BMC Evolutionary Biology

BioMed Central

Research article

No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly

I King Jordan, Yuri I Wolf and Eugene V Koonin*

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

Email: I King Jordan - jordan@ncbi.nlm.nih.gov; Yuri I Wolf - wolf@ncbi.nlm.nih.gov; Eugene V Koonin* - koonin@ncbi.nlm.nih.gov

* Corresponding author

Published: 6 January 2003

BMC Evolutionary Biology 2003, **3**:1

Received: 10 November 2002

Accepted: 6 January 2003

This article is available from: http://www.biomedcentral.com/1471-2148/3/1

Abstract

Background: It has been suggested that rates of protein evolution are influenced, to a great extent, by the proportion of amino acid residues that are directly involved in protein function. In agreement with this hypothesis, recent work has shown a negative correlation between evolutionary rates and the number of protein-protein interactions. However, the extent to which the number of protein-protein interactions influences evolutionary rates remains unclear. Here, we address this question at several different levels of evolutionary relatedness.

Results: Manually curated data on the number of protein-protein interactions among *Saccharomyces cerevisiae* proteins was examined for possible correlation with evolutionary rates between *S. cerevisiae* and *Schizosaccharomyces pombe* orthologs. Only a very weak negative correlation between the number of interactions and evolutionary rate of a protein was observed. Furthermore, no relationship was found between a more general measure of the evolutionary conservation of *S. cerevisiae* proteins, based on the taxonomic distribution of their homologs, and the number of protein-protein interactions. However, when the proteins from yeast were assorted into discrete bins according to the number of interactions, it turned out that 6.5% of the proteins with the greatest number of interactions evolved, on average, significantly slower than the rest of the proteins. Comparisons were also performed using protein-protein interaction data obtained with high-throughput analysis of *Helicobacter pylori* proteins. No convincing relationship between the number of protein-protein interactions and evolutionary rates was detected, either for comparisons of orthologs from two completely sequenced *H. pylori* strains or for comparisons of *H. pylori* and *Campylobacter jejuni* orthologs, even when the proteins were classified into bins by the number of interactions.

Conclusion: The currently available comparative-genomic data do not support the hypothesis that the evolutionary rates of the majority of proteins substantially depend on the number of protein-protein interactions they are involved in. However, a small fraction of yeast proteins with the largest number of interactions (the hubs of the interaction network) tend to evolve slower than the bulk of the proteins.

Background

Rates of protein evolution vary greatly and may be influenced by a variety of factors. Recently, it has been demonstrated that the magnitude of the fitness effects associated with deleterious mutations in protein-coding genes (i.e. proteins' dispensability) correlates with rates of protein evolution [1,2]. Essential proteins or those that are less dispensable to an organism tend to evolve slower than those that are more dispensable. It has also been suggested that proteins' evolutionary rates are determined by the proportion of amino-acids that are critical to their function [3]. According to this intuitively plausible notion, proteins with a greater fraction of amino acid residues that play an essential role in the protein's function are predicted to evolve slower than those with a smaller fraction of such crucial residues. Consistent with this prediction, a negative correlation has been reported between protein evolutionary rates, which were determined from evolutionary distances between orthologous proteins from yeast *Saccharomyces cerevisiae* and the nematode *Caenorhabditis elegans*, and the number of protein-protein interactions (i.e., physical interactions determined, primarily, using the yeast two-hybrid system) proteins are involved in [4]. Yeast proteins that have a large number of interacting partners were found to have evolved slower, on average, than those with fewer interacting partners, and this was presumed to be due to the fact that proteins with more interacting partners have a greater fraction of residues directly involved in function. However, these same data indicate that less than 6% of the variance in evolutionary rates is explained by the variance in the number of protein-protein interactions, suggesting that the influence of the number of interacting partners on protein evolutionary rates might not be substantial. We sought to further investigate this phenomenon by examining the relationship between the number of protein-protein interacting partners and protein evolutionary rates for the yeasts *S. cerevisiae* and *Schizosaccharomyces pombe* as well as for the proteobacteria *Helicobacter pylori* and *Camplyobacter jejuni*.

