Structural Determinants of the Rate of Protein Evolution in Yeast

Jesse D. Bloom, D. Allan Drummond,† Frances H. Arnold,* and Claus O. Wilke‡*

*Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California; †Program in Computation and Neural Systems, California Institute of Technology, Pasadena, California; and ‡Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas at Austin

We investigate how a protein's structure influences the rate at which its sequence evolves. Our basic hypothesis is that proteins with highly designable structures (structures that are encoded by many sequences) will evolve more rapidly. Recent theoretical advances argue that structures with a higher density of interresidue contacts are more designable, and we show that high contact density is correlated with an increased rate of sequence evolution in yeast. In addition, we investigate the correlations between the rate of sequence evolution and several other structural descriptors, carefully controlling for the strong effect of expression level on evolutionary rate. Overall, we find that the structural descriptors that we consider appear to explain roughly 10% of the variation in rates of protein evolution in yeast. We also show that despite the well-known trend for buried residues to be more conserved, proteins with a higher fraction of buried residues, nonetheless, tend to evolve their sequences more rapidly. We suggest that this effect is due to the increased designability of structures with more buried residues. Our results provide evidence that protein structure plays an important role in shaping the rate of sequence evolution and provide evidence to support recent theoretical advances linking structural designability to contact density.

Introduction

Protein sequences evolve largely through the gradual accumulation of amino acid substitutions, and the extent of sequence divergence is quantified by the number of nonsynonymous substitutions per site, *d*N. Over 40 years ago, Zuckerkandl and Pauling (1965) observed that *d*N for homologous proteins is proportional to the time since divergence, indicating that *d*N measures a roughly constant average rate of fixed amino acid substitutions. However, it has also long been clear that different proteins accumulate substitutions at markedly different rates: Zuckerkandl and Pauling (1965) remarked how the large values of *d*N for hemoglobin were "spectacularly at variance" with the small values of *d*N for cytochrome *c*. The availability of full genome sequences now allows for much more extensive comparisons of rates of protein sequence evolution, and such analyses have confirmed the widely differing rates noted by Zuckerkandl and Pauling (1965). For example, *d*N varies approximately a 1,000-fold between the fastest and slowest evolving proteins in yeast (Drummond et al. 2005).

Zuckerkandl and Pauling (1965) and subsequently Ohta and Kimura (1971) and others (King and Jukes 1969) argued that variation in *d*N was due to differences in the selective constraints on proteins' sequences. The basic argument is that most sequence divergence is due to the fixation of mutations with little or no effect on a protein's function, and so the rate at which substitutions accumulate is proportional to the average fraction of mutations that are effectively neutral (Ohta and Kimura 1971; Brookfield 2000). This argument has now gained widespread acceptance, and numerous studies have used high-throughput genomic data to attempt to pinpoint the biological constraints that underlie the different rates of sequence evolution. Numerous properties have been found to correlate with a protein's *d*N, including the dispensability or essentiality of its encoding gene (Hirsh and Fraser 2001; Jordan et al. 2002; Wall et al. 2005; Zhang and He 2005), the number of other proteins with which it interacts (Fraser et al. 2002; Lemos et al. 2005), its length (Marais and Duret 2001; Lemos et al. 2005), its centrality in the protein interaction network (Hahn and Kern 2005), and its expression level (Pal et al. 2001; Drummond et al. 2005, 2006). However, a casual analysis of these correlations is complicated by the fact that most of these biological properties are also correlated with each other (Drummond et al. 2006). At this point, the only clear conclusions are that, by far, the most dominant trend is for highly expressed proteins to evolve slowly (Pal et al. 2001; Drummond et al. 2005, 2006) and that the other correlations are either much weaker or potentially due to confounding factors (Hurst and Smith 1999; Bloom and Adami 2003; Jordan et al. 2003; Pal et al. 2003; Hahn et al. 2004; Agrafioti et al. 2005; Drummond et al. 2006).

The fact that expression level correlates with *d*N much more strongly than properties reflecting a protein's biological role is consistent with protein mutagenesis experiments showing that deleterious mutations usually act by hindering the formation of a properly folded protein rather than specifically altering a protein's function (Shortle and Lin 1985; Pakula et al. 1986; Loeb et al. 1989; Bloom et al. 2005, 2006). Therefore, *d*N should be largely determined by the fraction of mutations that prevent adequate protein expression and folding. Highly expressed proteins are under an increased requirement for fidelity in expression and folding due to the costs of misfolded proteins, meaning they have a smaller fraction of effectively neutral mutations and so evolve more slowly (Drummond et al. 2005). A protein's biophysical properties can also influence the fraction of mutations that allow for adequate expression and folding, for example, protein mutagenesis experiments have shown that increasing a protein's thermodynamic stability dramatically increases its tolerance to mutations (Bloom et al. 2005, 2006). Another factor that has received little consideration with respect to its effect on *d*N, but which may significantly affect a protein's mutational tolerance, is the characteristics of the native structure itself.

The relationship between a protein's native structure and its mutational tolerance has been extensively studied

Key words: designability, protein structure, evolutionary rate, protein evolution, principal component regression, yeast.

E-mail: cwilke@mail.utexas.edu.

Mol. Biol. Evol. 23(9):1751–1761. 2006
doi:10.1093/molbev/msl040
Advance Access publication June 16, 2006

