J Mol Evol (2008) 66:395–404
DOI 10.1007/s00239-008-9094-4

Contact Density Affects Protein Evolutionary Rate from Bacteria to Animals

Tong Zhou · D. Allan Drummond · Claus O. Wilke

Received: 9 November 2007 / Accepted: 25 February 2008 / Published online: 1 April 2008

Abstract The density of contacts or the fraction of buried sites in a protein structure is thought to be related to a protein's designability, and genes encoding more designable proteins should evolve faster than other genes. Several recent studies have tested this hypothesis but have found conflicting results. Here, we investigate how a gene's evolutionary rate is affected by its protein's contact density, considering the four species *Escherichia coli*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Homo sapiens*. We find for all four species that contact density correlates positively with evolutionary rate, and that these correlations do not seem to be confounded by gene expression level. The strength of this signal, however, varies widely among species. We also study the effect of contact density on domain evolution in multidomain proteins and find that a domain's contact density influences the domain's evolutionary rate. Within the same protein, a domain with higher contact density tends to evolve faster than a domain with lower contact density. Our study provides evidence that contact density can increase evolutionary rates, and that it acts similarly on the level of entire proteins and of individual protein domains.

Keywords Designability · Protein structure · Evolutionary rate · Protein evolution · Domain · Principal component regression

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9094-4) contains supplementary material, which is available to authorized users.

T. Zhou · C. O. Wilke (✉)
Center for Computational Biology and Bioinformatics,
Section of Integrative Biology, University of Texas at Austin,
Austin, TX 78731, USA
e-mail: cwilke@mail.utexas.edu

D. A. Drummond
FAS Center for Systems Biology, Harvard University,
Cambridge, MA 02138, USA

C. O. Wilke
Institute for Cell and Molecular Biology, University of Texas
at Austin, Austin, TX 78731, USA

Introduction

Understanding why protein-coding genes evolve at different rates is central for many fields, including molecular evolution, comparative genomics, and structural biology (Pal et al. 2006). In the past few years, numerous quantities have been found to correlate with evolutionary rate, including protein length, the number of a protein's interaction partners, the centrality in the protein interaction network, the dispensability of the gene, and gene expression level (Hurst and Smith 1999; Pal et al. 2001, 2003; Hirsh and Fraser 2001; Marais and Duret 2001; Jordan et al. 2002; Fraser et al. 2002; Rocha and Danchin 2004; Hahn and Kern 2005; Wall et al. 2005; Zhang and He 2005; Lemos et al. 2005; Agrafioti et al. 2005; Drummond et al 2005, 2006; Kim et al. 2006). Among these various variables, expression level seems to be the major determinant of evolutionary rate, at least in fast-replicating single-cellular organisms (Drummond et al. 2006).

A number of recent studies have addressed whether protein structure *per se* influences evolutionary rate (Shakhnovich 2006; Bloom et al. 2006; Lin et al. 2007), that is, whether certain aspects of a protein's structure such as its secondary structure composition and its accessible surface area affect the rate at which the protein's genetic sequence evolves. The main emphasis in these studies is the relationship between a gene's evolutionary rate and the

corresponding protein's contact density (mean number of residue-residue contacts per residue) or fraction of buried sites. The latter two quantities are predicted (Wolynes 1996; Shakhnovich 1998; England and Shakhnovich, 2003) to be measures of a protein's designability (Li et al. 1996), and more designable proteins should have more rapidly evolving genetic sequences (Bloom et al. 2006). Bloom et al. (2006) found indeed that contact density and fraction of buried sites were positively correlated with evolutionary rate in yeast. By contrast, Shakhnovich (2006) found a negative correlation between contact density and evolutionary rate in yeast and *Caenorhabditis elegans*. Finally, studying yeast, Lin et al. (2007) found no correlation between evolutionary rate and the fraction of buried sites when this fraction was calculated from protein crystal structures, but found a negative correlation when the fraction was predicted from the protein's amino-acid sequence using a support-vector machine. These discrepancies may be partly caused by differences in the definitions of the basic quantities studied. For example, Shakhnovich (2006) considered protein domains, whereas Bloom et al. (2006) considered entire proteins. Similarly, Lin et al. (2007) included interchain contacts in their calculation of relative accessible surface area (and thus fraction of buried sites), whereas Bloom et al. (2006) excluded those contacts. Some of the results may also be artifacts caused by insufficient data-set size. In particular, the study by Bloom et al. (2006) analyzed only 194 open reading frames (ORFs).

The purpose of the present study is fourfold. First, we verify the results of Bloom et al. (2006) in an independently derived and extended yeast data set. Second, we extend this analysis to three other species, *Escherichia coli*, *Drosophila melanogaster*, and *Homo sapiens*. Third, the theory by England and Shakhnovich (2003) suggests several quantities other than contact density as measures of a protein's designability, and we test whether these quantities provide additional information about evolutionary rate. Fourth, we assess whether the individual domains in multidomain proteins show rates of evolution related to their contact densities. Throughout this paper, we use the generic term *evolutionary rate* to refer to the rate of nonsynonymous substitutions per nonsynonymous site, *dN*.

Materials and Methods

Structural and Genomic Data

To match protein structures with the protein sequences in each analyzed species, we downloaded the structure prediction data from the GTOP (Genomes TO Protein structures and functions) database (Kawabata et al. 2002). For a given match in the GTOP database, if the region of similarity was longer than 80% of the protein length and the sequence identity was larger than 40% of the sequence in the Protein Data Bank (PDB), then the match was saved for further calculation. This process yielded 777, 363, 795, and 860 matches in *E. coli*, *S. cerevisiae*, *D. melanogaster*, and *H. sapiens*, respectively. We only considered entries in the PDB corresponding to experimentally determined structures.