Results and Discussion

Evolutionary rates and protein-protein interactions: yeast

A total of 1,879 pairs of orthologous proteins, one from *S. cerevisiae* and one from *S. pombe*, were identified (see Methods), and for 1,004 of these, there was data on protein-protein interactions of the *S. cerevisiae* member in the MIPS database [5]. For these 1,004 orthologous pairs, the number of protein-protein interactions detected for the *S. cerevisiae* protein was plotted against the calculated substitution rates between orthologs (Figure 1a). As with a previous survey that compared conserved *S. cerevisiae* and *C. elegans* orthologs [4], there is a negative correlation between the number of protein-protein interactions and the evolutionary rates. However, although this correlation is statistically significant (Table 1), the slope of the linear trend line (y = -0.012) fit to the data by least squares regression as well as the small r^2 value ($r^2 = 0.0065$) suggest that the influence of the number of interacting partners on rates of evolution is minor at best. Specifically, the r^2 value indicates that less than 1% of the variation in substitution rates between orthologous proteins is explained by the variation in the number of protein-protein interactions. Furthermore, when only the most conserved (≥ 40% sequence identity), and thus most reliably identified, pairs of orthologous proteins were considered, the slope of the linear trend line as well as the r^2 value decreased and the statistical significance disappeared (Figure 1b and Table 1). To account for the possibility that linear regression does not adequately reflect the structure of the data and the observed low correlation is due to a non-linear relationship between the number of interactions and evolutionary rate of a protein, we also calculated the rank correlation coefficients for these quantities. Under this approach, no statistically significant correlation was observed for either of the two analysed data sets (Table 1).

It is tempting to speculate that the difference between the results obtained here and those reported previously [4] can be attributed to the difference in the evolutionary relationships between the pairs of species compared in the two studies. The species compared here, *S. cerevisiae* and *S. pombe*, are much more closely related than *S. cerevisiae* and *C. elegans*, and orthologous proteins are likely to be more reliably inferred between the closely related genomes. However, we also performed comparisons for pairs of orthologous proteins identified between the more distantly related *S. cerevisiae* and *C. elegans* [6] and no significant relationship between evolutionary rates and protein-protein interactions was observed (data not shown).

Long-term evolutionary conservation and protein-protein interactions: yeast

To examine the relationship between protein-protein interactions and evolutionary conservation of proteins over longer periods of time, the numbers of interactions for *S. cerevisiae* proteins were assessed against the taxonomic distribution of their homologs, which were detected using BLAST searches of the Genbank non-redundant protein database with expect value ≤ 10^{-3}. Five distinct levels of taxonomic distribution categories, each including taxa that are successively more distant from *S. cerevisiae*, were considered: 1 – hits only to ascomycetes, 2 – hits to non-ascomycete fungi, 3 – hits to metazoa and plants, 4 – hits to non-crown-group eukaryotes, 5 – hits to archaea and/or bacteria. The broader the taxonomic distribution of homologs of a *S. cerevisiae* protein the more evolutionarily conserved it is considered to be. Each *S. cerevisiae* protein was assigned a taxonomic distribution category, and this value was compared to the number of protein-protein in-