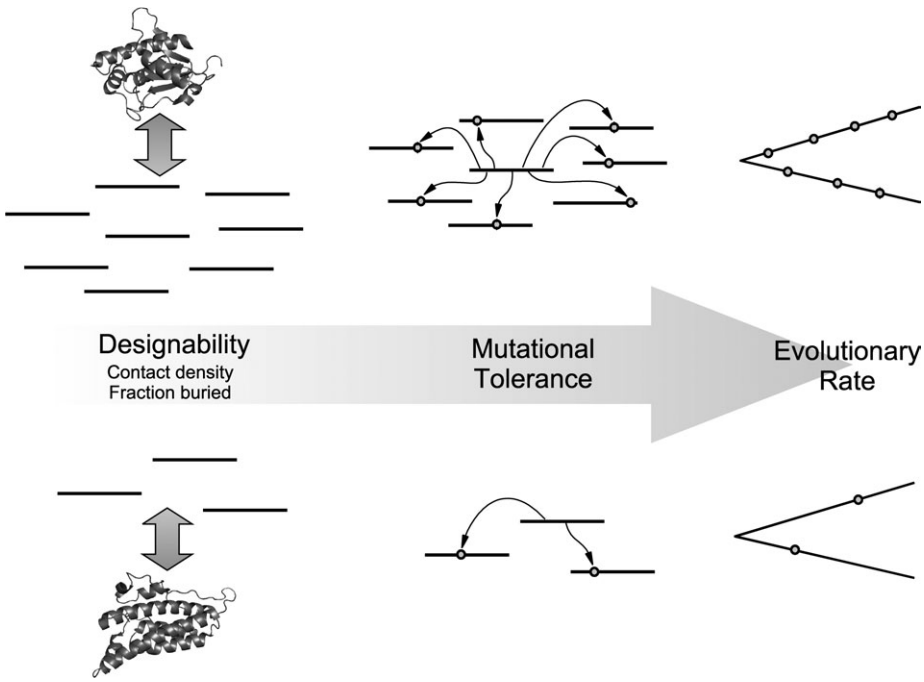

FIG. 1.—A model for how the designability of a protein's structure might affect the rate at which its sequence evolves. If many sequences fold into a given structure (highly designable), then many mutations preserve the structure, that is, the structure is mutationally tolerant. As a consequence, we observe rapid evolutionary divergence. Conversely, if few sequences fold into a given structure, then the structure is mutationally brittle, and evolutionary divergence is slow.

in the context of protein folding and design. These studies typically attempt to characterize a structure's "designability," defined as the total number of amino acid sequences that fold into that structure (Li et al. 1996; Kussell 2005). Because more designable structures are encoded by more sequences, proteins that fold into highly designable structures tend to be more tolerant to mutations and, thus, should evolve their sequences more rapidly (fig. 1). Most studies of protein designability have utilized simple computational models in which designability can be directly estimated by an extensive or exhaustive enumeration of different sequences (Li et al. 1996; Govindarajan and Goldstein 1996, 1997; Tiana et al. 2000; Chan and Bornberg-Bauer 2002; Irbäck and Troein 2002; Miller et al. 2002; Wingreen et al. 2004; Wroe et al. 2005; Zhang et al. 2005). The main conclusion of these model-protein simulations is that designability varies widely among structures; however, the simplicity of the models makes it difficult to extrapolate any quantitative measures of designability to real proteins. At the other end of the computational spectrum, a variety of studies have used state-of-the-art atomistic simulations to attempt to estimate the designabilities of real proteins (Zou and Saven 2000; Kono and Saven 2001; Voigt et al. 2001; Koehl and Levitt 2002; Larson et al. 2002). But these atomistic simulations are computationally expensive, and their accuracy is not known because there are no experimental measurements of protein designability with which they can be compared. Recently, England and Shakhnovich (2003) have proposed a general theory that relates designability to the pattern of contacts between residues in the native structure. Their approach is based on the argument of Wolynes (1996) and Shakhnovich (1998) that a structure's designability can be estimated as the number of sequences that fold into that structure with an energy below some threshold. By assuming that the energy of a structure is due to pairwise interactions between residues, England and Shakhnovich (2003) show that the designability D is given by a series in traces of powers of the structure's contact matrix $\mathbf{C}$,

$$D = \sum_{n=2}^{\infty} (\mathrm{Tr}\, \mathbf{C}^n) a_n. \qquad (1)$$

This theory has been verified with simulations on simple model proteins (England and Shakhnovich 2003; England et al. 2003; Tiana et al. 2004), but, unlike earlier simulations, it has a theoretical basis and so can, in principle, be applied broadly. However, the coefficients a_n in equation (1) cannot be calculated, but the designability can be estimated by truncating the series after the first term. (An alternative method of estimating the series in equation (1) as the maximum eigenvalue of the contact matrix gives comparable results, as discussed below.) The first term of equation (1) is just equal to the contact density (the average number of contacts per residue) (England and Shakhnovich 2003; England et al. 2003), and so, this truncation recovers the predictions of Wolynes (1996) and Shakhnovich (1998) that designability is approximated by contact density.

Here we use this predicted relationship between contact density and designability as the basis for exploring the contribution of protein structure to evolutionary rate in yeast. Although numerous earlier studies have shown that the conservation of residues at individual sites is influenced by structural characteristics such as secondary structure or solvent exposure (Overington et al. 1992; Koshi and

Goldstein 1995; Thorne et al. 1996; Goldman et al. 1998; Mirny and Shakhnovich 1999; Bustamante et al. 2000; Dokholyan and Shakhnovich 2001; Dean et al. 2002; Marsh and Griffiths 2005), our work looks at how a protein's structure affects its global rate of sequence evolution. We examine the correlation of *d*N with contact density and several other structural descriptors (fraction of buried residues, secondary structure composition, length, and fold classification), while statistically controlling for the effect of expression level. Our work shows that *d*N is influenced by protein structure in a way that suggests that proteins with more designable structures evolve their sequences more rapidly.

Materials and Methods

All 33,449 protein structures present in the protein data bank (PDB) on 13 November 2005 were downloaded as mmCIF files. The downloaded files were parsed to get the sequences of all the proteins, except for 1quz, 1jhq, 1zhe, 1quz, 2ad1, 1zir, and 2etg, which could not be parsed effectively. These parsed sequences included only those residues with coordinates—residues without coordinates were excluded from the sequences. Nonglycine residues that lacked any side-chain atoms were also excluded from the sequences. This procedure yielded a total of 73,121 protein sequences.

The sequences of all *Saccharomyces cerevisiae* open reading frames (ORFs) were downloaded from ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/orf_dna/ on 19 October 2004. All the genes that could be translated were considered. This procedure yielded 5,865 proteins. To match protein structures with these yeast proteins, we Blasted (Altschul et al. 1990) each protein against all the PDB protein sequences. Any matches with a Blast *E* value of at least 10^{-5} were then aligned using ClustalW (Thompson et al. 1994), and if the number of identities in the total length of the alignment was >80%, then the match was saved. If there were multiple such matches to a protein, then the best match was saved as the hit for that protein. This process yielded 275 matches. Because the number of proteins with solved structures is small compared with the number of proteins with known sequences, restricting our data set to yeast proteins with >80% similarity to a sequence in the PDB limits the size of our data set. However, we felt it was important to set this relatively stringent criterium for sequence identity to ensure that the PDB structures accurately represented the actual folded conformations of the yeast proteins with which they were matched.