For each protein with a match, we obtained the corresponding three-dimensional (3D) structural information from the PDB. From these crystal structures, we calculated residue-residue contact maps, secondary structure elements, and percentage solvent accessibility. We considered two residues in contact if any nonhydrogen atom of one residue was within a distance of 4.5 Å from any nonhydrogen atom of the other residue (Bloom et al. 2006). We excluded contacts between immediate neighbors in the polypeptide. We calculated the contact density for each protein by dividing the total number of contacts by the protein length. We calculated the secondary structure for each aligned residue using the DSSP (Dictionary of Protein Secondary Structure) program (Kabsch and Sander 1983). We simplified our data set by keeping track of four types of secondary structure elements only: helix (DSSP class H), sheet (DSSP class E), turn (DSSP classes S and T), and coil (DSSP classes B, G, I, and "."). We also calculated the percentage solvent-accessible surface area for each aligned residue with the DSSP program (Kabsch and Sander 1983). We normalized these results by the reference surface areas of an extended Gly-X-Gly peptide (Creighton 1992). We considered residues with $<25\%$ relative solvent accessibility as buried. A protein's fraction of buried sites is the number of buried sites divided by the protein length. We calculated relative solvent accessibility considering only the atoms within one protein chain, in agreement with Bloom et al. (2006), but in contrast to the procedure followed by Lin et al. (2007).

We calculated the evolutionary rates *dN* (nonsynonymous substitutions per nonsynonymous site) and *dS* (synonymous substitutions per synonymous site) between pairs of orthologues. For *E. coli*, we obtained orthologues between *E. coli* and *Salmonella typhimurium* from TIGR's Comprehensive Microbial Resource's multigenome homology comparison tool (http://www.cmr.tigr.org/). For yeast, we obtained orthologues between *S. cerevisiae* and *Saccharomyces bayanus* from the file ''fungalAlignCorrespondance.txt'' at the Saccharomyces Genome Database (ftp://genome-ftp.stanford.edu/). For fly, we obtained orthologues between *D. melanogaster* and *Drosophila yakuba* from the Drosophila 12-genome project AAAWiki at http://www.rana.lbl.gov/drosophila/. For human, we obtained orthologues between *H. sapiens* and *Mus musculus* from Biomart through the Ensembl Homology track (http://www.ensembl.org/). For each pair of orthologues, we obtained aligned nucleotide sequences based on the

alignment of the peptide sequences, which we generated with MUSCLE (Edgar 2004). Finally, we calculated *dN* and *dS* with PAML (Yang 1997), using one ω value (PAML parameters NSsites = 0 and fix_omega = 0) and the F3X4 codon frequency model (PAML parameter CodonFreq = 2). We obtained genomic sequences from the following sources: the Comprehensive Microbial Resource (http://www.cmr.tigr.org/) for *E. coli* and *S. typhimurium*, the *Saccharomyces* Genome Database (ftp://genome-ftp.stanford.edu/) for yeast, the Eisen Lab (http://www.rana.lbl.gov/drosophila/) for fly, and Ensembl (http://www.ensembl.org/) for human and mouse.

We used previously published expression data for each species: For *E. coli*, we obtained gene expression levels measured in mRNAs per cell from Covert et al. (2004); for yeast, we used expression data from Holstege et al. (1998); for fly, we used as expression level the geometric mean of expression data from different tissues obtained by Stolc et al. (2004); for human, we also measured expression level as the geometric mean of expression data from different tissues; we downloaded the human expression data from http://www.wombat.gnf.org/ (Su et al. 2004). After discarding all ORFs for which we did not have expression data and *dN* values, our final data sets contained 676, 339, 114, and 417 ORFs for *E. coli*, *S. cerevisiae*, *D. melanogaster*, and *H. sapiens*, respectively.

We obtained protein domain information from the Dali Domain Dictionary (http://www.ebi.ac.uk/dali/domain) (Dietmann et al. 2001). We calculated the contact density for each domain, excluding any contacts to residues outside of the domain. To calculate evolutionary rates for domains, we first had to identify domain boundaries in the aligned orthologues. We determined the location of these boundaries by aligning the two orthologues plus the sequence of the corresponding PDB structure with MUSCLE. We excluded from our data set those orthologous domain pairs with an alignment length <80%. Then we calculated *dN* and *dS* for the domains using PAML as described above.

Designability and Contact Density

A protein's designability is defined as the total number of amino acid sequences that fold into the given structure (Li et al. 1996; Kussell 2005). Designability varies widely among structures. Although atomistic simulations could be used to estimate the designabilities of real proteins, they are extremely time-consuming and therefore become infeasible for large numbers of proteomic sequences (but see Meyerguz et al. 2007). Assuming that the energy of a structure is due to pairwise interactions between residues, Shakhnovich and coworkers (England and Shakhnovich 2003; England et al. 2003) have proposed that a structure's designability D is related to traces of even powers of the structure's contact matrix **C**, that is, $\mathrm{Tr}C^2/L$, $\mathrm{Tr}C^4/L$, $\mathrm{Tr}C^6/L$, and so on, where L is the protein's length. (Note that $\mathrm{Tr}C^2/L$ is simply the average number of contacts per residue.) A suitable linear combination of these contact traces should estimate D. Alternatively, the largest eigenvalue of C should also estimate D (England and Shakhnovich 2003). Finally, Bloom et al. (2006) suggested that an alternative to measures derived from the contact matrix could be the fraction of buried residues in the protein structure, f_{bur}.

Here, we estimate D from the first three even contact traces, the largest eigenvalue of the contact matrix, and the fraction of buried sites. All five of these quantities are related to the density of contacts in a protein structure and are strongly correlated with each other (Table 1). Therefore, we use the term *contact density* as a generic term referring to all of them. When we want to refer to the quantity $\mathrm{Tr}C^2/L$, which is usually called contact density, we here use the term *average number of contacts per residue*.

