BMC Evolutionary Biology

Research article

Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein–protein interactions data sets

Jesse D Bloom*[1] and Christoph Adami[2]

Address: [1]Department of Chemistry and Digital Life Laboratory, 210-41, California Institute of Technology, Pasadena, CA 91125, USA and [2]Digital Life Laboratory and Jet Propulsion Laboratory, 136-93, California Institute of Technology, Pasadena, CA 91125, USA

Email: Jesse D Bloom* - bloom@caltech.edu; Christoph Adami - adami@caltech.edu

* Corresponding author

Published: 02 October 2003

BMC Evolutionary Biology 2003, **3**:21

This article is available from: http://www.biomedcentral.com/1471-2148/3/21

Received: 04 September 2003
Accepted: 02 October 2003

Abstract

Background: Several studies have suggested that proteins that interact with more partners evolve more slowly. The strength and validity of this association has been called into question. Here we investigate how biases in high-throughput protein–protein interaction studies could lead to a spurious correlation.

Results: We examined the correlation between evolutionary rate and the number of protein–protein interactions for sets of interactions determined by seven different high-throughput methods in *Saccharomyces cerevisiae*. Some methods have been shown to be biased towards counting more interactions for abundant proteins, a fact that could be important since abundant proteins are known to evolve more slowly. We show that the apparent tendency for interactive proteins to evolve more slowly varies directly with the bias towards counting more interactions for abundant proteins. Interactions studies with no bias show no correlation between evolutionary rate and the number of interactions, and the one study biased towards counting fewer interactions for abundant proteins actually suggests that interactive proteins evolve *more rapidly*. In all cases, controlling for protein abundance significantly decreases the observed correlation between interactions and evolutionary rate. Finally, we disprove the hypothesis that small data set size accounts for the failure of some interactions studies to show a correlation between evolutionary rate and the number of interactions.

Conclusions: The only correlation supported by a careful analysis of the data is between evolutionary rate and protein abundance. The reported correlation between evolutionary rate and protein–protein interactions cannot be separated from the biases of some protein–protein interactions studies to count more interactions for abundant proteins.

Background

Different proteins in the same organism evolve at different rates. An understanding of the factors that cause these differences in rates has important ramifications for genetics, molecular evolution, and evolutionary biology. Factors that are thought to influence a protein's evolutionary rate include its abundance [1], whether its function is encoded in a robust manner [2], and the amount of recombination that it undergoes [3]. Of these factors, abundance is the strongest correlate of evolutionary rate, and recent work has shown the importance of adequately controlling for abundance when examining correlations between evolutionary rate and other protein properties [4].

Another factor that has been suggested to influence a protein's evolutionary rate is its number of interaction partners, with recent studies claiming that interactive proteins evolve more slowly because they have more functionally constrained residues [5,6]. However, this reported association between evolutionary rate and the number of protein–protein interactions has proven controversial, with studies using different interactions data sets reaching different conclusions [7,6].

The original claim by Fraser *et al.* [5] that a protein's evolutionary rate depends on the number of different proteins it interacts with was based on a negative statistical correlation between evolutionary rate as determined from an alignment of orthologs, and the number of interactions as determined by pooling data from several studies. However, a second study by Jordan *et al.* [7] using different data sets for both protein–protein interactions and evolutionary rates failed to find a significant correlation between evolutionary rate and the number of interactions. A third study by Fraser *et al.* [6] using a much larger protein–protein interactions data set again found a correlation, and also showed that the conflicting results were due to differences in the interactions data sets rather than differences in the evolutionary rates. The authors of this last study suggested that Jordan *et al.* failed to observe a correlation because of an incomplete set of protein–protein interactions, yet they offered no explanation of why only some data sets should reveal a correlation.

The biophysical explanation proposed by Fraser *et al.* [5] for the tendency of proteins with more interactions to evolve more slowly was that interactive proteins have more residues involved in protein–protein interaction surfaces, and are therefore less tolerant of amino acid substitutions. However, an individual residue does not distinguish between contacts with other residues from the same or a different protein, so there is no obvious reason why residues involved in intermolecular contacts should be more evolutionary constrained than other residues with the same number of intramolecular contacts. Indeed, analysis of oligomeric proteins has shown that interacting residues are not under the strong selection constraints of enzymatic active site residues, but instead actually change more rapidly than typical core residues and only slightly more slowly than the average for the entire sequence [8]. On these grounds, one would expect the number of interaction partners to have at most a slight effect on the overall rate of sequence evolution, and that other factors such as the ratio of core to surface residues should be more important.

The sensitivity of the correlation between evolutionary rate and the number of interactions to the particular data set used, as well as the absence of a clear biophysical justification for why proteins with more interaction partners should evolve more slowly, prompted us to analyze the data more carefully. We find that the reported connection between evolutionary rate and the number of interactions is linked to the biases of some protein–protein interactions studies to count more interactions for abundant proteins.

Results and Discussion

Analysis of the different interactions data sets

Protein–protein interactions data for *S. cerevisiae* are derived from studies using a variety of distinct methods, each with its own strengths and weaknesses (for a comprehensive discussion, see [9]). In particular, several methods have been shown to be biased towards counting more interactions for abundant proteins [9]. Since abundant proteins are known to evolve more slowly [1], any examination of the relationship between interactions and evolutionary rate should control for biases towards counting more interactions for abundant proteins.