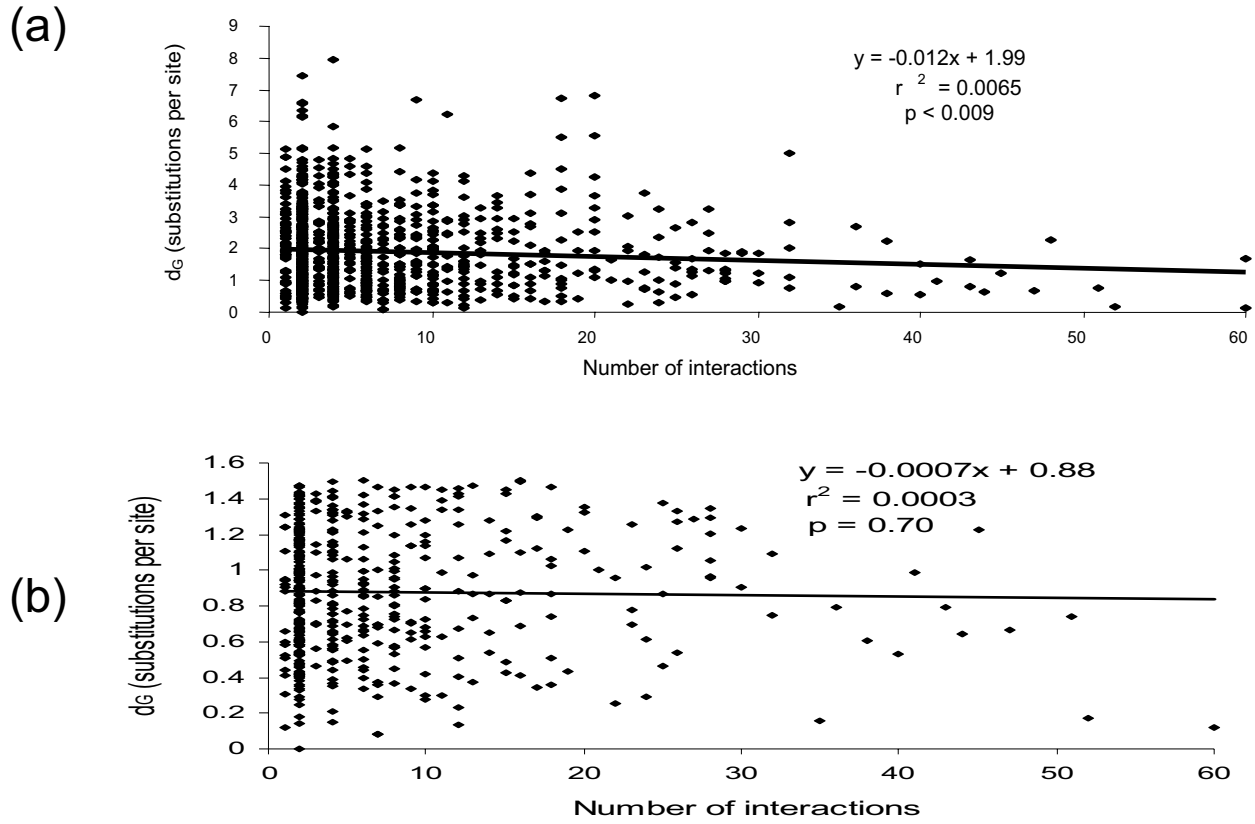

Figure 1

The relationship between the number of protein-protein interactions for *S. cerevisiae* proteins and the evolutionary rates between *S. cerevisiae* and *S. pombe* orthologs. Shown for each plot is the equation that describes the linear trend line, the r^2 value that describes the fraction of the variability in the evolutionary rates that is accounted for by the variability in the number of protein-protein interactions and the p value, which is the probability that the correlation between the number of protein-protein interactions and evolutionary rates could be due to chance. (a) All 1,004 observations. (b) 465 observations that correspond to orthologous protein pairs with ≥ 40% amino acid sequence identity.

teractions reported for the given protein. Correlation between these two features of *S. cerevisiae* proteins was not statistically significant ($r^2 = 0.007$, $p = 0.39$). Thus, as with the comparison between evolutionary rates and the number of interactions, no substantial relationship between long-term evolutionary conservation of *S. cerevisiae* proteins and the number of interactions was found.