Statistical Analysis

We carried out our statistical analyses with the statistics software R (Ihaka and Gentleman 1996). We calculated correlation coefficients and associated *P* values using the R function "cor.test()" with method "spearman." We used the package "pls" in R to perform principal component regression. Analyses were carried out on rank-transformed data unless specified otherwise.

In our correlation analyses, we adjusted for multiple testing separately for each species, using the method of the false discovery rate (Benjamini and Hochberg 1995).

Results

Protein Structure and Evolutionary Rate

First, we correlated with evolutionary rate the five measures of contact density: the average number of contacts

Table 1 Spearman correlations between the average number of contacts per residue (*den*) and the other four measures of contact density

Organism	*ev*	Tr4	Tr6	f_{bur}
E. coli	0.90***	0.98***	0.96***	0.90***
S. cerevisiae	0.86***	0.98***	0.94***	0.87***
D. melanogaster	0.90***	0.98***	0.94***	0.90***
H. sapiens	0.86***	0.98***	0.94***	0.83***

Note. ev, maximum eigenvalue of the contact matrix; Tr4, fourth-order contact trace; Tr6, sixth-order contact trace; f_{bur}, fraction of buried sites. *** $P < 0.001$. Adjustment for multiple tests does not affect significance levels

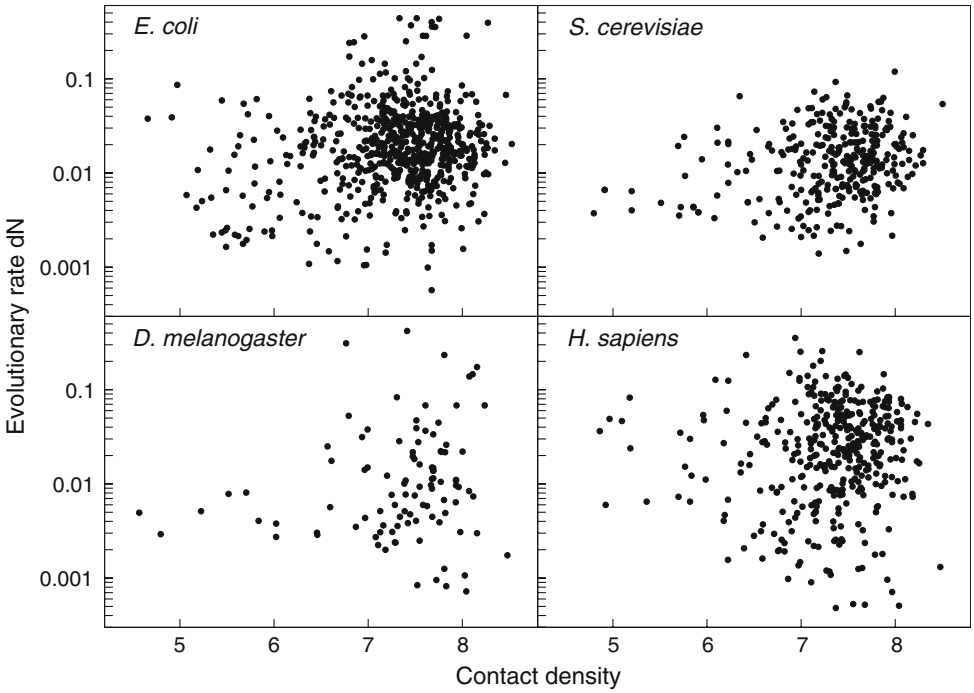

Fig. 1 Evolutionary rate *dN* as a function of the average number of contacts per residue (contact density)

Table 2 Spearman correlations of *dN* with expression level, measures of contact density, and measures of secondary structure content

Organism	*exp*	*den*	*ev*	Tr4	Tr6	f_{bur}	*len*	f_h	f_e	f_t	f_c
E. coli	−0.50***	0.17***	0.16***	0.16***	0.16***	0.20***	0.16***	0.02	0.03	−0.04	0.01
S. cerevisiae	−0.51***	0.25***	0.21***	0.25***	0.25***	0.23***	0.16***	0.05	0.01	−0.12(*)	0.05
D. melanogaster	−0.28**	0.32**(*)	0.27**	0.29**	0.28**	0.36***	0.34***	0.07	−0.08	0.08	−0.02
H. sapiens	−0.19***	0.14**	0.15**	0.15**	0.15**	0.16**(*)	0.05	−0.07	0.02	0.05	0.09

Note. exp, gene expression level; *den*, contact density; *ev*, maximum eigenvalue of the contact matrix; Tr4, fourth-order contact trace; Tr6, sixth-order contact trace; f_{bur}, fraction of buried sites; *len*, protein length; f_H, f_E, f_T, and f_C, fraction of sites with secondary structure helix, sheet, turn, and coil, respectively. Sample sizes are $n = 777$ (*E. coli*), $n = 363$ (*S. cerevisiae*), $n = 795$ (*D. melanogaster*), and $n = 860$ (*H. sapiens*). Significance levels: *** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$. Significance levels in parentheses disappear after correction for multiple testing

TrC^2/L (*den*), the maximum eigenvalue of the contact matrix (*ev*), the quantities TrC^4/L (Tr4) and TrC^6/L (Tr6), and the fraction of buried sites (f_{bur}). We found a significant positive correlation between evolutionary rate *dN* and these quantities in all four species (Fig. 1 and Table 2). The correlation coefficients ranged from 0.14 to 0.36, in agreement with the results of Bloom et al. (2006).

