BMC Evolutionary Biology

BioMed Central

Research article

A simple dependence between protein evolution rate and the number of protein-protein interactions

Hunter B Fraser*[1], Dennis P Wall[2] and Aaron E Hirsh[2]

Address: [1]Department of Molecular and Cell Biology, University of California, Berkeley, CA, 94720, USA and [2]Center for Computational Genetics and Biological Modeling, Department of Biological Sciences, Stanford University, Stanford, CA, 94305, USA

Email: Hunter B Fraser* - hunter@ocf.berkeley.edu; Dennis P Wall - dpwall@stanford.edu; Aaron E Hirsh - aehirsh@stanford.edu

* Corresponding author

Published: 23 May 2003

BMC Evolutionary Biology 2003, **3**:11

This article is available from: http://www.biomedcentral.com/1471-2148/3/11

Received: 27 March 2003
Accepted: 23 May 2003

Abstract

Background: It has been shown for an evolutionarily distant genomic comparison that the number of protein-protein interactions a protein has correlates negatively with their rates of evolution. However, the generality of this observation has recently been challenged. Here we examine the problem using protein-protein interaction data from the yeast *Saccharomyces cerevisiae* and genome sequences from two other yeast species.

Results: In contrast to a previous study that used an incomplete set of protein-protein interactions, we observed a highly significant correlation between number of interactions and evolutionary distance to either *Candida albicans* or *Schizosaccharomyces pombe*. This study differs from the previous one in that it includes all known protein interactions from *S. cerevisiae*, and a larger set of protein evolutionary rates. In both evolutionary comparisons, a simple monotonic relationship was found across the entire range of the number of protein-protein interactions. In agreement with our earlier findings, this relationship cannot be explained by the fact that proteins with many interactions tend to be important to yeast. The generality of these correlations in other kingdoms of life unfortunately cannot be addressed at this time, due to the incompleteness of protein-protein interaction data from organisms other than *S. cerevisiae*.

Conclusions: Protein-protein interactions tend to slow the rate at which proteins evolve. This may be due to structural constraints that must be met to maintain interactions, but more work is needed to definitively establish the mechanism(s) behind the correlations we have observed.

Introduction

What factors determine the rates at which different proteins evolve is a fundamental question in molecular evolution. With the advent of functional genomics, this question can now be addressed on a genome-wide scale. Different determinants of evolutionary rate revealed by analysis of functional genomic data include protein dispensability [1], transcript level [2], and number of protein-protein interactors [3].

Recently, Jordan *et al.* [4] suggested that the correlation between a protein's evolutionary rate and its number of protein interactions arises only because a few, highly interactive proteins evolve more slowly than all other proteins. In our original analysis, a distant genomic comparison of *S. cerevisiae* with *C. elegans* was used to find approximate evolutionary rates of putatively orthologous genes shared by these two species. One would expect that comparisons of more closely related species would increase the strength of the relationship, since more orthologs can be found and evolutionary rates can be estimated with greater precision. Surprisingly, when Jordan *et al.* compared orthologs between *S. cerevisiae* and another yeast, *S. pombe*, they found only an extremely weak relationship between number of protein interactions and evolutionary rate. Furthermore, they found that when proteins were binned by their number of interactions, only the bin containing the most highly interactive proteins showed any reduction in evolutionary rate. They concluded from this that there is no general correlation between number of protein interactions and evolutionary rate, and that the reduction of evolutionary rate observed in the most highly connected proteins may be an indirect effect of the relationship between protein dispensability and rate of evolution [4].

Here we show that the absence of a general correlation between protein interactions and evolutionary rate in the analysis of Jordan *et al.* can be attributed to an incomplete dataset. Our analysis differs from that of Jordan *et al.* in two basic ways. First, Jordan et al. used only protein-protein interactions from the MIPS database [5], which consists of individually reported interactions combined with data from the high-throughput screen of Uetz *et al.* [6]. While the MIPS database contains many high-confidence interactions, it is very small when compared to the total number of interactions known from all high-throughput screens. Second, Jordan *et al.* identified orthologs by taking reciprocal best BLAST hits, a method that leads to an incomplete list, because the top BLAST hit is often not the most closely related protein [7]. We used a method based on maximum likelihood estimation of evolutionary distances that results in a more complete list [8].