For each yeast protein with a match, we recorded the aligned residues, the secondary structure for each aligned residue, and the percent solvent-accessible area for each aligned residue. The latter number was computed using only atoms within that protein chain in the PDB structure (i.e., we did not consider surface area buried by atoms of other protein chains in the PDB structure). We first calculated the exposed surface area using the program given by McConkey et al. (2002) and then normalized these values by the reference surface areas of an extended Gly-X-Gly peptide, as given in Creighton (1992, Table 4.4, p. 142). We counted a residue as buried if it had less than 25% solvent accessibility and as exposed, otherwise.

We calculated contact maps of all the 275 PDB structures. We considered 2 residues in contact if any of their 2 heavy (nonhydrogen) atoms were within a distance of 4.5 Å from each other and if the 2 residues were not immediate sequence neighbors (i.e., we excluded trivial contacts). We then determined the contact density for each structure by calculating the average number of contacts per residue.

Protein chains were assigned to the Structural Classification of Proteins (SCOP) database (Murzin et al. 1995) classes given in version 1.69 available at http://scop.mrc-lmb.cam.ac.uk/scop/parse/index.html. For each protein chain, we first searched for the PDB structure ID in the downloadable dir.cla.scop.txt_1.69 file. If we found the ID, we then searched for the mmCIF file chain ID in the file. In cases where a structure had multiple chains with different SCOP classes, we took care to ensure that the mmCIF chain ID was matched appropriately with the correct SCOP class (the chain ID used in the mmCIF file is not always the same as the one used in the PDB file for the same structure, and the SCOP classifications are made according to the PDB chain ID). Some chains have different regions assigned to distinct SCOP classes. If we found multiple entries for different regions of the chain, we recorded the SCOP class only if all regions of the chain were assigned to the same class.

We calculated evolutionary rates (*d*N) using the reciprocal-shortest-distance method (Wall et al. 2003). All ORFs in *S. cerevisiae* were Blasted against those in *Saccharomyces bayanus* and vice versa. Pairwise hits with an *E* value of $<10^{-20}$ were retained and aligned with ClustalW, using the aligned protein sequences to align the nucleotide sequences. Evolutionary rates, the numbers of nonsynonymous substitutions per nonsynonymous site (*d*N) and synonymous substitutions per synonymous site (*d*S), were computed for these hits using the Phylogenetic Analysis by Maximum Likelihood (Yang 1997) program codeml operating on codons with a 9–free-parameter model for codon frequencies. Pairs with less than 80% aligned residues were discarded because there are no well-established methods for dealing with gaps when calculating evolutionary distances. Remaining aligned gene pairs having each other as the shortest-distance (smallest *d*N) hit were designated orthologs and used in our analysis. Among the previously identified 275 ORFs with a match to a PDB structure, we found *S. bayanus* orthologs in 203 cases.

We calculated evolutionary rates at buried/exposed sites and sites of specific secondary structure by discarding all but the relevant portions of the ortholog alignments generated to compute overall evolutionary rates. For example, to compute the evolutionary rates at buried sites, we considered only buried residues (identified as described above) and assembled the corresponding codons into a reduced pair of ortholog sequences, from which evolutionary rates were calculated exactly as described above. This procedure was carried out for buried and exposed residues and for residues corresponding to the 4 secondary structure types of helix (DSSP class H), sheet (DSSP class E), turn (DSSP classes S and T), and coil (DSSP classes B, G, I, and "."). (DSSP is the "Dictionary of Protein Secondary Structure," Kabsch and Sander 1983.)

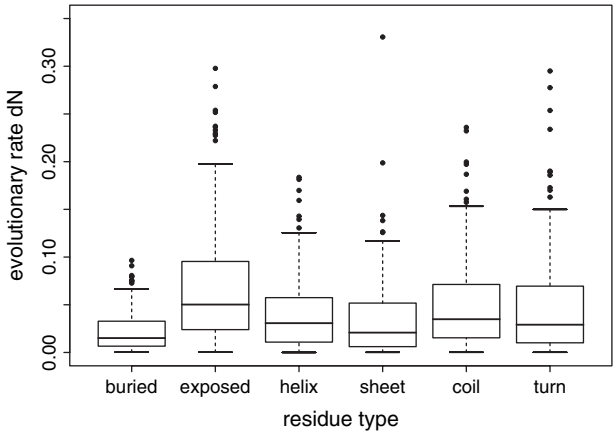

FIG. 2.—Distributions of evolutionary rate, *d*N, for different residue types, shown as modified boxplots. (Boxes enclose all data between the first and third quartile and are divided at the median. Whiskers at top extend to the maximum observation or 1.5 times the box height from the top of the box, whichever is smaller. Similarly, whiskers at bottom extend to the minimum observation or 1.5 times the box height from the bottom of the box, whichever is larger. All data points outside the range of the whiskers are drawn individually. See, for example, Sokal and Rohlf [1994]). For all residue types, the distribution of *d*N is heavily skewed and, thus, not well characterized by its mean and standard deviation. For the 2 residue types "exposed" and "helix," one outlier each falls outside the top boundary of the graph. The evolutionary rates in all groups are significantly different from each other with $P < 0.01$ (Wilcoxon signed-rank test), apart from helix and coil ($P = 0.01$) and coil and turn ($P = 0.37$, not significant).

We calculated codon adaptation indices (CAIs) exactly as described by Sharp and Li (1987), using tabulated codon relative adaptiveness values for *S. cerevisiae*. We used expression data from Holstege et al. (1998). After discarding all ORFs for which we did not have expression data, we arrived at our final data set of 194 ORFs. This data set is given as supplementary table, Supplementary Material online.

All statistical analyses were carried out with the statistics software R, version 2.1.1 (R Development Core Team 2005). Principal component regression was done using the R package "pls." Unless mentioned otherwise, all analyses were carried out on ranks, rather than on the actual values of the quantities. In particular, all correlations are Spearman correlations, and all principal component regression analyses were carried out on rank-transformed data.

Results

Differences in Substitution Rates of Different Classes of Residues

For each ORF, we determined the evolutionary rate (*d*N) of the buried and of the exposed residues and also of residues belonging to each of the 4 secondary structure classes of helix, sheet, turn, and coil. We found that exposed sites evolve substantially faster than buried sites, whereas secondary structure has little effect on evolutionary rate (fig. 2). For all residue types, the distribution of evolutionary rates was highly skewed. For example, the fastest evolving buried sites evolve much faster than the median exposed site. Overall, our findings for type-specific evolutionary rates confirmed the consensus in the literature that solvent accessibility has a strong effect on the conservation of individual residues, whereas secondary structure type has at most a weak effect (Goldman et al. 1998; Bustamante et al. 2000; Dean et al. 2002).