We compiled *S. cerevisiae* protein–protein interactions sets from nine studies using seven different high-throughput methods, taking data from two studies that identified interactions by mass spectrometry [10,11], two studies that identified interactions with the yeast two-hybrid system [12,13], and studies that identified interactions by correlated mRNA expression (synexpression) [9], identification of conserved gene neighborhoods [14,15,9], cooccurrence of genes in sequenced genomes [16,9], identification of gene fusion events [17,18,9], and synthetic lethality in knockouts [19,9]. The mass spectrometry studies [10,11] involved tagging and overexpression of one of the proteins, which may lead to non-native interactions, so for these studies we also compiled data sets that counted the interactions only for the untagged proteins. We also compiled a comprehensive list of all of the interactions from all studies, as well as the interactions found independently by two and three of the studies.

We also gathered information on the evolutionary rates and abundances of *S. cerevisiae* proteins. The sequence evolution rates were based on alignments with orthologs from *Candida albicans* compiled by Fraser *et al.* [6] according to the method of [20]. We used two established proxies for protein abundance: mRNA transcript levels from gene microarrays [21,22] and codon adaptation indices (CAI) calculated from gene sequences [23,24]. We used this information to create sets of proteins that participated

Table 1: The correlations among evolutionary rate, the number of interactions, and protein abundance for all studies when abundance is measured by (A) mi-croarray expression level or (B) CAI. N_p and N_i are the number of proteins and interactions for each data set. The Kendall's rank correlations between variables are given by τ_{EI}, τ_{AI}, and τ_{EA}. The Kendall's partial rank correlation between evolutionary rate and the number of interactions when abundance is controlled for is given by $\tau_{EI.A}$. All correlations have two-tailed significances of $P < 10^{-3}$ unless another P value is given in parentheses.

(A)

Study	N_p	N_i	τ_{EI}	τ_{AI}	τ_{EA}	$\tau_{EI.A}$
Ito [12]	505	1007	0.03 (0.30)	-0.03 (0.36)	-0.34	0.02 (0.51)
Uetz [13]	607	1183	0.01 (0.63)	-0.01 (0.67)	-0.35	0.01 (0.75)
2H studies [12,13]	893	2034	0.02 (0.29)	-0.02 (0.40)	-0.36	0.02 (0.48)
Gavin [10]	1039	15224	-0.12	0.09	-0.45	-0.09
Gavin [10] untagged	1018	8568	-0.08	0.09	-0.44	-0.04 (0.04)
Ho [11]	1183	5879	-0.18	0.15	-0.41	-0.13
Ho [11] untagged	991	2990	-0.28	0.30	-0.40	-0.18
MS studies [10,11]	1698	20708	-0.13	0.12	-0.42	-0.09
MS studies [10,11] untagged	1543	11424	-0.14	0.16	-0.42	-0.08
synexpression [9]	1114	19188	0.09	-0.12	-0.40	0.05 (0.02)
gene neighborhood [9]	765	9882	-0.10	0.15	-0.44	-0.04 (0.15)
synthetic lethality [9]	524	1463	0.02 (0.50)	0.01 (0.71)	-0.43	0.03 (0.40)
gene cooccurrence [9]	298	1718	-0.15	0.07 (0.08)	-0.36	-0.13 (0.002)
gene fusion [9]	222	535	0.04 (0.40)	-0.03 (0.51)	-0.37	0.03 (0.56)
all interactions	2846	54258	-0.16	0.13	-0.39	-0.12
two studies	1112	2792	-0.17	0.14	-0.42	-0.13
three studies	329	556	0.03 (0.43)	-0.03 (0.42)	-0.36	0.02 (0.65)

(B)

Study	N_p	N_i	τ_{EI}	τ_{AI}	τ_{EA}	$\tau_{EI.A}$
Ito [12]	528	1049	0.03 (0.34)	-0.04 (0.21)	-0.31	0.02 (0.61)
Uetz [13]	630	1213	0.01 (0.65)	-0.06 (0.02)	-0.32	-0.01 (0.78)
2H studies [12,13]	931	2104	0.02 (0.27)	-0.06 (0.003)	-0.33	0.00 (0.90)
Gavin [10]	1055	15836	-0.12	0.08	-0.38	-0.10
Gavin [10] untagged	1033	8648	-0.08	0.07	-0.37	-0.06 (0.01)
Ho [11]	1209	5941	-0.18	0.16	-0.42	-0.13
Ho [11] untagged	1013	3019	-0.28	0.33	-0.42	-0.16
MS studies [10,11]	1735	20930	-0.13	0.11	-0.40	-0.10
MS studies [10,11] untagged	1575	11531	-0.14	0.15	-0.40	-0.09
synexpression [9]	1163	20291	0.09	-0.09	-0.41	0.05 (0.06)
gene neighborhood [9]	790	10186	-0.09	0.08	-0.49	-0.06 (0.02)
synthetic lethality [9]	533	1505	0.03 (0.36)	-0.02 (0.60)	-0.38	0.02 (0.49)
gene cooccurrence [9]	309	1767	-0.14	0.07 (0.07)	-0.44	-0.13 (0.02)
gene fusion [9]	233	559	0.04 (0.41)	-0.01 (0.74)	-0.40	0.03 (0.50)
all interactions	2960	56058	-0.16	0.12	-0.38	-0.13
two studies	1131	2822	-0.17	0.08	-0.41	-0.15
three studies	332	562	0.03 (0.45)	-0.07 (0.06)	-0.41	0.00 (0.99)

in at least one interaction and for which evolutionary rate and abundance information were available; the size of the coverage sets for each interactions study is shown in Table 1.