Using our more complete lists of both protein-protein interactions and orthologs, we show here that the correlation we originally reported in the *C. elegans* – *S. cerevisiae* comparison is indeed even stronger when more closely related genome sequences are compared. We use orthologs between *S. cerevisiae* and *S. pombe,* as well as the more closely related yeast, *C. albicans*, to probe this relationship in greater detail than we did in our previous study. We find a simple monotonic relationship between number of protein interactions and evolutionary rate, and we find that this relationship applies to proteins with few interactions, as well as to those with many.

Results and Discussion

Protein-protein interactions and evolutionary rates in yeast

We compiled a list of *S. cerevisiae* protein-protein interactions from every major high-throughput study published to date [6,9–11], as well as individually reported interactions from the MIPS database [5]. The final non-redundant set consists of all interactions used in our previous study [3], and contains 13,925 interactions involving 3575 proteins. This is a more comprehensive data set than that analyzed by Jordan *et al.* [4], which contained fewer than 2500 interactions once duplicate interactions were removed [I.K. Jordan, pers comm].

Using the genome sequences of *C. albicans* and *S. pombe* for comparison with *S. cerevisiae*, we identified putative orthologs using a maximum likelihood-based approach [8], which identified 3727 orthologs between *S. cerevisiae* and *C. albicans*, and 2988 orthologs between *S. cerevisiae* and *S. pombe*. All data will be made available upon request.

Taking the intersections of our interaction and ortholog data sets, we plotted the number of protein-protein interactions vs. evolutionary rate for all genes for which we had both types of data. For all genes in the *S. pombe*–*S. cerevisiae* comparison, we found a highly significant relationship (Figure 1a; n = 2119, Spearman Rank r = -0.24, P = 5.8×10^{-30}). This correlation is stronger than the rank correlation that we reported in our original study [3], and is over 27 orders of magnitude more statistically significant, due to both the increased strength and the far greater number of genes involved. Thus our expectation of a more significant correlation from a closer genomic comparison is borne out by the data.

A conclusion of Jordan *et al.* [4] was that only the proteins with the most interactors showed any reduction in evolutionary rate-i.e., the relationship between interactions and evolutionary rate was confined to those proteins with the most interactors. As shown in Figure 1b, when a more complete set of interactions and orthologs is used, the relationship can be seen to extend over the entire range of number of interactions. It takes the form of a simple monotonic relationship. This supports the idea that regardless of how many protein-protein interactions a protein participates in, each interaction affects the protein's rate of evolution.

This same analysis can be repeated using an *S. cerevisiae*–*C. albicans* genomic comparison, and the same set of *S. cerevisiae* protein-protein interactions. When we perform this analysis, the results are even stronger than for the *S. pombe* comparison. As shown in Figure 2a, a significant correlation is found (n = 2496, Spearman Rank r = -0.25, P = 5.2×10^{-38}). Separating the data into bins by their number of interactors also shows the same relationship as for the *S. pombe* comparison, with a clearly monotonic relationship observable over the entire range of protein interactions per protein (Figure 2b).

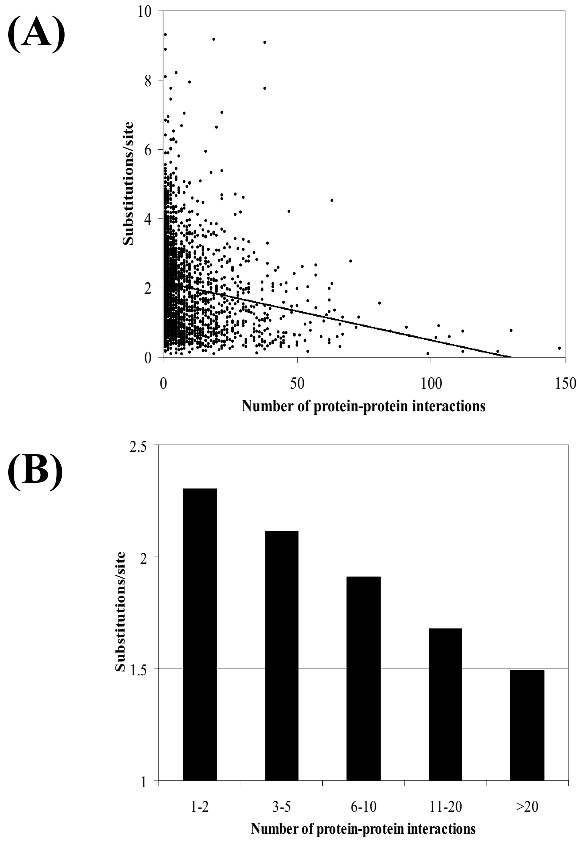

Figure 1
The relationship between number of protein-protein interactions and evolutionary rate between *S. cerevisiae* and *S. pombe*. (a) The relationship between number of protein-protein interactions and evolutionary rate for all 2119 orthologs with protein interaction data. Several outliers are not shown but were included in the analysis. (b) Average evolutionary rates of genes binned by their number of protein-protein interactions.