Effect of Contact Density on Evolutionary Rate

Because buried sites are more conserved, we might expect that proteins with a larger fraction of buried sites should evolve slower. On the other hand, the prediction of England and Shakhnovich (2003) is that proteins with a higher contact density are more designable and, thus, would be expected to evolve faster. To find out which of the two views is correct, we correlated the overall evolutionary rate *d*N with contact density and the fraction of residues that are buried. The former is the average of the number of contacts per residue and is the quantity treated theoretically by England and Shakhnovich (2003) (the first term of eq. 1). The latter does not explicitly count contacts but is strongly correlated with contact density (table 1). Throughout this paper, we usually report results, which hold for both measures, for only one measure, the contact density. An alternative method for estimating the designability as given by equation (1) is to use the maximum eigenvalue of the contact matrix (England and Shakhnovich 2003); doing so yields results that are highly similar to those for contact density for all correlations shown in table 1 (the Spearman correlation between the maximum eigenvalue and *d*N is $\rho = 0.25$).

Table 1
Spearman Correlations between Variables Considered in This Study

	*d*N	*x*	*c*	f_{bur}	*d*	*L*	f_H	f_E	f_T
x	−0.58***								
c	−0.40***	0.71***							
f_{bur}	0.29***	−0.09	0.20*(*)						
d	0.24**(*)	−0.01	0.20*(*)	0.80***					
L	0.19*(*)	−0.14(*)	0.13	0.79***	0.57***				
f_H	0.03	−0.01	0.05	−0.05	0.13	0.07			
f_E	−0.01	0.06	−0.05	0.10	0.06	−0.04	−0.80***		
f_T	−0.10	0.10	0.17*	0.19*(*)	0.19*(*)	0.06	−0.46***	0.28***	
f_C	0.02	−0.10	0.02	−0.02	−0.33***	−0.01	−0.40***	−0.06	0.11

NOTE.—Significance levels: ***$P < 0.001$; **$P < 0.01$; *$P < 0.05$. *d*N: nonsynonymous evolutionary rate; *x*: gene expression level; *c*: CAI; f_{bur}: fraction of buried sites; *d*: contact density; *L*: protein length; f_H, f_E, f_T, and f_C: fraction of sites with secondary structure helix, sheet, turn, and coil, respectively. Significance levels in parentheses disappear after correction for multiple testing (Benjamini and Hochberg 1995).

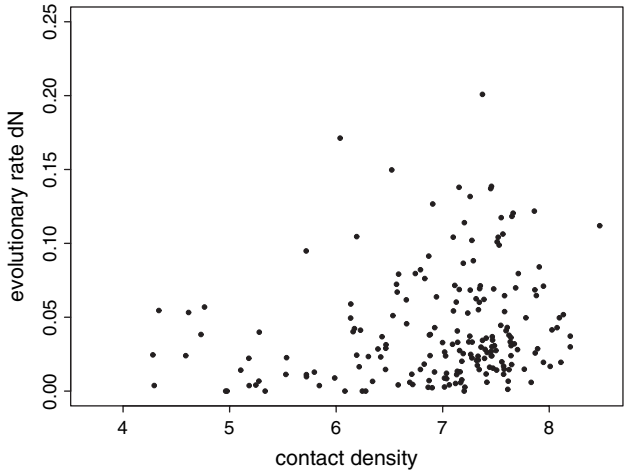

FIG. 3.—Evolutionary rate *d*N as a function of a protein's contact density.

As predicted by the hypothesis that more designable proteins should evolve faster (fig. 1), we found that contact density and the fraction of buried sites correlated significantly with *d*N (table 1, fig. 3). However, this correlation does not imply per se that the increased rate of evolution is caused by contact density. It is well known that more highly expressed proteins evolve more slowly in yeast (Pal et al. 2001), and therefore, we always have to control for expression level before we can conclude that any quantity has an effect on *d*N (Bloom and Adami 2003; Pal et al. 2003; Drummond et al. 2006). We calculated the correlations between contact density and both expression level as measured by DNA microarrays (Holstege et al. 1998) and CAI, another proxy for gene expression (table 1). Expression level is not significantly correlated with contact density. There was a weak correlation between CAI and contact density, but this correlation had the opposite sign from what we would expect if a correlation between CAI and contact density were to cause the correlation between contact density and *d*N. These results indicate that the correlation between contact density and evolutionary rate is not caused by an underlying correlation between contact density and expression level.

The results from the analysis of evolutionary rate at buried and exposed sites (fig. 2) seem to be at odds with the results from the analysis of the overall evolutionary rate of the proteins (fig. 3). Buried sites tend to be more conserved than exposed sites, but the overall evolutionary rate increases with increasing contact density (and also fraction of buried sites). However, this apparent paradox can be understood if we consider the effect of high contact density on buried and exposed sites separately (fig. 4). Whereas the *d*N at buried sites shows a moderate increase with increasing contact density, the *d*N at exposed sites grows dramatically. Even though proteins with high contact density have a reduced fraction of exposed residues, the residues that are exposed in these proteins evolve very rapidly. Therefore, the reduction in the fraction of exposed residues is more than compensated for by the increased variability of exposed residues in proteins with high contact density.