We confirmed that abundant proteins evolved more slowly in all coverage sets (Table 1, Figure 1A), in agreement with established results [1]. The tendency for abundant proteins to evolve more slowly was both substantial and robust, with all coverage sets showing significant correlations (Kendall's τ ranged from -0.31 to -0.49, $P < 10^{-3}$) regardless of whether abundance was measured by expression level or CAI.

We also confirmed [9] that some of the interactions studies are biased towards counting more interactions for abundant proteins (Table 1, Figure 1B). Among the experimentally-based studies that look for direct evidence of interactions, the mass spectrometry studies [10,11] were consistently biased towards counting more interactions for abundant proteins (τ ranged from 0.07 to 0.33, $P < 10^{-3}$, Table 1), while the yeast two-hybrid studies [12,13] showed no substantial bias towards counting more

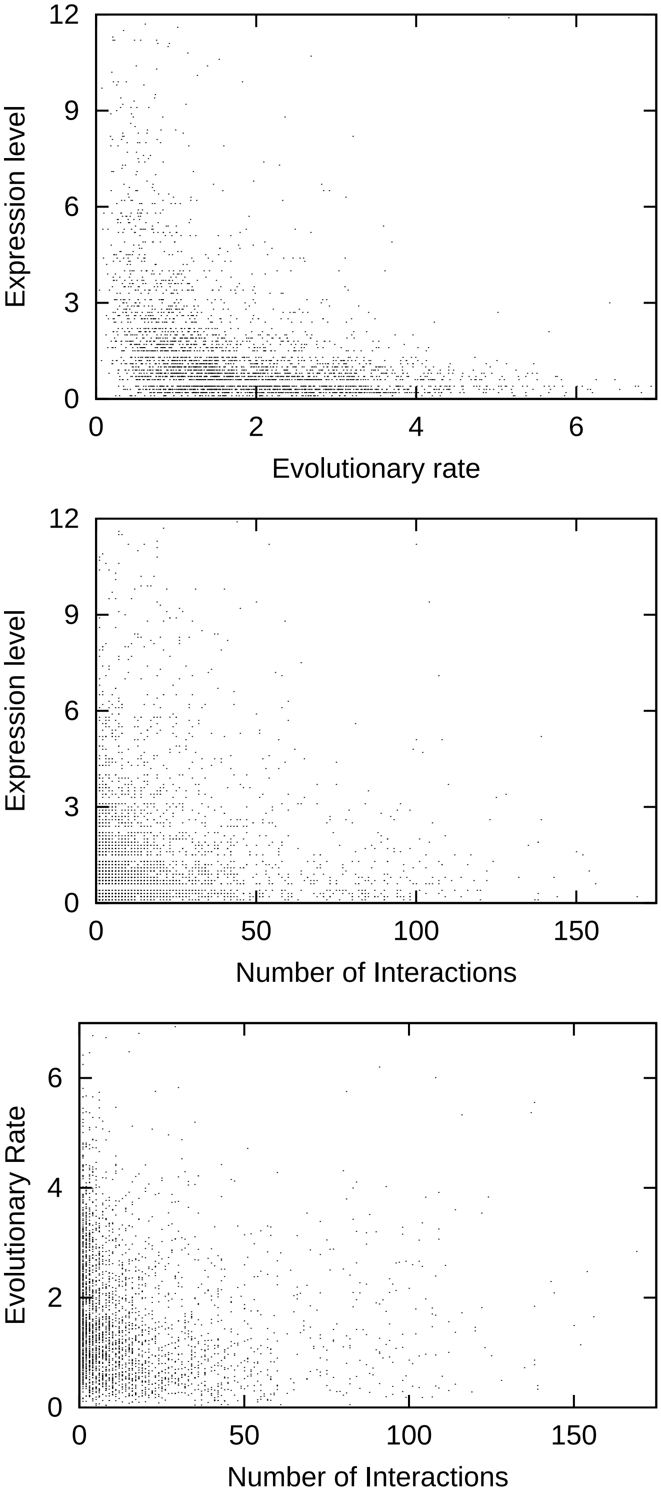

Figure 1
(A) shows the relationship between evolutionary rate and expression level as measured by gene microarrays [21]. (B) shows the relationship between expression level and the total number of interactions from all studies. (C) shows the relationship between evolutionary rate and the total number of interactions from all studies. Some outlying data points are not shown, but are included in the calculations of the correlations in Table 1.

interactions for abundant proteins ($P > 0.25$, Table 1). The existence of a bias in the mass spectrometry but not the yeast two-hybrid studies can be explained by considering the experimental methods. The yeast two-hybrid studies involve over-expression of both interacting proteins, and so the probability of observing an interaction is unrelated to a protein's native concentration. In contrast, in the mass-spectrometry studies only the tagged protein is over-expressed, and so the probability of observing an interaction depends on the choice of which proteins to tag, as well as the native concentrations of the untagged proteins.