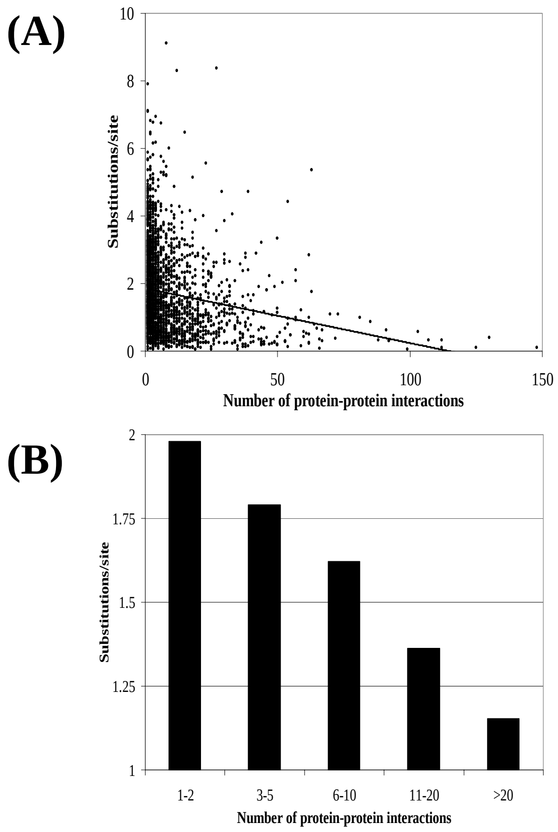

Figure 2
The relationship between number of protein-protein interactions and evolutionary rates between *S. cerevisiae* and *C. albicans*. (a) The relationship between number of protein-protein interactions and evolutionary rate for all 2496 orthologs with protein interaction data. Several outliers are not shown but were included in the analysis. (b) Average evolutionary rates of genes binned by their number of protein-protein interactions.

What is the source of the difference between the two studies?

Our finding of a strong correlation where Jordan *et al.* [4] did not find one raises the question of what causes the difference. There are two possibilities: our lists of protein-protein interactions, or our lists of orthologs and the associated evolutionary rates. To answer this question, we first tested the correlation between our list of protein interactions and Jordan *et al.*'s list of orthologs and evolutionary rates. We observed a significant correlation between the two (Spearman Rank r = -0.22, P = 8.5 × 10^{-24} Figure 3a), only slightly weaker than our correlation in Figure 1a. Next we plotted Jordan *et al.*'s protein interaction data against our list of evolutionary rates. We found no significant correlation between the two data sets (Spearman Rank r = -0.01, P = 0.79; Figure 3b). This demonstrates that the difference in our findings was due to the difference in our protein interaction lists, and not in our list of orthologs or evolutionary rates, and it underscores the importance of using datasets that are as complete as possible in this type of analysis.

Is it an indirect correlation?

Jordan *et al.* speculate that the reduction in evolutionary rate of the most highly connected proteins could be due to their greater likelihood of being essential for viability of the cell [1,12]. However in our original analysis we