Evolutionary rates and protein-protein interactions: bacteria

High throughput analysis of protein-protein interactions has also been conducted [7] on the proteobacterium H. pylori (the causative agent of gastric ulcers), for which complete genome sequences of two strains are available [8,9]. Thus it is possible to assess the effect of protein-protein interactions on the rates of evolution over much shorter periods of time (within species) compared to the analysis of the yeast proteins described above. Towards this end, orthologs between the two completely sequenced H. pylori strains were identified and the substitution rates between pairs of orthologous proteins were calculated (see Methods). The number of protein-protein interactions was plotted against the amino acid substitution rates and no significant relationship between the two was detected (Figure 2a and Table 1). The same conclusion was reached when the rank correlation coefficient was determined (Table 1). In this case, the lack of correlation between evolutionary rates and the number of interacting partners might simply be due to the small amount of evolutionary diversification that has occurred since the two H. pylori strains separated from their common ancestor. To evaluate this possibility, orthologous protein pairs

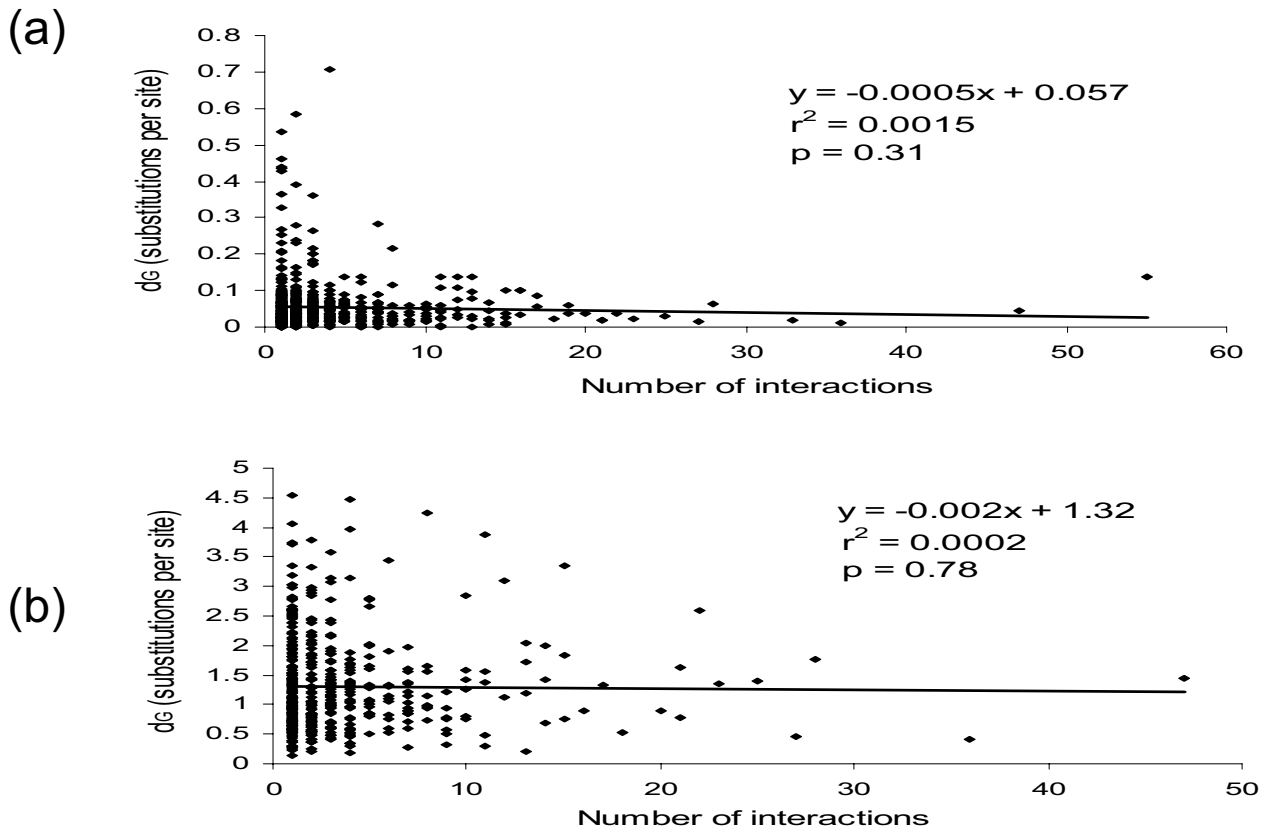

Figure 2
The relationship between the number of protein-protein interactions for *H. pylori* and the evolutionary rates between (a) *H. pylori* strain 26695 and *H. pylori* strain J99 orthologs and (b) *H. pylori* strain 26695 and *C. jejuni* orthologs. The values shown in each plot are the same as in Figure 1.