Among the bioinformatics-based methods, the gene neighborhood data are substantially biased towards counting more interactions for abundant proteins (τ = 0.15 or 0.08, $P < 10^{-3}$, Table 1), the gene cooccurrence data are mildly biased towards counting more interactions for abundant proteins (τ = 0.07, P = 0.07 or 0.08, Table 1), while the synexpression data are actually biased towards counting fewer interactions for abundant proteins (τ = -0.12 or -0.09, $P < 10^{-3}$, Table 1). The synthetic lethality and gene fusion studies are unbiased with respect to protein abundance ($P > 0.5$, Table 1), as is the set of interactions found independently by three studies ($P > 0.05$, Table 1). The set of interactions found by two studies and the set of all interactions are both biased towards counting more interactions for abundant proteins (τ ranged from 0.15 to 0.19, $P < 10^{-5}$, Table 1), presumably because both of these sets are dominated by interactions found by the mass spectrometry studies (see Table 1 and discussion below).

We found that proteins with more interactions appeared to evolve more slowly only when the interactions data set was biased towards counting more interactions for abundant proteins (Table 1, Figure 1C). The yeast-two hybrid, the synthetic lethality, the gene fusion, and the interactions found by three studies are all unbiased with respect to abundance, and none of these data sets suggested any significant correlation between evolutionary rate and the number of interactions ($P > 0.25$ in all cases, Table 1). The mass spectrometry, the gene neighborhood, the gene cooccurrence, the interactions found by two studies, and the combined sets are all biased towards counting more interactions for abundant proteins, and data from all of these studies suggested that proteins with more interactions evolve more slowly (τ ranges from -0.08 to -0.28, $P < 10^{-3}$, Table 1). The synexpression data is biased towards counting fewer interactions for abundant proteins, and it suggests that proteins with more interactions actually evolve more rapidly (τ = 0.09, $P < 10^{-3}$, Table 1).

If the bias of some studies to count more interactions for abundant proteins explains the correlation between the number of interactions and the evolutionary rate, then there should be a direct relationship between the bias and the observed correlation. We examined this relationship for all 17 data sets in Table 1, and confirmed that there was a simple linear relationship between the correlation of abundance with the number of interactions and the correlation of the number of interactions with the evolutionary rate, as shown in Figure 2.

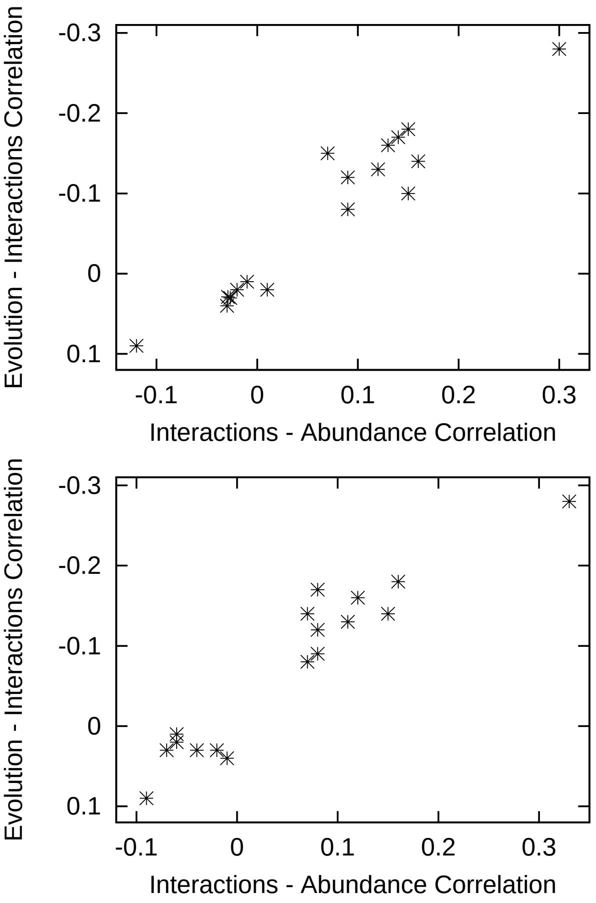

Figure 2
The correlation between evolutionary rate and the number of interactions is directly related to the bias towards counting more interactions for abundant proteins, both when abundance is measured by (A) gene microarray expression levels and (B) CAI. Correlations are Kendall's rank correlation τ, and points are for all data sets listed in Table 1.

The trends described here are not sensitive to the evolutionary rates used. When evolutionary rates are derived from alignments of *S. cerevisiae* and *Schizosaccharomyces pombe* orthologs by Fraser *et al.* [6], there is again a consistent correlation between evolutionary rate and abundance, but a correlation between evolutionary rate and