Effect of Protein Length on Evolutionary Rate

We found a significant correlation between protein length and *d*N and a strong correlation between length and contact density (table 1). This observation prompted us to investigate the relationship between contact density and protein length. We found that the correlation between these 2 quantities stems primarily from short proteins (fig. 5). For very short proteins, there is a large variation in contact density. In this regimen, contact density can be as low as 4 or as high as 7. As the protein length increases, there is an overall increase in contact density, but at the same time the variability in contact density decreases. Eventually, contact density levels off and remains in a range between approximately 6 and 8. Therefore, we next calculated the correlations between contact density, fraction of buried sites, protein length, and *d*N separately for short (<250 residues) and for long (≥ 250 residues) proteins (table 2). For short proteins, both the fraction of buried sites and the protein length showed a significant positive correlation with *d*N. For long proteins, on the other hand, only the correlation between the fraction of buried sites and *d*N was positive and significant; the correlation between length and *d*N turned negative and lost significance. The correlation between contact density and *d*N was similar to, but weaker

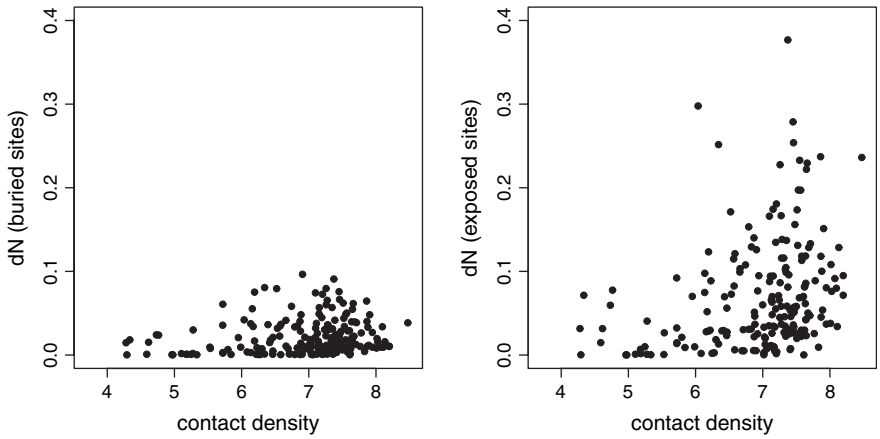

FIG. 4.—Evolutionary rate *d*N for buried sites (left panel) or exposed sites (right panel) only, as a function of the protein's contact density.

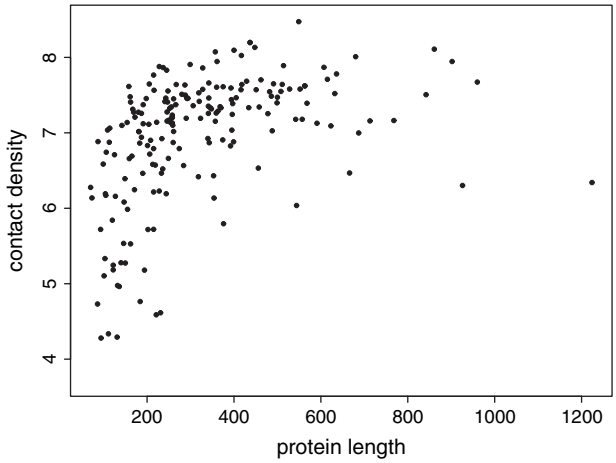

FIG. 5.—Contact density as a function of protein length.

than, the correlation of the fraction of buried sites and *d*N but was not significant at the 5% level. Finally, we noted that the correlation between length and either contact density or fraction of buried sites was stronger for short proteins (table 2).

We then wanted to know whether the correlation between *d*N and protein length we observed was potentially caused by a biased selection of protein structures because our data set was biased toward short proteins (median length is 263.5 residues for the 194 proteins with structural information and a *S. bayanus* ortholog vs. 440 residues in all 4,532 *S. cerevisiae* proteins with a *S. bayanus* ortholog). Therefore, we calculated the correlation between length and *d*N for all ORFs, including those without structural information. We found that the overall correlation between length and *d*N was significant but weak (table 3), as previously reported (Drummond et al. 2006). When we considered long and short proteins separately, we found a similar picture as before. Length correlates much more strongly with *d*N for short proteins than for long proteins. In fact, the amount of variance in *d*N explained by length alone is approximately 10 times larger for short proteins than for long proteins.

Because longer proteins are known to be expressed at lower levels (Coghlan and Wolfe 2000; Munoz et al. 2004), a positive correlation between length and *d*N could also be caused indirectly by expression-level differences. To ascertain whether differences in expression level could explain the difference in correlation between length and *d*N for short and long proteins, we also calculated the correlation between expression and length for short and long proteins separately. For all ORFs with evolutionary rate data, this correlation was almost identical for short and long proteins (table 3). For the 194 ORFs with structural information, neither length nor contact density correlated significantly with expression level when we considered short and long proteins separately (not shown). Therefore, it is unlikely that the increased correlation between length and *d*N for short proteins is an artifact of expression-level differences in these proteins.

Table 2
Spearman Correlations ρ for All ORFs with Structural Information Calculated Separately for Short and Long Proteins

	$L < 250$ ($n = 85$)	$L \geq 250$ ($n = 109$)
$\rho(d, d\text{N})$	0.20^{+}	0.14
$\rho(f_{\text{bur}}, d\text{N})$	0.27*	0.21*
$\rho(L, d\text{N})$	0.22*	−0.14
$\rho(d, L)$	0.46***	0.22*
$\rho(f_{\text{bur}}, L)$	0.61***	0.47***

NOTE.—Significance levels: $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$; $^{+}P < 0.01$. *n* is the number of ORFs in each group, and other symbols are defined as in table 1.

Table 3
Spearman Correlations ρ for All ORFs with Evolutionary Rate Data Calculated Separately for Short and Long Proteins

	All Lengths ($n = 4{,}523$)	$L < 250$ ($n = 935$)	$L \geq 250$ ($n = 3{,}597$)
$\rho(L, d\text{N})$	0.07***	0.13***	0.04*
$\rho(L, x)$	−0.24***	−0.17***	−0.15***

NOTE.—Significance levels: $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$. *n* is the number of ORFs in each group, and other symbols are defined as in table 1.

Effect of Secondary Structure Composition on Evolutionary Rate

We also asked whether a protein's composition of secondary structure types has an influence on evolutionary rate. For example, do proteins that are composed primarily of helices evolve faster or slower than other proteins? To this end, we correlated *d*N with the fraction of helix sites f_{H}, fraction of sheet sites f_{E}, fraction of turn sites f_{T}, and fraction of coil sites f_{C} (table 1). We found that none of these quantities correlated significantly with *d*N and neither did they correlate with expression level or CAI (apart from a marginally significant correlation between CAI and the fraction of turn sites). However, not surprisingly, we found several strong and significant correlations among the different secondary structure measures (table 1).