The correlations between contact density and *dN* are only meaningful if they are not confounded by gene expression level, which is a major predictor of evolutionary rate in most species (Duret and Mouchiroud 2000, Pal et al. 2001; Herbeck et al. 2003; Rocha and Danchin 2004; Subramanian and Kumar 2004; Drummond et al. 2005, 2006). In our data set, expression level correlates negatively with *dN* in all four species (Table 2). To detect potentially confounding effects of expression level, we correlated it with the five measures of contact density. Our results were ambiguous (Table 3). While we found weak but highly significant negative correlations with expression level for all measures of contact density in *E. coli* and yeast, correlations were not significant in fly (likely a consequence of small sample size) and absent in human. A partial correlation analysis between *dN* and measures of contact density controlling for expression level yielded results similar to the raw correlations shown in Table 2 (data not shown). This result provides further support for our hypothesis but is not conclusive, because partial correlation analysis has a tendency to produce false positives when the variable that is being controlled for is noisy (Drummond et al. 2006). We return to the question whether expression level confounds our results in the next subsection.

We also considered protein length. Bloom et al. (2006) found that contact density is related to protein length, that only short proteins tend to have low contact densities, and that protein length correlates positively with evolutionary

Table 3 Spearman correlations of expression level with measures of contact density

Organism	*den*	*Ev*	Tr4	Tr6	f_{bur}
E. coli	−0.13**(*)	−0.15***	−0.15***	−0.15***	−0.16***
S. cerevisiae	−0.19**(*)	−0.19**(*)	−0.18**(*)	−0.18**(*)	−0.17**(*)
D. melanogaster	−0.22*	−0.13	−0.19(*)	−0.17	−0.20(*)
H. sapiens	−0.01	−0.00	−0.03	−0.03	−0.02

Note. den, contact density; *ev*, maximum eigenvalue of the contact matrix; Tr4, fourth-order contact trace; Tr6, sixth-order contact trace; f_{bur}, fraction of buried sites. Significance levels: *** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$. Significance levels in parentheses disappear after correction for multiple testing

rate. Our present results agree with these findings (Tables 2 and Supplementary Table S1). Finally, we tested whether proteins' secondary structure composition had a relationship to evolutionary rate. We found no such signal (Table 2), in agreement with the results of Bloom et al. (2006).

Principal Component Regression

The simple correlation and partial correlation analyses of the previous subsection could not rule out that the relationship between contact density and evolutionary rate is confounded by expression level. As an alternative approach to this question, we carried out a principal component regression (Mandel 1982; Drummond et al. 2005) on our data.

In this regression, we used *dN* as response variable and included gene expression level, length, and the five measures of contact density as predictors. Because secondary structure composition was not correlated with *dN*, we excluded the former from the predictors to simplify the analysis. However, our conclusions remain virtually unchanged if secondary structure composition is included (data not shown).

We first carried out a principal component analysis of the seven predictor variables. We found that the component composition was nearly identical in the four species (Fig. 2). The first three components have a clear

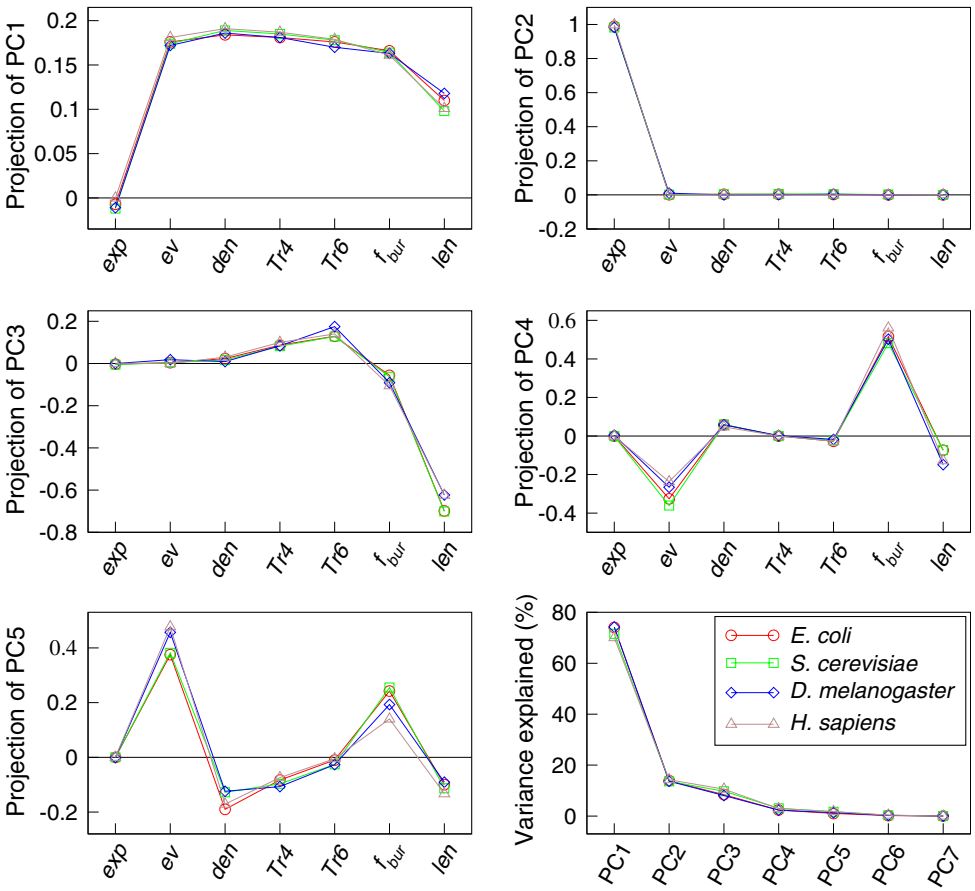

Fig. 2 Projection patterns of the first five principal components for each species. We analyzed seven predictor variables including expression level, protein length, and five measures of contact density. *exp*, gene expression level; *den*, contact density; *ev*, maximum eigenvalue of the contact matrix; Tr4, fourth-order contact trace; Tr6, sixth-order contact trace; f_{bur}, fraction of buried sites; *len*, protein length

interpretation. PC 1 represents contact density, PC 2 represents expression level, and PC 3 represents length.