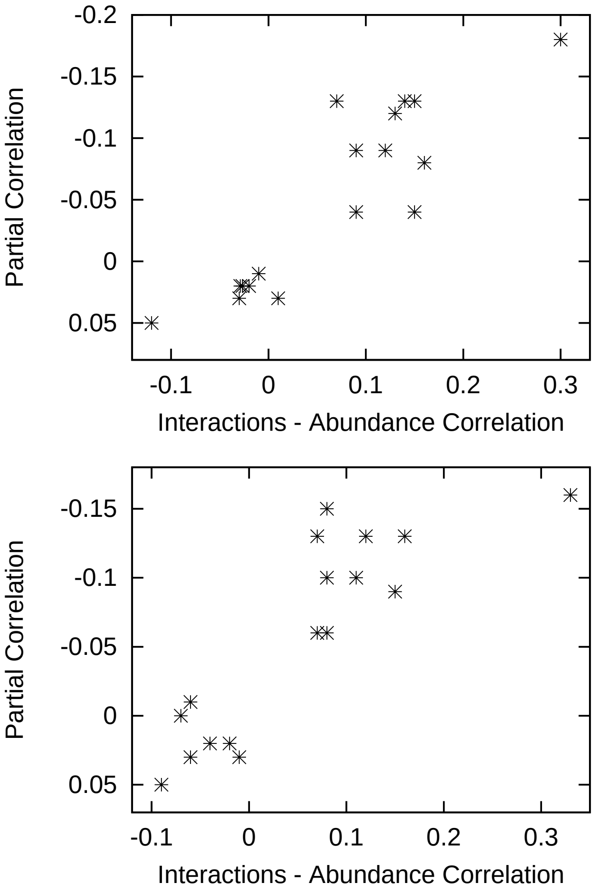

Figure 3
Controlling for abundance reduces the magnitude of the correlations between evolutionary rate and the number of interactions from those shown in Figure 2, and the remaining partial correlation still depends on the bias towards counting more interactions for abundant proteins, both when abundance is measured by (A) gene microarray expression levels and (B) CAI. The partial correlations are Kendall's partial τ, the correlation between interactions and abundance is Kendall's rank correlation τ, and points are for all data sets listed in Table 1.

interactions emerges only for interactions data sets biased towards counting more interactions for abundant proteins (data not shown).

Controlling for bias reduces apparent correlation between evolutionary rate and interactions

The relationship between the correlation of evolutionary rate with the number of interactions and the bias towards counting more interactions for abundant proteins (Figure 2) suggests that the bias contributes to the observed correlation. To obtain a statistical view of this effect, we used a partial correlation statistic (Kendall's partial τ) to measure the correlation between evolutionary rate and the number of interactions when protein abundance is controlled for. In all data sets where there is a significant correlation between evolutionary rate and the number of interactions, controlling for protein abundance reduces the magnitude of the correlation (Table 1). We determined the significance of this reduction by performing 10^4 randomizations of the protein abundances. In none of the cases where there was a highly significant correlation between evolutionary rate and the number of interactions (the mass spectrometry, synexpression, gene neighborhood, gene cooccurrence, two study, and combined data sets) did the randomized abundances give a partial τ with a magnitude as small as for the actual data, demonstrating that the reductions in the correlation due to controlling for abundance were highly significant ($P < 10^{-4}$).

Although controlling for protein abundance always reduces the magnitude of any significant correlation between evolutionary rate and the number of interactions, in some cases the remaining partial correlation is still statistically significant. However, this remaining correlation appears to be due to an incomplete correction for protein abundance rather than a real correlation between evolutionary rate and the number of interactions. As Figure 3 shows, the remaining partial correlation between evolutionary rate and the number of interactions is still directly related to the bias towards counting more interactions for abundant proteins, suggesting that this bias is still the primary factor causing the partial correlation. Note also that the partial correlation between evolutionary rate and the number of interactions for the synexpression data set still suggests that proteins with more interactions evolve more rapidly, again suggesting that the partial correlation statistic does not completely correct for biases in the interactions data set.

There are several reasons why the partial correlation statistic may be unable to completely correct for experimental biases. Both microarray expression data and CAI are imperfect proxies for true protein abundance (indeed, the Spearman correlation between these two proxies is only 0.62) [22,24], and so statistically controlling with these variables does not completely correct for effects due to actual protein abundances or expression levels. In addition, the evolutionary rates and expression data for the large set of proteins considered here may underestimate the true tendency for abundant proteins to evolve more slowly. Pal *et al* [1] analyzed the correlation between evolutionary rate and protein abundance using a carefully culled set of well-characterized proteins, and reported Pearson correlations of evolutionary rate with the logarithm of microarray expression levels and with CAI of -

0.584 and -0.617 respectively ($P < 10^{-6}$). In comparison, the same Pearson correlations are substantially smaller (-0.423 and -0.356 respectively, $P < 10^{-6}$) for the set of all proteins considered here, possibly because the larger set of proteins here necessitates using less clean data. Such an underestimation of the strength of the relationship between evolutionary rate and abundance would cause the partial correlation statistic to incompletely correct for the bias. The fact that the remaining partial correlation still directly depends on the extent of the bias is evidence for this incomplete correction.

In addition, the different native concentrations of proteins is only one source of bias in the counting of interactions by the mass spectrometry studies. There also is an inherent asymmetry in the counting of interactions in the mass-spectrometry studies because some proteins are tagged and over-expressed while others are only present at their native levels. If the experimenter tends to select more abundant proteins for tagging, biases towards counting more interactions for abundant proteins would be amplified in a way that cannot be controlled for by transcript level. One way to examine this effect is to only consider interactions for the untagged proteins in the mass spectrometry studies. When this is done for study [10], the bias to count more interactions for abundant proteins is slightly reduced and there is a concomitant decrease in the association between evolutionary rate and the number of interactions (Table 1). But when this is done for study [11], the bias to count more interactions for abundant proteins increases and the association between evolutionary rate and the number of interactions becomes larger (Table 1). Therefore, the effect of the experimental choice of tagged proteins differs between the studies, but in both cases, an increased tendency to count more interactions for abundant proteins increases the apparent correlation between evolutionary rate and interactions.