Table 1: Correlation between the number of protein-protein interactions and the evolutionary rate

Data set	Linear correlation coefficient (r)/ P-value	Rank correlation coefficient (R)/P-value
S. cerevisiae – *S. pombe* (all orthologs, N = 1044)	-0.081/0.009	-0.029/0.352
S. cerevisiae – *S. pombe* (only orthologs with >40% identity, N = 465)	-0.018/0.697	0.074/0.111
H. pylori J99 – *H. pylori* 26695 (N = 672)	-0.039/0.310	0.020/0.610
H. pylori – *C. jejuni* (N = 458)	-0.013/0.787	0.015/0.747

were identified between H. pylori and a more distantly related bacterium, C. jejuni [10]. These two species are close enough (both belong to the epsilon subdivision of proteobacteria) to ensure accurate identification of orthologs, but distant enough for substantial sequence divergence to have accumulated between orthologs. Nevertheless, comparison between these two bacteria showed no discernable correlation between the number of protein-protein interactions and the rates of substitution between orthologs, measured either directly or using the rank correlation approach (Figure 2b and Table 1).

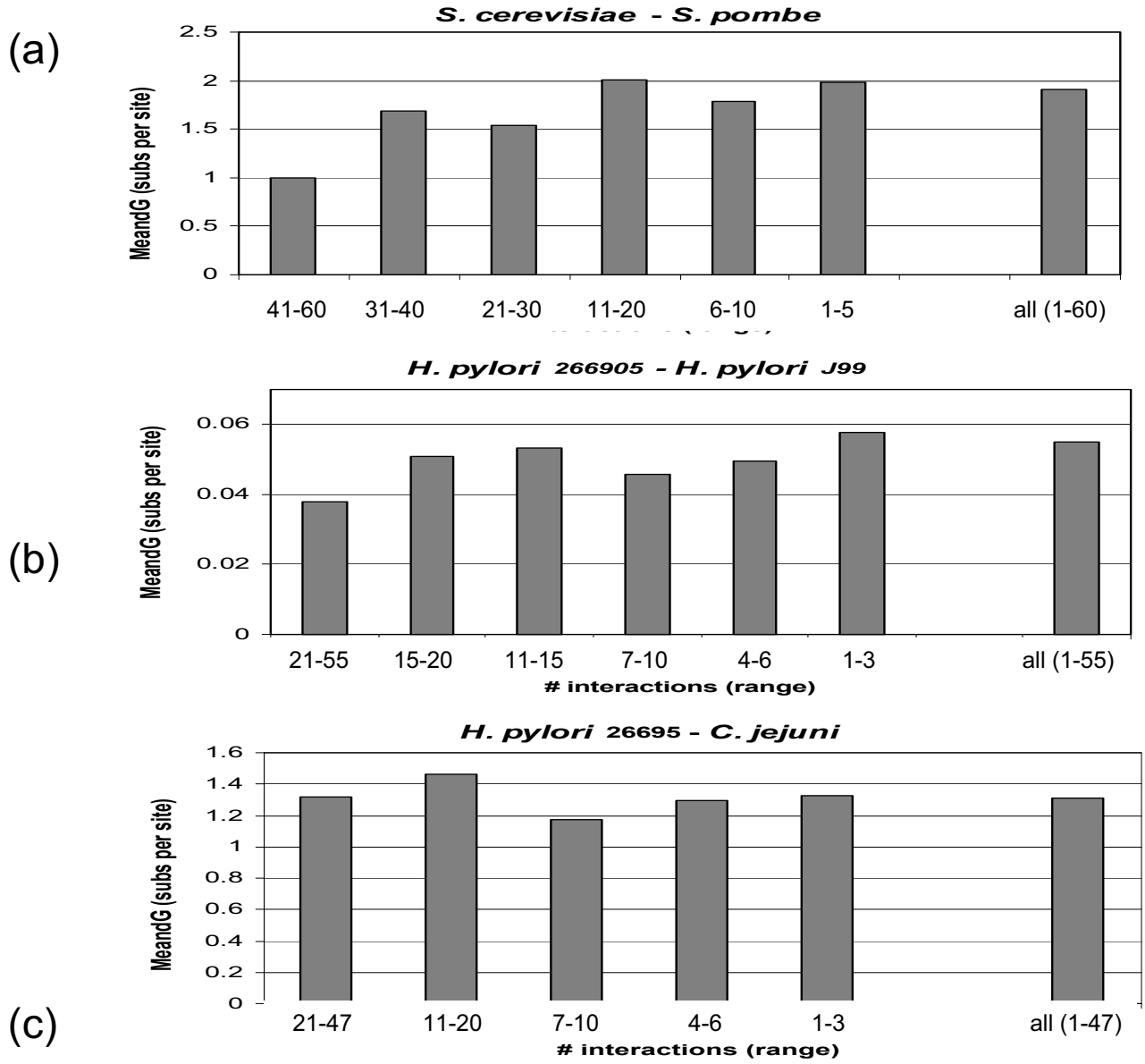

Figure 3

Mean evolutionary rates for bins of proteins with different number of interactions. Shown for each graph is the range of the number protein-protein interactions for each bin (x-axis) and the mean evolutionary rate (substitutions per site) for each bin (y-axis). (a) *S. cerevisiae* and *S. pombe* orthologs. (b) *H. pylori* strain 26695 and *H. pyori* strain J99 orthologs. (c) *H. pylori* strain 26695 and *C. jejuni* orthologs.

Yeast proteins with the greatest number of interactions appear to evolve slowly

The observations described above seem to indicate that the number of interaction partners a given protein has does not make an important contribution to the evolutionary rate. One could speculate, however, that whatever minor correlation is seen (Fig. 1a, 2a), is not spread evenly, as a miniscule difference in the evolutionary rates, among all proteins, but rather reflects a substantial slowdown of evolution among a small fraction of proteins that