For each species, we next regressed evolutionary rate *dN* against all components. In all four species, we found that both the contact density (PC 1) and the expression level (PC 2) made a significant contribution to the regression (Fig. 3). (For fly, the significance of *dN*'s regression against PC 2 was marginal.) Contact density explained between 2.3% and 11.8% of the variance in *dN*, independently of expression level. These findings are in agreement with the results of Bloom et al. (2006). We also regressed the PCs against the synonymous evolutionary rate *dS*, expecting that contact density would have no explanatory power for *dS*. Yet we found a significant correlation between PC 1 and *dS* in *E. coli* and fly. In *E. coli*, contact density predicts *dS* as well as it predicts *dN*. In fly, contact density predicts *dS* better than *dN*, explaining almost 20% of the variance in *dS*.

For fly, the PC regression seems to imply that contact density is a more important predictor of evolutionary rate than expression level. We caution against this conclusion, however, because sample size in fly is small and fly genes with structural information form a biased sample of all fly genes. Expression level explains 12.3% of the variance in *dN* if we consider all genes for which we have both expression level and *dN* values ($\rho = -0.35$, $P \ll 0.0001$; $n = 2{,}277$) but explains only 7.8% of the variance if we restrict the data set to genes with structural information ($\rho = -0.28$, $P = 0.003$; $n = 114$). The contrast becomes even stronger if we consider the fraction of optimal codons (F_{op}, known to correlate strongly with expression level in fly, e.g. Duret and Mouchiroud 1999) as surrogate for expression level. F_{op} explains 27% of *dN*'s variance in the full data set ($\rho = -.52$, $P \ll 0.0001$, $n = 2{,}277$) but only 11.6% in the data set with structural information ($\rho = -0.34$, $P = 0.0002$; $n = 114$). Both expression level and F_{op} are significantly higher for the genes with structural data than for those without (*t*-tests, $P = 0.0006$ and $P \ll 10^{-10}$, respectively).

For *E. coli*, *S. cerevisiae*, and *D. melanogaster*, we also carried out the principal component regression with F_{op} added as second indicator of gene expression level. Our results were essentially unchanged from those shown in Fig. 3.

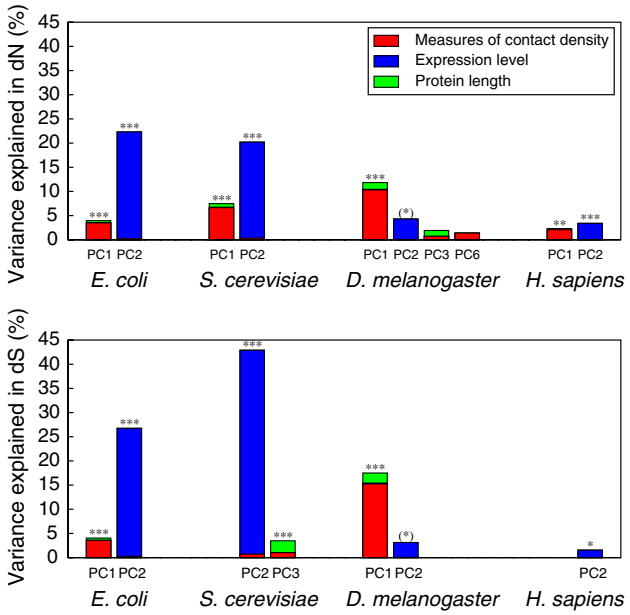

Fig. 3 Variance in *dN* and *dS* explained by principal components. We only show components that explain at least 1% of the variance in *dN* or *dS*. The following components were statistically significant ($P < 0.05$) but excluded from the graph by this criterion: PC5 in human regressed against *dN*, PC4 in *E. coli*, and PC1 in yeast regressed against *dS*. The principal component structure is shown in Fig. 2. Significance levels: ***$P < 0.001$; **$P < 0.01$; *$P < 0.05$. Significance levels in parentheses disappear after correction for multiple testing (Benjamini and Hochberg 1995)

Contact Density and Evolutionary Rate of Protein Domains

Even though the PC regression seems to indicate that contact density influences evolutionary rate independently of expression level, it is worrisome that contact density predicts the synonymous rate of evolution *dS* in *E. coli* as well as, and in fly better than, the nonsynonymous rate *dN*. This result could indicate that contact density is confounded with the true, evolutionarily relevant expression level, which may differ from the expression level obtained experimentally under laboratory growth conditions. Alternatively, factors other than expression level may also be confounding contact density.

To control for a large class of possibly confounding factors, we shifted our analysis to protein domains. Different domains have different designabilities, and if a protein consists of a high-designability and a low-designability domain, we expect the former to evolve faster than the latter. Since the two domains are physically linked, we expect them to experience virtually identical mutation pressure, expression level, selection for translational efficiency, or functional importance. Even though two domains within a single protein may still experience different selective pressures (see *Discussion*), by analyzing domains we exclude all confounding factors that cause differences among proteins but act uniformly within a single protein.

We calculated contact densities and evolutionary rates for each protein domain, and found that the correlations between these quantities mirrored those for protein-wide quantities but tended to be weaker (Table 4). (In this paragraph, we use *contact density* exclusively in its narrow sense, to refer to the average number of contacts per residue.) We then correlated the log-transformed ratio of the per-domain and per-protein evolutionary rates with the difference of the per-domain and

Table 4 Correlations between a domain's *dN* and a domain's contact density

Organism	Domain		Protein		Pearson correlation R [$\ln(dN_{dom}/dN_{pro})$, $(den_{dom} - den_{pro})$]
	n	ρ (dN_{dom}, den_{dom})	n	ρ (dN_{pro}, den_{pro})	
E. coli	833	0.08*	521	0.13**	0.14***
S. cerevisiae	390	0.18***	229	0.23***	0.15**
D. melanogaster	633	0.24***	387	0.29***	0.22***
H. sapiens	752	0.03	456	0.09*	0.09*