Principal Component Regression

The correlation analysis presented in the previous subsections is useful to get an initial understanding of the data and to find broad trends but cannot detect more subtle interactions between the various predictor variables or quantify the amount of variance in *d*N these predictors explain independently of each other. Therefore, we carried out a principal component regression (Mandel 1982; Drummond et al. 2006) of *d*N against the 9 predictor variables of expression level, CAI, fraction of buried sites, contact density, protein length, and the fractions of the 4 secondary structure types. Table 4 summarizes the results from the principal component regression. We found 5 components that made a significant contribution to the regression, and the total amount of variance explained was 43.34%. The component composition is given in figure 6. Among the components that contributed significantly to the regression, Component 1 measures primarily the contact density of a protein. Component 2 measures primarily aspects of secondary structure. Component 3 represents a protein's expression level. Component 6 measures primarily the difference between contact density and length. Finally, Component 7 measures the

Table 4
Percent Variance Explained in *d*N and in the Predictor Variables as Found by a Principal Component (PC) Regression of *d*N against 9 Predictor Variables

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	Total
*d*N	**2.81**	**1.32**	**33.59**	0.00	0.00	**3.54**	**1.27**	0.65	0.15	43.34
Predictor space	**28.83**	**23.86**	**19.10**	12.47	7.94	**3.63**	**1.27**	0.65	0.15	100.00

NOTE.—Predictor variables are x, c, f_{bur}, d, L, f_H, f_E, f_T, and f_C, with symbols defined as in table 1. Principal components that make a statistically significant contribution to the variation in *d*N are shown in boldface ($P < 0.05$ for all components in boldface). The percent values for the individual components do not sum exactly to the numbers given under total because of rounding errors. See figure 6 for component composition.

difference between expression level and CAI. The components that did not contribute significantly to the regression represent secondary structure (Components 4, 5, and 9, not shown) or differences between contact density and the fraction of buried sites (Component 8). Thus, the component measuring expression level explained approximately 34% of the variation in *d*N, whereas all other components together explained approximately 10% of the variation in *d*N.

We also regressed *d*N separately against expression level and CAI and against the 7 structural variables to determine how much variance these 2 groups of variables explained individually. Our results were very much in agreement with those of the joint regression against all 9 predictor variables. The regression of *d*N against expression level and CAI explained 34.03% of the variance, whereas regression of *d*N against the 7 structural variables explained 11.97% of the variance. Thus, the total amount of variance explained in the regression against all 9 variables is approximately the sum of the amounts of variance explained from the 2 individual regressions, and therefore, the regression of *d*N against structural variables is unlikely to be confounded by expression-level effects.

Does Structure Classification Determine Contact Density or Evolutionary Rate?

We investigated the relationship between protein structure classification and both contact density and evolutionary

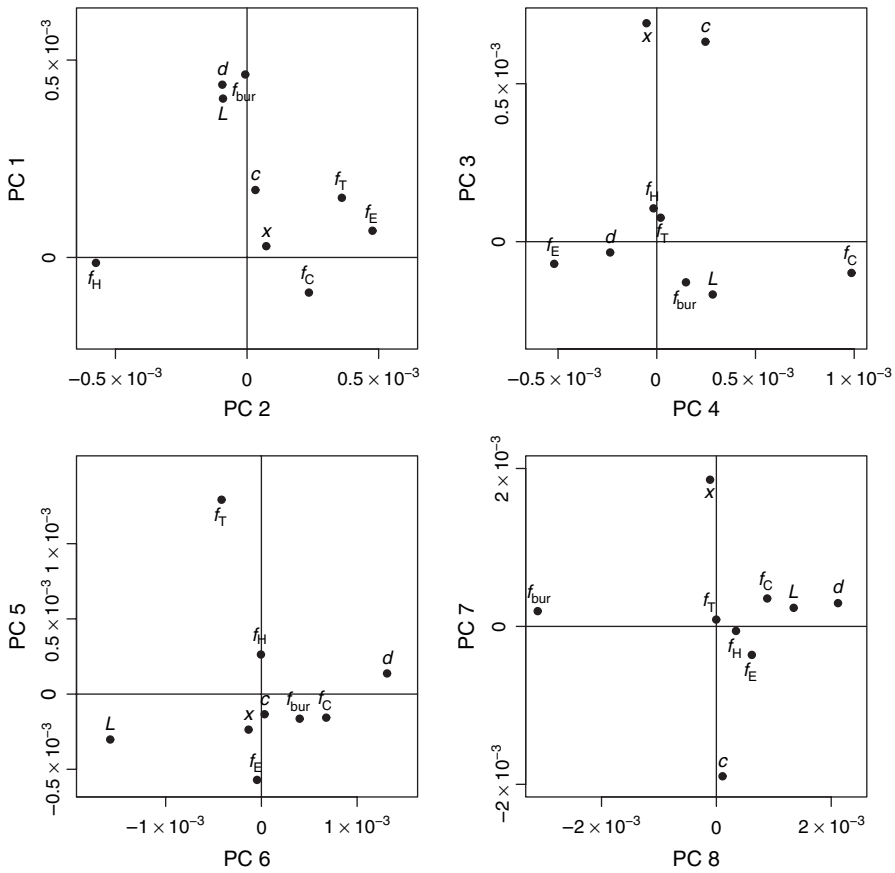

FIG. 6.—Composition of the first 8 principal components (PCs). Symbols representing predictor variables are as defined in table 1. Each dot represents the contribution of the corresponding predictor variable to the principal components. For example, in the top-left panel, we see that L makes a strong, positive contribution to PC 1 and a weak, negative contribution to PC 2.

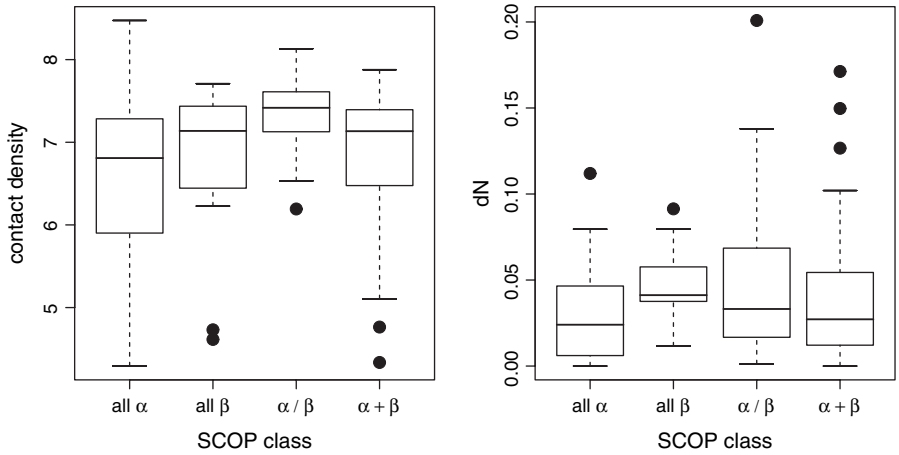

FIG. 7.—Modified boxplots of the distributions of contact density (left) and evolutionary rate (right) for different SCOP classes. The following are not shown: 3 proteins classified as multidomain (α and β), 2 classified as membrane and cell-surface proteins and peptides, 2 for which the classification was ambiguous, and 55 for which no classification could be determined.