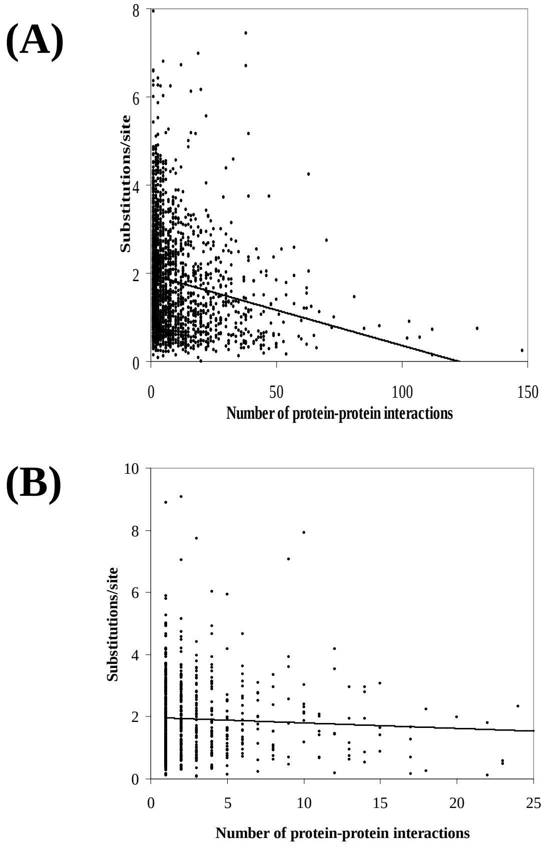

Figure 3
Testing the different lists of protein-protein interactions and evolutionary rates from the two studies. (a) A significant correlation is found when using evolutionary rates of orthologs from Jordan *et al*. [4] with our list of protein-protein interactions. Several outliers are not shown but were included in the analysis. (b) No correlation is seen when using our evolutionary rates of orthologs with Jordan *et al*.'s list of protein-protein interactions.

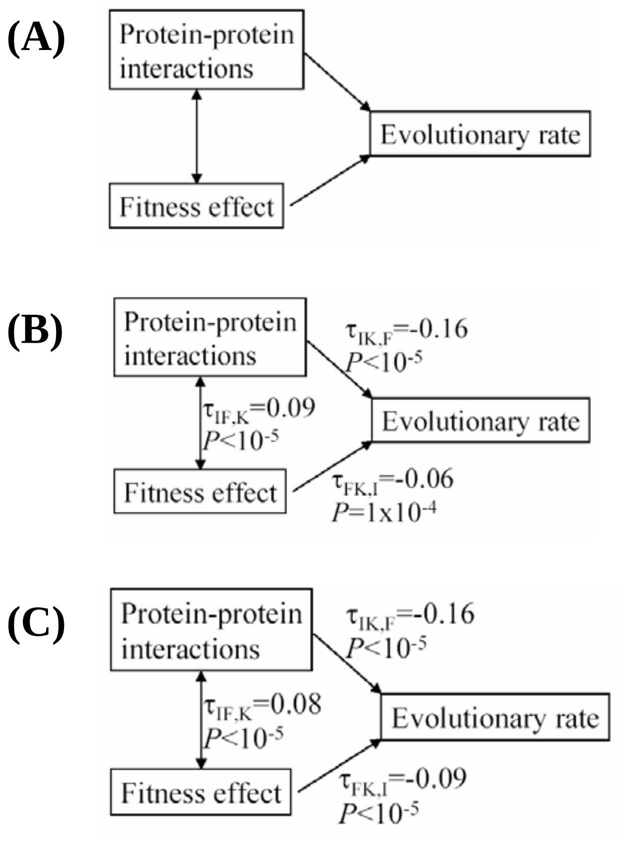

Figure 4
Diagram of correlations between number of protein-protein interactions, evolutionary rates, and fitness effects (a) Each arrow represents the correlation between the two variables it connects. Whether or not the correlation is statistically significant by Kendall's Partial Tau is shown by the *P*-values next to each arrow in (b) and (c). (b) The significance of each correlation for the *S. cerevisiae-S. pombe* comparison. Note that the arrow connecting number of protein-protein interactions and evolutionary rates is highly significant, with none of the 10^5 randomizations of the data having a stronger correlation. (c) The significance of each correlation for the *S. cerevisiae–C. albicans* comparison. Note that the arrow connecting number of protein-protein interactions and evolutionary rates is highly significant, with none of the 10^5 randomizations of the data having a stronger correlation.

showed that the effect on cell fitness when a gene is deleted cannot explain the correlation between number of protein interactions and evolutionary rate [3]. In order to investigate this question for these less distant genomic comparisons, we repeated the analysis from our original study. We used Kendall's Partial Tau [13], a metric of partial correlation that allows one to quantify the magnitude of a correlation between two variables when a third, potentially related variable is statistically held constant. For example, in Figure 4a, a diagram is shown in which the arrows connecting the three variables represent the relationships among them. We used Kendall's Partial Tau to assign a *P*-value (by 10^5 randomization tests of the data) to each arrow, representing the probability that the arrow represents a correlation that is significantly different from zero when the third variable is statistically controlled. When we use this method to analyze the number of protein-protein interactions, evolutionary rate, and fitness effect of each gene, we find that fitness effect cannot explain the relationship between number of protein-protein interactions and evolutionary rate for either of our genomic comparisons (Figure 4b,4c), consistent with our original study [3].