Note. Correlations were calculated for all ORFs with protein structure and domain data, including ORFs without expression data. n, number of sequences; dN_{dom} and dN_{pro}, the domain's and the corresponding protein's evolutionary rates; den_{dom} and den_{pro}, the domain's and the corresponding protein's contact densities. Significance levels: *** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$

per-protein contact densities, and found significant positive correlations in all species (Table 4). This result indicates that a domain's contact density can speed up the domain's evolutionary rate in comparison to the protein-wide *dN*. As a second test of the same principle, we considered only two-domain proteins, identified the domain with the higher contact density (den_{high}) and the one with the lower contact density (den_{low}), and correlated the log-transformed ratio of the corresponding evolutionary rates $\ln(dN_{high}/dN_{low})$ with the difference $den_{high}-\backslash den_{low}$. Again, we found a significant positive Pearson correlation in all species (Table 4; see also Fig. 4). We also considered the data set obtained by pooling the data from all species, and found a positive Pearson correlation as well ($r = 0.20$, $P < 10^{-4}$). By contrast, we found no correlation when we carried out the same analysis for *dS* ($r = -0.04$, $P = 0.460$). Our results did not change significantly when we extended the analysis from two-domain to all multidomain proteins (data not shown).

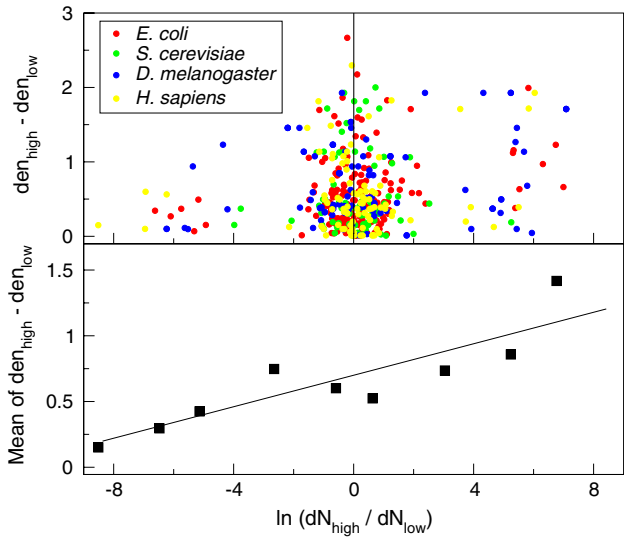

Fig. 4 The correlation between the domain's contact density difference and the corresponding log-transformed *dN* ratio in two-domain proteins. In the lower graph, each data point represents a bin in the logarithmic *dN* ratio with a step size of 2.0

Accuracy of Crystal Structures and Sequence Alignments

To assess to what extent our results depend on the sequence identity between ORF and PDB structure and on the accuracy of the PDB structure, we repeated our analysis with reduced, high-confidence data sets in which we excluded all cases with a sequence identity <80% or with a structure that had a resolution >2.5 Å or was not determined by X-ray diffraction. These criteria led to a dramatic reduction in data-set size, from a factor of 2.4 in *E. coli* to a factor of 160 in *D. melanogaster*. Consequently, statistical significance was greatly reduced or lost completely for many correlations (Supplementary Tables S2 and S3). Yet the sign and strength of the correlations were largely unaffected, and all whole-gene correlations remained significant in yeast (Table S2). For fly, the reduced data set was so small (five ORFs with both structural and expression data) that a correlation analysis on it became meaningless. For the per-domain results—domains with contact density exceeding the protein average experience accelerated evolution compared to the entire gene—significance was preserved in *E. coli* and *S. cerevisiae*.

We found that the size of the reduced data sets follows broadly the number of species-specific structures in the PDB. As of November 8, 2007, there are 411 structures for *E. coli*, 180 for *S. cerevisiae*, and 895 for *Homo sapiens*, but only 10 for *D. melanogaster*. Note that we should not expect perfect agreement between these numbers and the sizes of our data sets, because the PDB frequently contains several structures for the same protein, for example, with different ligands or in different mutant forms.

Discussion

We have analyzed whether a protein structure's contact density affects the evolutionary rate of the encoding gene. We used the average number of contacts per residue, the maximum eigenvalue of the contact matrix, the fourth- and

sixth-order contact trace, and the fraction of buried sites to estimate contact density. These five quantities were strongly correlated with each other and had roughly comparable predictive power for evolutionary rate. We found that proteins with higher contact density tended to correspond to more rapidly evolving genes. This tendency is consistent with the hypothesis that contact density is a measure of designability and that more designable proteins evolve faster. We tested whether these results were possibly caused by confounding effects of gene expression level but did not find strong evidence supporting this alternative hypothesis.

The effect of contact density on evolutionary rate differed substantially among the four species we studied. Contact density was a much better predictor of *dN* in yeast and fly than in *E. coli* and human. In particular, in fly contact density explained more than 10% of the variation in *dN*, whereas in human it only accounted for approximately 2%. Our results for yeast were largely consistent with the previous results of Bloom et al. (2006), but the present study used a data set almost twice as large as the previous one (339 yeast ORFs in the present study versus 194 ORFs in Bloom et al. [2006]). Furthermore, with respect to secondary structure composition and protein length, our results in all four species also agreed with those of Bloom et al. (2006). Secondary structure composition seems to be largely irrelevant for the rate of evolution. Protein length is strongly correlated with contact density but does not seem to make an independent contribution to *dN*.

Solvent-inaccessible residues tend to be evolutionarily more conserved than solvent-accessible residues (Koshi and Goldstein 1995; Goldman et al. 1998; Mirny and Shakhnovich 1999; Dean et al. 2002), yet we find that proteins with a larger fraction of buried sites evolve more rapidly as a whole. The resolution to this apparent contradiction is simple: just because buried sites in one protein evolve slower than exposed sites in the same protein, these buried sites do not necessarily evolve slower than buried or exposed sites in another protein. Regions of high contact density form stabilizing cores of conserved highly interacting amino acids, and these cores allow other, more exposed residues to mutate more freely (Shakhnovich et al. 2005; Bloom et al. 2006). As a result, the protein-wide evolutionary rate increases with the fraction of buried sites in the protein.