rate for 137 proteins (out of our final data set of 194) for which we could determine the class of the structure according to the SCOP (Murzin et al. 1995). Figure 7 shows that there are some differences in both contact density and evolutionary rate among proteins with different structure classes but that these differences are relatively minor. We found all-α proteins to have the lowest median contact density and the highest variability in contact density, whereas α/β proteins had the highest median contact density and the lowest variability. The median contact density was significantly higher for α/β proteins than for all-α proteins ($P = 0.0001$) or for α + β proteins ($P = 0.0008$) but not for all-β proteins ($P = 0.070$) (Wilcoxon rank-sum tests with false discovery rate correction [Benjamini and Hochberg 1995] for multiple testing). The median evolutionary rate was the highest for all-β proteins, but the individually largest evolutionary rates were observed for α/β and α + β proteins. However, neither all-β proteins nor any other class of proteins had a significantly elevated evolutionary rate after correction for multiple testing (Wilcoxon rank-sum tests with false discovery rate correction).

Discussion

Our results show that protein structure has a moderate effect on protein evolutionary rate. Furthermore, this effect is consistent with the idea that proteins with more designable structures, as indicated by higher contact densities (England and Shakhnovich 2003), tend to evolve more rapidly. Specifically, contact density, the fraction of buried sites, and protein length each showed a significant correlation with *d*N, with R^2 values between 4% and 8%. These correlations were not caused by confounding effects due to cocorrelations with protein expression level, which is the dominant determinant of evolutionary rate in yeast (Pal et al. 2001; Drummond et al. 2006). The contact density and the fraction of buried sites were strongly correlated with each other and had roughly comparable predictive power for the evolutionary rate, although the fraction of buried sites tended to be a slightly better predictor than contact density. We found that secondary structure composition and protein-fold classification had almost no effect on evolutionary rate. Protein length was significantly correlated with both *d*N and contact density, making it difficult to fully elucidate the separate contributions of length and contact density to evolutionary rate. However, the fact that length is positively correlated with *d*N only for short proteins, but that higher contact density leads to faster evolutionary rates for both long and short proteins, supports the notion that higher contact density increases evolutionary rate independent of length effects.

We corroborated the correlations we observed with a principal component regression of *d*N against all structural predictor variables plus protein expression level and CAI. We found that the principal components that measured primarily the aspects of contact density or fraction of buried residues explained 6.35% of the variation in *d*N. In comparison, the principal component related to secondary structure explained only 1.32% of the variation in *d*N. Overall, the principal component regressions indicated that the structural characteristics we considered explained between 10% and 12% of the variation in evolutionary rate and that the important variables in this explanation were contact density (or the highly correlated variable of fraction of buried residues) and protein length.

We had to restrict our analysis to those yeast proteins to which we could confidently assign structures. Although there are tools that use homology modeling or other computational methods to predict the structure adopted by a protein sequence, we chose to consider only those yeast proteins that matched with at least 80% identity to an experimentally determined structure. This choice eliminated the possibility of introducing biases due to inaccuracies of the structure prediction tools, but it also substantially reduced the size of our data set because the number of proteins with experimentally solved structures is relatively low. Overall, we were able to match structures with only 194 of the 4,223 yeast proteins for which we had expression and evolutionary information. The proteins that we matched with structures tended to be both more highly expressed and slower evolving than all yeast proteins for which we had expression and evolutionary information

(the median transcript per cell levels were 2.7 and 0.8, respectively; the median *d*N values were 0.03 and 0.08, respectively). One clear consequence of limiting ourselves to proteins with experimentally solved structures is that we are excluding those (usually faster evolving) proteins that contain large regions that are intrinsically disordered (Brown et al. 2002). It is possible that the subset of yeast proteins with structures also contains other biases that affect the correlation between contact density and evolutionary rate. However, it is impossible to assess any such effects with the currently available protein structures, and so, further analysis of this question will have to await the experimental determination of more protein structures.

At first glance, our findings that proteins with higher contact densities (and therefore a higher fraction of buried residues) evolve more slowly seem at odds with the tendency for buried residues to be more conserved (Koshi and Goldstein 1995; Goldman et al. 1998; Mirny and Shakhnovich 1999; Bustamante et al. 2000; Dean et al. 2002). The key point is to realize that although buried residues are generally more conserved than exposed ones, increasing the fraction of buried residues leads to an overall increase in the evolutionary rate of all residues in the protein, primarily via a dramatic increase in *d*N for the exposed residues. We suggest that the increase in designability that accompanies high contact density enhances the mutational tolerance of exposed residues enough to more than offset the higher fraction of slower evolving buried sites. A potential reason for the elevated evolutionary rate at exposed sites is increased protein stability. Highly designable proteins tend to be more stable (Wingreen et al. 2004), and stability promotes mutational tolerance (Bloom et al. 2005, 2006). A larger fraction of buried residues suggests a more robust protein core, whose stability may thus allow loops and other surface features to mutate more freely. In other words, regions of high contact density form stabilizing cores of conserved, highly interacting amino acids that allow other exposed regions of the sequence to mutate more freely (Shakhnovich et al. 2005).

Overall, our results support the notion that proteins with more designable structures tend to evolve their sequences more rapidly. These findings suggest that the structures of real proteins may differ substantially in their mutational tolerances and that this effect is manifested in the rates of sequence evolution across the yeast proteome. However, the overall contribution of protein structure to evolutionary rate that we detect is still much smaller than that made by protein expression level: the principal component representing protein expression explains 3 times more variance in *d*N than all the structural components. Earlier work has shown that highly expressed proteins are more likely to adopt mixed α-helix and β-sheet folds (Jansen and Gerstein 2000). However, this tendency did not lead to a net relationship between contact density and expression level in our data set because we found these 2 variables to be uncorrelated. Therefore, protein structure appears to make an independent contribution to evolutionary rate, although expression level remains the more dominant force in determining the rate of sequence evolution.

