evolution.

Beyond describing the relation between a protein's interactions and its rate of evolution, the correlations presented here could find application in the rapid assessment of functional genomic data. Much as gene expression levels have recently been used to assess protein-protein interaction data sets (23), the correlation between protein interaction and evolutionary rate may allow one to use simple genomic sequence comparisons to statistically assess the quality of large interaction data sets. More generally, correlations between protein interaction, fitness effect, and evolutionary rate may provide a means by which multiple bioinformatic data sets can be quickly cross-referenced to assess the reliability of any single method or data set.

References and Notes

1. A. C. Wilson, S. S. Carlson, T. J. White, *Annu. Rev. Biochem.* **46**, 573 (1977).
 2. E. A. Winzeler et al., *Science* **285**, 901 (1999).
 3. S. A. Chervitz et al., *Science* **282**, 2022 (1998).
 4. A. E. Hirsh, H. B. Fraser, *Nature* **411**, 1046 (2001).

5. E. Zuckerkandl, *J. Mol. Evol.* **7**, 167 (1976).
 6. R. E. Dickerson, *J. Mol. Evol.* **1**, 26 (1971).
 7. B. Schwikowski, P. Uetz, S. Fields, *Nature Biotechnol.* **18**, 1257 (2000).
 8. P. Uetz et al., *Nature* **403**, 623 (2000).
 9. T. Ito et al., *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4569 (2001).
 10. Our data set of yeast protein-protein interactions was compiled from (7), the "core data" of (9), and 625 interactions identified by the two-hybrid array method [described in (8)] listed at <http://depts.washington.edu/sfields/yplm/data/new2h.html>.
 11. Putative orthologs were defined as reciprocal best hits (4, 24) with $P < 10^{-10}$ and >80% protein sequence alignment in reciprocal searches using Washington University Basic Local Alignment Search Tool for Proteins (WU-BLASTP) (3). Thus, if a query of all 19,099 nematode open reading frames (ORFs) with yeast ORF *i* yielded the set of hits {*W*} with $P < 10^{-10}$ and >80% alignment, and a query of all 6217 yeast ORFs with nematode ORF *j* yielded the set of hits {*Y*} with $P < 10^{-10}$ and >80% alignment, then the pair (ORF *i*, ORF *j*) was considered a putative ortholog only if *i* was the member of {*Y*} with the lowest *P* value and *j* was the member of {*W*} with the lowest *P* value.
 12. To estimate the number of substitutions per amino acid site, we numerically solved the equation $q = [\ln(1 + 2K)]/2K$, where *q* is the proportion of identical sites between aligned sequences. This method is described in (25) and tested further in (26).
 13. Unless otherwise specified, the correlation coefficient $r_{X,Y}$ represents Spearman's rank correlation between variables *X* and *Y*.
 14. Y. Ho et al., *Nature* **415**, 180 (2002).
 15. A. C. Gavin et al., *Nature* **415**, 141 (2002).
 16. Supplementary fig. 1 is available on Science Online at www.sciencemag.org/cgi/content/full/296/5568/750/DC1.
 17. V. Smith, K. N. Chou, D. Lashkari, D. Botstein, P. O. Brown, *Science* **274**, 2069 (1996).
 18. H. Jeong, S. P. Mason, A. L. Barabasi, Z. N. Oltvai, *Nature* **411**, 41 (2001).
 19. The r_{IF} and r_{FK} correlations were also significant when genetic footprint data (27) were used instead of parallel analysis data. When essential proteins are excluded from the analysis, which is expected to strengthen the correlation between fitness effect and evolutionary rate (4), $r_{FK} = -0.18$, $P = 1.3 \times 10^{-8}$, for 944 nonessential proteins in parallel analysis.
 20. D. Kaplan, in *Advanced Quantitative Techniques in the Social Sciences*, J. de Leeuw, R. Berk, Eds. (Sage Publications, Thousand Oaks, CA, 2000), pp. 13–39.
 21. J. D. Gibbons, in *Sage University Papers*, M. S. Lewis-Beck, Ed. (Sage Publications, Newbury Park, CA, 1993), pp. 3–29.
 22. Examples of interprotein coevolution are described in (28, 29).
 23. A. Grigoriev, *Nucleic Acids Res.* **29**, 3513 (2001).
 24. M. C. Rivera, R. Jain, J. E. Moore, J. A. Lake, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6239 (1998).
 25. N. V. Grishin, *J. Mol. Evol.* **41**, 675 (1995).
 26. D. Feng, R. Doolittle, *J. Mol. Evol.* **44**, 361 (1997).
 27. V. Smith, D. Botstein, P. O. Brown, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 6479 (1995).
 28. P. D. Rawson, D. A. Brazeau, R. S. Burton, *Gene* **248**, 15 (2000).
 29. K. K. Koretke, A. N. Lupas, P. V. Warren, M. Rosenberg, J. R. Brown, *Mol. Biol. Evol.* **17**, 1956 (2000).
 30. B. Dunn provided genetic footprinting data. P. Uetz assisted with interaction data. J. Davis, M. Eisen, B. Kerr, D. Petrov, and C. Winter provided helpful discussion and comments on the manuscript.

5 December 2001; accepted 14 March 2002

Genetic Dissection of Transcriptional Regulation in Budding Yeast

Rachel B. Brem,* Gaël Yvert,* Rebecca Clinton, Leonid Kruglyak†

To begin to understand the genetic architecture of natural variation in gene expression, we carried out genetic linkage analysis of genomewide expression patterns in a cross between a laboratory strain and a wild strain of *Saccharomyces cerevisiae*. Over 1500 genes were differentially expressed between the parent strains. Expression levels of 570 genes were linked to one or more different loci, with most expression levels showing complex inheritance patterns. The loci detected by linkage fell largely into two categories: cis-acting modulators of single genes and trans-acting modulators of many genes. We found eight such trans-acting loci, each affecting the expression of a group of 7 to 94 genes of related function.

Genetic linkage analysis has traditionally focused on mapping loci that affect one or a small number of organism-level phenotypes. DNA microarray technology (1, 2) makes it possible to apply such analysis to global patterns of gene expression, with the transcript abundance of each of thousands of genes

treated as a quantitative phenotype (3). Although it has recently become clear that genetic variation has a strong effect on gene expression (4–7), little is known about the genetic basis of natural variation in expression levels (the number and type of loci involved, the effect of each locus, and the interaction between loci).

We carried out linkage analysis of global expression levels in a cross between two strains of the budding yeast *Saccharomyces cerevisiae*. The parents were haploid derivatives of a standard laboratory strain (BY) and a wild isolate from a California vineyard

Fred Hutchinson Cancer Research Center (FHRC), 1100 Fairview Avenue North, D4-100, Seattle, WA 98109, USA and Howard Hughes Medical Institute.

*These authors contributed equally to this work.
 †To whom correspondence should be addressed. E-mail: leonid@fhrc.org