Protein–protein interactions and evolutionary rates in bacteria

We suggest a simple explanation for the failure of a previous analysis to observe a correlation between evolutionary rate and the number of interactions in the bacteria *Helicobacter pylori* [7]. This analysis was based on protein–protein interactions data obtained from a yeast two-hybrid study [25], and so based on our analysis here we would expect this data to have no bias towards counting-more interactions for abundant proteins, and therefore to show no correlation between evolutionary rate and the number of interactions.

Data set size or accuracy are not plausible explanations for absence of correlation

The most recent study by Fraser *et al.* [6] claiming a correlation between evolutionary rate and the number of interactions suggested that the correlation may not be apparent if the interactions data set is too small, and stresses the importance of always using the largest possible data set. In order to evaluate this claim, we investigated the effect of data set size on the correlation between evolutionary rate and the number of interactions.

If the dependence of evolutionary rate on the number of interactions only becomes obvious for large interactions data sets, we would expect that larger data sets would show a greater correlation. Figure 4(A) shows how the correlation depends on the size of the interactions data set. There is no obvious trend of larger data sets yielding a larger correlation – indeed, the strongest correlation is found using relatively small data sets with strong biases towards counting more interactions for abundant proteins.

In order to investigate how the spread in observed correlations between evolutionary rate and the number of interactions would be expected to depend on data set size if the bias towards counting more interactions for abundant proteins was unimportant, we performed sampling simulations on the set of all interactions mimicking both the methods of the mass spectrometry studies (counting all interactions for selected proteins) and the yeast two-hybrid studies (counting only interactions between pairs of selected proteins). The results of these simulations are shown in Figure 4(B) – they show that the observed correlation should be roughly constant regardless of the interactions data set size. Although the spread does increase for smaller data sets, this increase is not large enough to explain the observed spread in correlations. This demonstrates that differences in the data set sizes or sampling methods do not explain the variation in the observed correlations.

The inadequacy of data set size as an explanation for the failure to observe a correlation for some sets is most obvious in a comparison of Figures 2 and 4(A). Data set size bears no clear relationship to the correlation between evolutionary rate and the number of interactions, but the experimental bias towards counting more interactions for abundant proteins is an excellent predictor of this correlation.

We also considered the possibility that the accuracy of the interactions data might affect the strength of the observed correlation. In their review of protein interactions studies, von Mering and coworkers [9] provide estimates of the accuracies of the different studies. According to their measure of accuracy, synthetic lethality is the single most accurate method for detecting interactions, interactions detected by two different studies are more accurate than those detected by any one study, and interactions detected

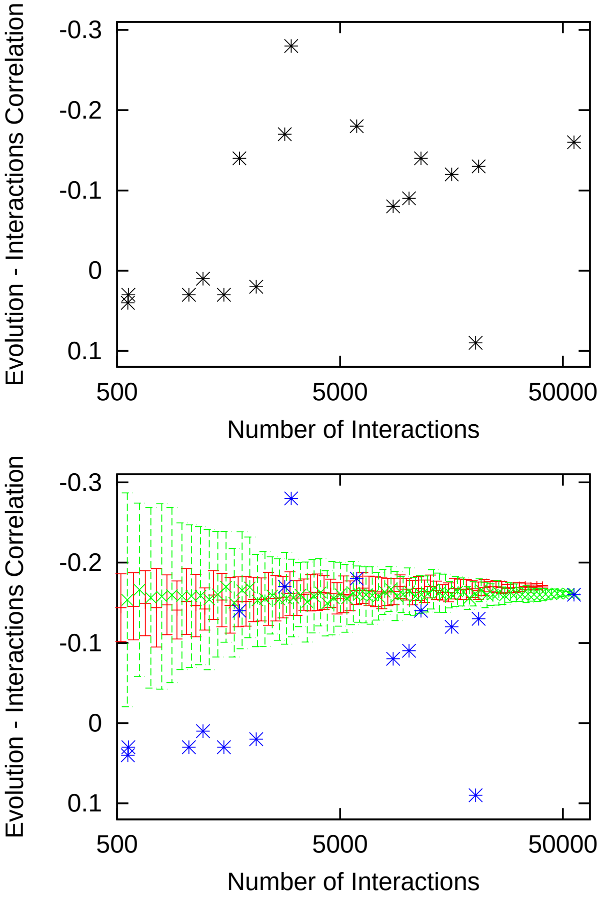

Figure 4
The correlation between evolutionary rate and the number of interactions does not depend on the size of the interactions data set as it would in the absence of bias in the counting of interactions. (A) shows the correlation and data set sizes for all sets in Table 1. (B) shows how the mean and standard deviation of the correlation should depend on the data set size in the absence of experimental bias in the counting of interactions, based on sampling simulations of the mass spectrometry (green) and yeast two-hybrid (red) method of counting interactions.

by three studies are more accurate still [9]. The set of interactions detected by two different studies does show a tendency for more interactive proteins to evolve more slowly, however it is also strongly biased towards counting more interactions for abundant proteins (Table 1). This can be explained by noting that over 54% of the interactions in this data set were identified only by the two mass spectrometry studies, and that for 69% of the interactions in this set, one of the two identifications was by a mass spectrometry study. When this heavy slant towards the mass spectrometry studies is ameliorated by requiring the interactions to be identified by three different studies (meaning that at least one of the studies must use a method other than mass spectrometry), both the bias towards counting more interactions for abundant proteins and the tendency of interactive proteins to evolve more slowly disappear (Table 1). The data from the synthetic lethality method show no bias towards counting more interactions for abundant proteins and no tendency for abundant proteins to evolve more slowly (Table 1). We also note that Jordan *et al* [7] observed no significant correlation between evolutionary rate and the number of interactions when they used a set of manually curated interactions that might be expected to be of higher accuracy than those from any single high-throughput method. Therefore, the accuracy of the interactions data does not appear to explain the apparent correlation between evolutionary rate and the number of interactions.