In *E. coli*, yeast, and fly, we found a weak tendency for proteins with higher contact density to have lower expression levels. This tendency may simply reflect the negative correlation between expression level and gene length (Supplementary Table S1). Since more highly expressed genes tend to have lower *dN* values, some of the positive correlation between contact density and *dN* may actually reflect the negative correlation between contact density (or length) and expression level. Nevertheless, because of our results from the PC regression and from protein domain evolution (see also below), we are confident that at least some of the correlation between contact density and evolutionary rate is genuine.

In *E. coli* and fly, we also found a significant correlation between contact density and the rate of synonymous substitution *dS*. While previous work has shown that protein structure can cause codon bias (Orešič and Shalloway 1998; Gu et al. 2004), these effects are generally weak and restricted to a small number of sites in a protein and, therefore, are likely not causing the correlations we observed. We currently do not have a good explanation for our finding, or why it should occur in *E. coli* and fly but not in yeast or human.

Most proteins whose length exceeds approximately 300 residues contain two or more structural domains. The individual domains in such proteins usually show a high degree of structural independence, with relatively weak interactions between the domains. We tested whether differences in contact density were correlated with differences in evolutionary rate for domains within a single protein, and found that such a correlation, even though weak, does indeed exist. This result serves as an important control for factors that might have confounded our results for entire proteins. Numerous quantities may confound a comparison of evolutionary rates among genes, such as expression level, GC bias, gene function, or positive selection, and it is difficult to properly control for all of them in a comparison among genes. By contrast, we expect that many of these factors, in particular, expression level and GC bias, are identical for different domains of the same protein. Even so, there are potentially confounding factors acting on individual domains. Different domains in a protein frequently have separate functions (Ren et al. 1995; Holstein et al. 1996; Appelgren et al. 2003), and certain types of domains could experience increased positive or purifying selection pressures. Domains with a larger proportion of sites involved in protein-protein interactions may experience stronger purifying selection (Kim et al. 2006). Purifying selection is also expected to act stronger on thermodynamically less stable domains (Bloom et al. 2005).

For any study involving protein structure, the question arises how reliable the data are. The main results of the present work were obtained under criteria that maximize data-set size over data quality. We considered all ORFs for which we could find a structure with at least 40% sequence identity and considered all experimentally determined protein structures, regardless of resolution or experimental method. Under these criteria, we can assume that we generally assign the correct protein fold to each ORF, but the atomic details of the PDB structure will usually differ from those of the actual protein we are interested in. To

assess whether the quality of the alignment or structure affected our results, we also considered high-confidence data sets in which we excluded distantly related structures and structures with low resolution. Our results for these data sets were consistent across all species and generally followed the pattern we would expect from a reduction in data-set size. Correlations did not change sign and had only moderate changes in magnitude, but statistical significance declined or disappeared completely.

Two recent studies seem to conflict with our results. Shakhnovich (2006) analyzed domains of protein-coding genes in yeast and *Caenorhabditis elegans*, and found a negative correlation between *dN*/*dS* and contact density. While we analyzed *dN* rather than *dN*/*dS* here, our results remain largely unchanged if we consider the latter (data not shown). The differences between Shakhnovich's work and ours seems to arise from the assignment of sequences to structures. While his and our evolutionary rates correlate strongly, we found little overlap between our sequence-to-structure assignment and his. In support of our results, we emphasize that the sequence-to-structure assignment of Bloom et al. (2006) was carried out independently from the one we use here, with largely identical results.

Lin et al. (2007) reported a positive correlation between evolutionary rate and the fraction of exposed sites, thus implicitly reporting a negative correlation with the fraction of buried sites. There are two important differences between Lin et al.'s work and ours. First, Lin et al. (2007) counted both intra- and interchain contacts when calculating exposed surface area, whereas we excluded all interchain contacts. We believe that mixing these two types of contacts is problematic, because intrachain contacts are predicted to speed up evolution (England and Shakhnovich 2003; Bloom et al. 2006), whereas interchain contacts seem more likely to slow down evolutionary rate, even if the magnitude of this effect is highly debated (Fraser et al. 2002; Bloom and Adami 2003; Fraser 2005; Mintseris and Weng 2005; Kim et al. 2006; Drummond et al. 2006; Hakes et al. 2007). Second, whereas our work used experimentally determined protein structures, the central results of Lin et al. (2007) relied on machine learning algorithms to predict the fraction exposed in sequences that were only distantly related to proteins with known structures. Lin et al. (2007) obtained their main result from amino acid sequences alone, using a support-vector machine (SVM) to predict residues' surface exposure. Lin et al.'s Table 1 shows that a strong positive correlation between the fraction of exposed sites (as predicted by SVM) and the evolutionary rate arose only in the limit where the SVM was least likely to work, for ORFs that align poorly with sequences of known crystal structures.

In this study, we have addressed the question whether more designable proteins evolve faster. We have found that such a relationship seems to exist but is weak. Two major obstacles impede more conclusive studies. First, the number of proteins with known crystal structure remains comparatively small. Even for well-studied model organisms such as *E. coli* and yeast, we have crystal structures for only <20% of all proteins encoded in these organisms' genomes. We hope that structural data with more complete coverage will be available in the future. Second, we are still lacking an accurate estimator of protein designability. Contact density is a convenient but crude estimator of designability; we would prefer to use a direct estimate of the number of sequences folding into a particular structure, obtained, for example, from an accurate simulation of protein folding. Fortunately, such data are now becoming available (Meyerguz et al. 2007). In future studies, it should be possible to compare directly the relationship between designability and evolutionary rate.