Table 2: Statistical significance of the differences in evolutionary rates between groups of proteins with different numbers of interactions.

Bin (# interactions) comparisons[a]	P[b]
S. cerevisiae – S. pombe	
41 – 60 vs. 1 – 40	8.3×10^{-4}
31 – 60 vs. 1 – 30	2.4×10^{-2}
21 – 60 vs. 1 – 20	1.7×10^{-4}
H. pylori 26695 – H. pylori J99	
21 – 55 vs. 1 – 20	1.5×10^{-1}
15 – 55 vs. 1 – 14	1.8×10^{-1}
11 – 55 vs. 1 – 10	3.2×10^{-1}
H. pylori 26695 – C. jejuni	
21 – 47 vs. 1 – 20	9.8×10^{-1}
11 – 47 vs. 1 – 10	5.1×10^{-1}

[a] Orthologous pairs of proteins were placed into bins based on the number of protein-protein interactions (Figure 3). [b] *P*-value for the Student's ttest comparing the mean evolutionary rates between orthologs for bins with distinct ranges in the number of protein-protein interactions.

have the greatest number of interactions. To test this hypothesis, we grouped proteins from S. cerevisiae and H. pylori into separate bins, with each bin containing proteins whose number of interactions fell within a given range. Comparison of the evolutionary rates for proteins in different bins showed that yeast proteins in the bins with the greatest number of interactions, on average, evolved slower than the bulk of the proteins (Fig. 3a). The difference was less than twofold even for the top bin, but was statistically significant for each of the top three bins or their combination (Table 2). The proteins with a large number of interactions placed in the top bins comprise only 6.5% of the yeast proteins. In contrast, for the bulk of the proteins, which have a small to moderate number of interactions, there did not seem to be any dependence at all between the number of interactions and the evolutionary rates (Fig. 3a). H. pylori proteins with the greatest number of interactions also appear to have evolved slower on average between strains than the majority of the proteins. However, the difference was not significant and this effect was not seen in the comparison of H. pylori and C. jejuni orthologs (Table 2 and Fig 3b,3c).