Conclusions

We have examined the relationship among evolutionary rate, protein abundance, and the number of protein–protein interactions for data from different high-throughput studies. We have shown that while there is a consistent tendency for abundant proteins to evolve more slowly, proteins with more interactions only appear to evolve more slowly when using interactions data from studies biased towards counting more interactions for abundant proteins. The strength of the correlation between evolutionary rate and the number of interactions is directly dependent on the strength of the bias towards counting more interactions for abundant proteins – when there is no bias, there is no correlation, and in the one case where the bias is towards counting fewer interactions for abundant proteins, interactive proteins actually appear to evolve more rapidly instead. We have shown that this effect is not explained by the size or accuracy of the interactions data sets. This suggests that the apparent tendency of interactive proteins to evolve more slowly is due to the fact that abundant proteins evolve more slowly, combined with a bias towards counting more interactions for abundant proteins.

Our findings underscore the importance of considering experimental methods when analyzing biological data. The failure of Jordan *et al.* [7] to observe a correlation between evolutionary rate and the number of interactions in a data set of several thousand interactions should have raised a red flag, yet the approach of Fraser *et al.* [6] was simply to pool all available data and recalculate the correlations. But while pooling data may yield higher statistical confidences, statistics are only as good as the quality of the data to which they are applied. In our analysis of data from individual studies, it appears that the correlation is contingent on a bias towards counting more interactions for abundant proteins. Since this bias cannot be properly controlled for with the presently available data, there is no basis to conclude that there is any association between evolutionary rate and the number of interactions.

Recent advances in genomic and proteomic technologies are providing vast amounts of information about proteins and genes, including their sequences and chromosomal locations, expression levels [21], recombination rates [26], functions and dispensability [27], evolutionary rates [28], and interactions [9]. Many of these properties are interdependent, and in addition many of the high-throughput studies are subject to systematic biases. A major challenge of bioinformatics is to adequately correct for these interdependencies and biases in order to extract meaningful trends from the available data sets [4]. We have shown here how careful consideration of the biases of individual studies can explain correlations in pooled biological data sets.

Methods

Gathering of Data

Protein evolutionary rate data were obtained from Fraser [6] compiled according to the method of [20], and are based on the alignment of *S. cerivisiae* and *C. albicans* orthologs. Information on gene expression was taken from [21], where the authors have estimated the number of mRNA molecules per cell based on microarray analysis of yeast grown to the mid-log phase in YPD (yeast-extract, peptone, dextrose) media and presented this data online at http://web.wi.mit.edu/young/pub/data/orf_transcriptome.txt. CAI for the yeast genes were calculated [23] using gene sequences from the MIPS yeast database [29]. Mass spectrometry protein–protein interaction data from [10] were parsed from Table S3 of the supplementary material, counting only binary interactions between the tagged and untagged proteins in a complex.

Mass spectrometry protein–protein interaction data from [11] were taken from http://www.mdsp.com/yeast/, again counting interactions as binary between the tagged and untagged proteins in a complex. The mass spectrometry data set in Table 1 were the combined results of these two studies [10,11]. In the untagged-only mass spectrometry data sets for these studies, the interactions were counted only for the untagged proteins in a complex. Yeast two-hybrid protein–protein interaction data from [13] were parsed from Table 2 of the paper. Yeast two-hybrid protein–protein interaction data from [12] were downloaded from the core data list at http://genome.c.kanazawa-u.ac.jp/Y2H/. The yeast two-hybrid data set in Table 1 was the combined results of these two studies [13,12]. The high confidence, synexpression, gene neighborhood, synthetic lethality, gene cooccurrence, and gene fusion interactions data sets were parsed from supplementary Table 4 of [9]. The combined data sets of all proteins was formed from all interactions from these nine studies. The sets of interactions found by two and three of these studies were independently listed in at least that many of the nine studies. In the interaction counts listed in Table 1, a binary interaction was counted once for each partner except for self-interactions, in which case the interaction was only counted once. The interaction counts given in Table 1 are the sums of the number of interactions assigned to all proteins in the data sets for which both evolutionary rate and abundance (expression or CAI) information was available. When combining interactions data sets, duplicate interactions were removed. All data will be made available upon request.

Statistical Analysis

Statistical analyses were performed using Kendall's τ rank correlation coefficients and two-tailed P values were calculated as described in [30]. Briefly, the Kendall's correlation between x and y was calculated as $\tau_{xy} = \frac{2(C-D)}{\sqrt{(n^2-n-2T)(n^2-n-2U)}}$ where C is the number of concordant pairs, D is the number of discordant pairs, n is the number of pairs, and $T = \frac{1}{2}\sum(t^2-t)$ and $U = \frac{1}{2}\sum(u^2-u)$ are corrections for tied values computed by summing over the number of observations t and u that are tied at any given value for the x and y data sets respectively. Kendall's partial correlation between x and y controlling for z was calculated as $\tau_{xy.z} = \frac{\tau_{xy} - \tau_{xz}\tau_{yz}}{\sqrt{(1-\tau_{xz}^2)(1-\tau_{yz}^2)}}$. For Kendall's partial τ correlation, two-tailed P values were calculated using 10^4 randomizations of the abundances and the evolutionary rates. The calculations of the significances of the change in Kendall's partial τ correlation were performed by determining what fraction of 10^4 randomizations of the abundances (preserving the interactions and evolutionary rates) yielded an increase or decrease in the partial τ larger than that observed for the actual data.