Acknowledgments This work was supported by NIH Grant AI 065960. D.A.D. received support through an NIH center grant to the FAS Center for Systems Biology. We would like to thank Jesse Bloom for helpful comments on this work.

References

Agrafioti I, Swire J, Abbott J, Huntley D, Butcher S, Stumpf MPH (2005) Comparative analysis of the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* protein interaction networks. BMC Evol Biol 5:23

Appelgren H, Kniola B, Ekwall K (2003) Distinct centromere domain structures with separate functions demonstrated in live fission yeast cells. J Cell Sci 116:4035–4042

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy tat Soc B 57:289–300

Bloom JD, Adami C (2003) Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. BMC Evol Biol 3:21

Bloom JD, Drummond DA, Arnold FH, Wilke CO (2006) Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol 23:1751–1761

Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH (2005) Thermodynamic prediction of protein neutrality. Proc Natl Acad Sci USA 102:606–611

Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429:92–96

Creighton TE (1992) Proteins: structures and molecular properties, 2nd edn. Freeman, New York

Dean AM, Neuhauser C, Grenier E, Golding GB (2002) The pattern of amino acid replacements in $\alpha\backslash\beta$-barrels. Mol Biol Evol 19:1846–1864

Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L (2001) A fully automatic evolutionary classification of protein folds: Dali domain dictionary version 3. Nucleic Acids Res 29:55–57

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA 102:14338–14343

Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23:327–337

Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc Natl Acad Sci USA 96:4482–4487

Duret L, Mouchiroud D (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol 17:68–74

Edgar RC (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

England JL, Shakhnovich EI (2003) Structural determinant of protein designability. Phys Rev Lett 90:218101

England JL, Shakhnovich BE, Shakhnovich EI (2003) Natural selection of more designable folds: a mechanism for thermophilic adaptation. Proc Natl Acad Sci USA 100:8727–8731

Fraser HB (2005) Modularity and evolutionary constraint on proteins. Nature Genet 37:351–352

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

Goldman N, Thorne JL, Jones DT (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149:445–458

Gu W, Zhou T, Ma J, Sun X, Lu Z (2004) The relationship between synonymous codon usage and protein structure in *Escherichia coli* and *Homo sapiens*. Biosystems 73:89–97

Hahn MW, Kern AD (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22:803–806

Hakes L, Lovell SC, Oliver SG, Robertson DL (2007) Specificity in protein interactions and its relationship with sequence diversity and coevolution. Proc Natl Acad Sci USA 104:7999–8004

Herbeck JT, Wall DP, Wernegreen JJ (2003) Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont *Wigglesworthia*. Microbiology 149:2585–2596

Hirsh AE, Fraser HB (2001) Protein dispensability and rate of evolution. Nature 411:1046–1049

Holstege FCP, Jennings E, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95:717–728

Holstein SE, Ungewickell H, Ungewickell E (1996) Mechanism of clathrin basket dissociation: separate functions of protein domains of the DnaJ homologue auxilin. J Cell Biol 135:925–937

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

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314

Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res 12:962–968

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637

Kawabata T, Fukuchi S, Homma K, Ota M, Araki J, Ito T, Ichiyoshi N, Nishikawa K (2002) Gtop: a database of protein structures predicted from genome sequence. Nucleic Acids Res 30:294–298

Kim PM, Lu LJ, Gerstein MB (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314:1882–1883

Koshi JM, Goldstein RA (1995) Context-dependent optimal substitution matrices. Protein Eng 8:641–645

Kussell E (2005) The designability hypothesis and protein evolution. Protein Peptide Lett 12:111–116

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL (2005) Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol 22:1345–1354

Li H, Helling R, Tang C, Wingreen N (1996) Emergence of preferred structures in a simple model of protein folding. Science 273:666–669

Lin YS, Hsu WL, Hwang JK, Li WH (2007) Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. Mol Biol Evol 24:1005–1011

Mandel J (1982) Use of the singular value decomposition in regression analysis. Am Stat 36:15–24

Marais G, Duret L (2001) Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. J Mol Evol 52:275–280

Meyerguz L, Kleinberg J, Elber R (2007) The network of sequence flow between protein structures. Proc Natl Acad Sci USA 104:11627–11632

Mintseris J, Weng Z (2005) Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci USA 102:10930–10935

Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 291:177–196

Orešič M, Shalloway D (1998) Specific correlations between relative synonymous codon usage and protein secondary structure. J Mol Biol 281:31–48

Pal C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158:927–931

Pal C, Papp B, Hurst LD (2003) Rate of evolution and gene dispensability. Nature 421:496–497

Pal C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. Nat Rev Genet 7:337–348

Ren M, Villamarin A, Shih A, Coutavas E, Moore MS, LoCurcio M, Clarke V, Oppenheim JD, D'Eustachio P, Rush MG (1995) Separate domains of the Ran GTPase interact with different factors to regulate nuclear protein import and RNA processing. Mol Cell Biol 15:2117–2124

Rocha EPC, Danchin A (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol 21:108–116

Shakhnovich BE (2006) Relative contributions of structural designability and functional diversity in molecular evolution of duplicates. Bioinformatics 22:e440–e445

Shakhnovich BE, Deeds E, Delisi C, Shakhnovich E (2005) Protein structure and evolutionary history determine sequence space topology. Genome Res 15:385–392

Shakhnovich EI (1998) Protein design: a perspective from simple tractable models. Fold Des 3:R45–R58

Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE, Bussemaker HJ, White KP (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science 306:655–660

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101:6062–6067

Subramanian S, Kumar S (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168:373–381

Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW (2005) Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci USA 102:5483–5488

Wolynes PG (1996) Symmetry and the energy landscapes of biomolecules. Proc Natl Acad Sci USA 93:14249–14255

Yang ZH (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Zhang J, He X (2005) Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol 22:1147–1155