Discussion and conclusions

The hypothesis that a protein's rate of evolution is determined by the fraction of residues that are critical to its function, and this, in turn, is likely to be proportional to the number of interactions a protein is involved in, seems to make perfectly good sense. Indeed, a recent report is consistent with this idea in suggesting that the number of protein-protein interactions significantly affects rates of evolution [4]. However, upon investigation of this relationship at multiple levels of evolutionary relatedness, we found that there was only a slight correlation, at best, between evolutionary rates and the number of protein-protein interactions. In fact, examination of the actual data presented in support of the previous claim of a connection between the number of interactions and evolutionary rates [4] also shows a weak correlation, albeit greater than the one observed in this study. Thus, differences in the number of interaction partners seem to explain, at best, only a small part of the great variation of the evolutionary rates of proteins encoded in each genome [11].

Why does the number of interaction partners apparently have only a slight effect on the evolutionary rate? The first and most obvious possibility to consider would be that the low quality of protein-protein interaction data might obscure the signal. Indeed, a recent comparison of protein-protein interaction data sets from high-throughput studies suggested that more than half of all interactions determined by large scale experiments are likely to be false positives [12]. However, at least for the yeast data, we relied on manually curated protein-protein interaction data from the MIPS database, which are expected to have a substantially lower error rate. Second, one could speculate that, even if the majority of the analyzed interactions actually do occur, they are selectively (nearly) neutral; the number of such real but functionally irrelevant interactions would not affect the rate of evolution. Third, the possibility exists that, even if many of the observed interactions are functionally important and, by inference, the respective binding sites are subject to purifying selection, the binding sites for different partners tend to overlap such that the number of amino residues in these sites increases only slowly with the increase in the numbers of interactions.

The latter two possibilities are not incompatible with each other and with the other aspect of the observations reported here. We found that the small fraction of yeast proteins that have the greatest number of interaction partners do,

on average, evolve slower than the bulk of the proteins, which are involved in a moderate or small number of interactions. This effect was less pronounced, if observed at all, for *H. pylori*, but it has to be noticed that the top bins of the *H. pylori* interaction data included proteins with fewer interactions than the respective bins in the yeast data (compare Fig. 3b,3c and 3a). Protein-protein interactions form scale-free networks, which show the characteristic power-law distribution of the node degrees; simply put, there is a small number of highly connected proteins (hubs), whereas the majority have a small number of partners (the most abundant class are proteins that are involved in just one interaction) [13,14]. Scale-free networks are highly tolerant to error (elimination of nodes at random) but are vulnerable to attack, i.e. elimination of the hubs [15] and, indeed, it has been found that the most highly connected proteins in yeast interaction networks tend to be essential [13]. This might explain the present findings, namely that a small number of yeast protein-protein interaction hubs evolve slowly due to strong purifying selection, whereas, for the great majority of the proteins, there is no discernible connection between the number of interactions and evolutionary rates.