Protein-protein interactions and evolutionary rates in bacteria

Jordan *et al.* [4] also note that they cannot detect a correlation between number of protein-protein interactions and evolutionary rate in *Helicobacter pylori*. Based on this observation, they conclude that the relationship between interactions and evolutionary rate does not apply to bacteria. However, substantial caution must be exercised in interpreting results that are based on a single protein interaction study [14]. Indeed, when using either one of the first two published high-throughput yeast protein interaction data sets [6,9] alone, it is not possible to find a significant correlation between the number of interactors and evolutionary rate; it is only through a compilation of several data sets that a significant relationship emerges for yeast. Until this is possible for *H. pylori*, we should be reluctant to conclude whether or not such a relationship exists in this organism.

Conclusions

We have shown that the previously reported relationship between protein-protein interactions and evolutionary rates of proteins is even stronger when comparing different yeast species than it is when comparing yeast with *C. elegans*. The fact that the relationship can be detected at all with a genomic comparison of species separated by approximately 1 billion years of evolution (*S. cerevisiae* and *C. elegans*), as well as with the comparisons of the more closely related species presented here, underscores the robustness of the relationship. That the correlation cannot be detected when using a smaller set of protein-protein interactions, as in the study by Jordan *et al.* [4], demonstrates the importance of using data that are as complete as possible when correlating diverse genomic data. Since no such complete data set is available for any organism other than *S. cerevisiae*, it is not yet possible to judge whether the relationship applies to prokaryotes as well as eukaryotes.

It was correctly noted by Jordan *et al.* that the correlation we previously observed explains only about 6% of the variance in evolutionary rates [3]; the correlations presented here are only slightly stronger. However, when one considers the various and unavoidable sources of noise in the analysis (e.g., identifying orthologs, aligning orthologs, estimating evolutionary distances, and perhaps most importantly, false positives and negatives in the protein-protein interaction data), as well as confounding biological factors (e.g., the fact that protein-protein interactions will not be invariable between the species whose genomes are compared, so interactions recently evolved in the *S. cerevisiae* lineage will not show a significant effect on evolutionary rate), it seems surprising that the correlations are as strong as they are. In view of the sources of noise presently unavoidable in evolutionary analysis of functional genomic data, the fraction of variance in evolutionary rate that is explained by any one functional paramater – such as protein interactions, dispensability, or expression – cannot yet be taken as an accurate estimate of the relative importance of that factor's role in determining the rate of evolution. It will be interesting to see how much the strength of the correlations examined here increases, and whether the relationships take informative functional forms, as more high-quality protein-protein interaction data sets and genome sequences are published.

Authors' Constributions

HBF performed the correlational and statistical analyses, and wrote the manuscript. DPW found putative orthologs and evolutionary rates, and edited the manscript. AEH edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

HBF is supported by an NSF predoctoral fellowship. DPW is supported by an NSF postdoctoral fellowship in bioinformatics. AEH and DPW are supported by NIH grants GM 28016 and 28424 to Marcus W. Feldman.

References

1. Hirsh AE and Fraser HB **Protein dispensability and rate of evolution** *Nature* 2001, **411:**1046-1049
2. Pal C, Papp B and Hurst LD **Highly expressed genes in yeast evolve slowly** *Genetics* 2001, **158:**927-931
3. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C and Feldman MW **Evolutionary rate in the protein interaction network** *Science* 2002, **296:**750-752
4. Jordan IK, Wolf YI and Koonin EV **No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly** *BMC Evolutionary Biology* 2003, **3:**1
5. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S and Weil B **MIPS: a database for genomes and protein sequences** *Nucleic Acids Res* 2002, **30:**31-34
6. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M and Pochart P **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*** *Nature* 2000, **403:**623-627
7. Koski LB and Golding GB **The closest BLAST hit is often not the nearest neighbor** *Journal of Molecular Evolution* 2001, **52:**540-542
8. Wall DP, Fraser HB and Hirsh AE **An improved method for detecting putative orthologs** *Bioinformatics* 2003,
9. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y **A comprehensive two-hybrid analysis to explore the yeast protein interactome** *Proc Natl Acad Sci USA* 2001, **98:**4569-4574
10. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM and Cruciat CM **Functional organization of the yeast proteome by systematic analysis of protein complexes** *Nature* 2002, **415:**141-147
11. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K and Boutilier K **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry** *Nature* 2002, **415:**180-183

12. Jeong H, Mason SP, Barabasi AL and Oltvai ZN **Lethality and centrality in protein networks** *Nature* 2001, **411:**41-42
13. Gibbons JD **Nonparametric Measures of Association** *Newbury Park, UK: Sage* 1993,
14. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J and Schachter V **The protein-protein interaction map of *Helicobacter pylori*** *Nature* 2001, **409:**211-215