For the sampling simulations, we began with a list of all non-duplicate interactions from the combined data set. For the mass-spectrometry simulation, we randomly selected n proteins and all of their interactions to add to the interactions sample set, where n was iterated from 3341 to 10, performing $\frac{3341}{n}$ trials at each n. For the yeast two-hybrid simulation, we selected proteins in the same way, but only counted an interaction if both of the proteins participating in the interaction were among the selected proteins. Kendall's τ correlation between the evolutionary rate and the number of interactions for each sample set was calculated, and results were binned according to the total number of interactions in the sample set into bins of exponentially scaled size with centers shown in Figure 4(B). The mean and standard deviation of the correlation were calculated for each bin.

Authors' Contributions

JDB gathered the data, performed the statistical analysis, and wrote the manuscript. CA provided guidance on the analysis and edited the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

We thank Frances H. Arnold for helpful comments and advice. We also thank an anonymous reviewer for insightful comments that greatly improved our work. JDB is supported by a Howard Hughes Medical Institute Predoctoral Fellowship. CA is supported by the NSF under contract number DEB-9981397. Part of this work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

1. Pal C, Papp B and Hurst LD: **Highly expressed genes in yeast evolve more slowly.** *Genetics* 2001, **158:**927-931.
2. Wilke CO and Adami C: **Evolution of mutational robustness.** *Mut Res* 2003, **523:**3-11.
3. Pal C, Papp B and Hurst LD: **Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer.** *Mol Biol Evol* 2001, **18:**2323-2326.
4. Pal C, Papp B and Hurst LD: **Rate of evolution and gene dispensability.** *Nature* 2003, **421:**496-498.
5. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C and Feldman MW: **Evolutionary rate in the protein interaction network.** *Science* 2002, **296:**750-752.
6. Fraser HB, Wall DP and Hirsh AE: **A simple dependence between protein evolution rate and the number of protein–protein interactions.** *BMC Evol Biol* 2003, **3:**11.
7. 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 Evol Biol* 2003, **3:**1.
8. Grishin NV and Phillips MA: **The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences.** *Protein Sci* 1994, **3:**2455-2458.
9. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P: **Comparative assessment of large–scale data sets of protein–protein interactions.** *Nature* 2002, **417:**399-403.
10. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM and Cruciat CM *et al.*: **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 *et al.*: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spec-trometry.** *Nature* 2002, **415:**180-183.
12. Ito T, Chiba T, Ozawa T, Yoshida M, Hattoria M and Sakaki Y: **A comprehensive two–hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci USA* 2001, **98:**4569-4574.
13. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M and Pochart P *et al.*: **A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403:**623-627.
14. Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N: **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896-2901.
15. Huynen M, Snel B, Lathe W III and Bork P: **Predicting protein functioin by genomic context: quantatitive evaluation and qualitative inferences.** *Genome Res* 2000, **10:**1204-1210.
16. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96:**4285-4288.
17. Enright AJ, Iliopoulos I, Kyrpides NC and Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402:**86-90.
18. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO and Eisenberg D: **Detecting protein function and protein–protein interactions from genome sequences.** *Science* 1999, **285:**751-753.
19. Tong AHY, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizedeh S, Hogue CWV, Bussey H, Andrews B, Tyers M and Boone C: **Systematic genetic analysis of ordered arrays of yeast deletion mutants.** *Science* 2001, **294:**2364-2368.
20. Wall DP, Fraser HB and Hirsh AE: **Detecting putative orthologs.** *Bioinformatics* 2003, **19:**1710-1711.
21. Holstege FCP, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES and Young RA: **Dissecting the regulatory circuitry of a eukaryotic genome.** *Cell* 1998, **95:**717-728.
22. Gygi SP, Rochon Y, Franza BR and Aebersold R: **Correlation between protein and mRNA abundance in yeast.** *Mol Cell Biol* 1999, **19:**1720-1730.
23. Sharp PM and Li WH: **The codon adaptation index – a measure of directional synonomous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15:**1281-1295.
24. Coghlan A and Wolfe KH: **Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*.** *Yeast* 2000, **16:**1131-1145.
25. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J and Schachter V *et al.*: **The protein–protein interaction map of *Helicobacter pylori*.** *Nature* 2001, **409:**211-215.
26. Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO and Petes TD: **Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccaromyces cerevisiae*.** *Proc Natl Acad Sci USA* 2000, **97:**11383-11390.
27. Winzeler EA, Shoemaker DD, Astromoff A and Liang H: **Function characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis.** *Science* 1999, **285:**901-906.
28. Brookfield RFY: **What determines the rate of sequence evolution?** *Curr Biol* 2000, **10:**R410-R411.
29. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S and Well B: **MIPS: a databse for genomes and protein sequences.** *Nucleic Acids Res* 2002, **30:**31-34.
30. Gibbons JD: **Nonparametric Measures of Association.** In *Quantitative Applications in the Social Sciences Volume 91. Sage Publications*; 1993.