Methods

Comparison of evolutionary rates and protein-protein interactions

Sets of protein sequences encoded by the complete genome sequences of the yeasts *S. cerevisiae* [16] and *S. pombe* [17], the nematode *C. elegans* [6] and the proteobacteria *H. pylori* strain 26695 [9], *H. pylori* strain J99 [8] and *C. jejuni* [10] were downloaded from the National Center of Biotechnology Information's Genbank ftp site ftp://ftp.ncbi.nlm.nih.gov/genomes/. Protein sets (proteomes) from the following pairs of complete genome sequences were compared in order to identify orthologous sequences: *S. cerevisiae* – *S. pombe*, *S. cerevisiae* – *C. elegans*, *H. pylori* strain 26695 – *H. pylori* strain J99, *H. pylori* strain 26695 – *C. jejuni*. Pairs of proteomes were compared using the BLASTP program [18], with post-processing of results done using the SEALS package [19]. For each proteome, individual proteins were used as queries in BLASTP searches against the entire proteome of the other analyzed species (or strain). Symmetrical best hits in these BLAST searches (expectation value $\leq 10^{-3}$) were taken to be orthologs [20]. Pairs of orthologous proteins were aligned using the ClustalW program [21] and their substitution (evolutionary) rates were calculated using the gamma distance correction [22]. The data on protein-protein interactions for the *S. cerevisiae* proteome were obtained from the Munich Information Center for Protein Sequences (MIPS) [5] Comprehensive Yeast Genome Database http://mips.gsf.de/proj/yeast/CYGD/db/index.html. This database includes a manually curated catalogue of binary protein-protein interactions that is considered to be a reliable reference set [12]. Protein-protein interactions for the *H. pylori* proteome [7] were taken from the PIMRider functional proteomics software platform http://pim.hybrigenics.fr/pimrider/pimriderlobby/PimRiderLobby.jps.

Authors' contributions

IKJ performed the comparisons between evolutionary rates and the number of protein-protein interactions and drafted the manuscript. YIW determined the evolutionary conservation levels for *S. cerevisiae* proteins and contributed to the statistical analysis. EVK helped to conceive of the study, participated in its design and coordination and revised the manuscript. All authors read and approved the final manuscript.

References

1. Hirsh AE and Fraser HB **Protein dispensability and rate of evolution.** *Nature* 2001, **411:**1046-1049
2. Jordan IK, Rogozin IB, Wolf YI and Koonin EV **Essential genes are more evolutionarily conserved than are nonessential genes in bacteria.** *Genome Res* 2002, **12:**962-968
3. Brookfield JF **What determines the rate of sequence evolution?** *Curr Biol* 2000, **10:**R410-R411
4. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C and Feldman MW **Evolutionary rate in the protein interaction network.** *Science* 2002, **296:**750-752
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. The C. elegans Sequencing Consortium **Genome sequence of the nematode C. elegans: a platform for investigating biology.** *Science* 1998, **282:**2012-2018
7. 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
8. Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC and deJonge BL **Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori.** *Nature* 1999, **397:**176-180
9. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S and Dougherty BA **The complete genome sequence of the gastric pathogen *Helicobacter pylori*.** *Nature* 1997, **388:**539-547
10. Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T and Holroyd S **The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences.** *Nature* 2000, **403:**665-668
11. Grishin NV, Wolf YI and Koonin EV **From complete genomes to measures of substitution rate variability within and between proteins.** *Genome Res* 2000, **10:**991-1000
12. 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
13. Jeong H, Mason SP, Barabasi AL and Oltvai ZN **Lethality and centrality in protein networks.** *Nature* 2001, **411:**41-42
14. Lappe M, Park J, Niggemann O and Holm L **Generating protein interaction maps from incomplete data: application to fold assignment.** *Bioinformatics* 2001, **17(Suppl 1):**S149-156
15. Albert R, Jeong H and Barabasi AL **Error and attack tolerance of complex networks.** *Nature* 2000, **406:**378-382
16. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C and Johnston M **Life with 6000 genes.** *Science* 1996, **274:**563-547
17. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J and Baker S **The genome sequence of Schizosaccharomyces pombe.** *Nature* 2002, **415:**871-880
18. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ **Gapped BLAST and PSI-BLAST: a new generation**

of protein database search programs. *Nucleic Acids Res* 1997, **25:**3389-3402
19. Tatusov RL, Koonin EV and Lipman DJ **A genomic perspective on protein families.** *Science* 1997, **278:**631-637
20. Walker DR and Koonin EV **SEALS: a system for easy analysis of lots of sequences.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**333-339
21. Higgins DG, Thompson JD and Gibson TJ **Using CLUSTAL for multiple sequence alignments.** *Methods Enzymol* 1996, **266:**383-402
22. Ota T and Nei M **Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites.** *J Mol Evol* 1994, **38:**642-643