However, it is possible that our analysis underestimates the contribution of protein structure to evolutionary rate. In attempting to apply the designability theory of England and Shakhnovich (2003), our ignorance of the a_n coefficients in equation (1) has forced us to make the severe approximation of truncating all higher order powers of the contact matrix and estimating designability solely from contact density. The effect of this truncation is unknown, but contact density is surely less informative than the full series of equation (1). Furthermore, the derivation of equation (1) by England and Shakhnovich (2003) makes the twin assumptions that proteins are stabilized only by pairwise contacts and that designability is simply equal to the number of sequences that fold to a structure with an energy below some cutoff—both these assumptions are unlikely to be completely true for real proteins. For these reasons, contact density is clearly an imperfect proxy for designability, and our inability to more accurately quantify designability probably causes us to underestimate its true contribution to protein evolutionary rate.

Despite these caveats, our work makes an important contribution by providing some of the first evidence about how a protein's structure influences its evolutionary rate. Structure is clearly only one of the many factors that determine the extent of constraint on a protein's sequence, but its effect appears to be significant. Future progress in developing theoretical treatments of structural designability and in better characterizing the other factors that constrain sequence evolution should eventually allow for improved measurements of the net effect of structure on protein evolution. For now, we simply add a structure's designability to the pantheon of factors that shape the rate of protein sequence evolution.

Supplementary Material

The data set of 194 *S. cerevisiae* ORFs analyzed in this work is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

J.D.B. was supported by a Howard Hughes Medical Institute predoctoral fellowship, and C.O.W. was supported by National Institutes of Health grant AI 065960.

Literature Cited

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.

Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–10.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat 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, Labthavikul ST, Otey CR, Arnold FH. 2006. Protein stability promotes evolvability. Proc Natl Acad Sci USA 103:5869–74.

Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C, Arnold FH. 2005. Thermodynamic prediction of protein neutrality. Proc Natl Acad Sci USA 102:606–11.

Brookfield JFY. 2000. What determines the rate of sequence evolution? Curr Biol 10:R410–1.
Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol 55:104–10.
Bustamante CD, Townsend JP, Hartl DL. 2000. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. Mol Biol Evol 17:301–8.
Chan HS, Bornberg-Bauer E. 2002. Perspectives on protein evolution from simple exact models. Appl Bioinformatics 1:121–44.
Coghlan A, Wolfe KH. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. Yeast 16:1131–45.
Creighton TE. 1992. Proteins: structures and molecular properties. 2nd ed. New York: Freeman.
Dean AM, Neuhauser C, Grenier E, Golding GB. 2002. The pattern of amino acid replacements in α/β-barrels. Mol Biol Evol 19:1846–64.
Dokholyan NV, Shakhnovich EI. 2001. Understanding heirarchical protein evolution from first principles. J Mol Biol 312:289–307.
Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA 102:14338–43.
Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol 23:327–37.
England JL, Shakhnovich BE, Shakhnovich EI. 2003. Natural selection of more designable folds: a mechanism for thermophilic adaptation. Proc Natl Acad Sci USA 100:8727–31.
England JL, Shakhnovich EI. 2003. Structural determinant of protein designability. Phys Rev Lett 90:218101.
Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science 296:750–2.
Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149:445–58.
Govindarajan S, Goldstein RA. 1996. Why are some protein structures so common? Proc Natl Acad Sci USA 93:3341–5.
Govindarajan S, Goldstein RA. 1997. The foldability landscape of model proteins. Biopolymers 42:427–38.
Hahn MW, Conant GC, Wagner A. 2004. Molecular evolution in large genetic networks: does connectivity equal constraint? J Mol Evol 58:203–11.
Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol 22:803–6.
Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. Nature 411:1046–9.
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 95:717–28.
Hurst LD, Smith NGC. 1999. Do essential genes evolve slowly? Curr Biol 9:747–50.
Irbäck A, Troein C. 2002. Enumerating designing sequences in the HP model. J Biol Phys 28:1–15.
Jansen R, Gerstein M. 2000. Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. Nucleic Acids Res 28:1481–8.
Jordan IK, Rogozin IB, Wolf YI, Koonin EV. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res 12:962–8.
Jordan IK, Wolf YI, Koonin EV. 2003. 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 3:1.
Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–637.
King JL, Jukes TH. 1969. Non-darwinian evolution. Science 165:788–98.
Koehl P, Levitt M. 2002. Protein topology and stability define the space of allowed sequences. Proc Natl Acad Sci USA 99: 1280–5.
Kono H, Saven JG. 2001. Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. J Mol Biol 306:607–28.
Koshi JM, Goldstein RA. 1995. Context-dependent optimal substitution matrices. Protein Eng 8:641–5.
Kussell E. 2005. The designability hypothesis and protein evolution. Protein Pept Lett 12:111–6.
Larson SM, England JL, Desjarlais JR, Pande VS. 2002. Thoroughly sampling sequence space: large-scale protein design of structural ensembles. Protein Sci 11:2804–13.
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 22:1345–54.
Li H, Helling R, Tang C, Wingreen N. 1996. Emergence of preferred structures in a simple model of protein folding. Science 273:666–9.
Loeb DD, Swanstrom R, Everitt L, Manchester M, Stamper SE, Hutchison CA. 1989. Complete mutagenesis of the HIV-1 protease. Nature 340:397–400.
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 52:275–80.
Marsh L, Griffiths CS. 2005. Protein structural influences in rhodopsin evolution. Mol Biol Evol 22:894–904.
McConkey BJ, Sobolev V, Edelman M. 2002. Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure. Bioinformatics 18:1365–73.
Miller J, Zeng C, Wingreen NS, Tang C. 2002. Emergence of highly designable protein-backbone conformations in an off-lattice model. Proteins Struct Funct Genet 47:506–12.
Mirny LA, Shakhnovich EI. 1999. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 291: 177–96.
Munoz ET, Bogarad LD, Deem MW. 2004. Microarray and EST database estimates of mRNA expression levels differ: the protein length versus expression curve for *C. elegans*. BMC Genomics 5:30.
Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–40.
Ohta T, Kimura M. 1971. On the constancy of the evolutionary rate of cistrons. J Mol Evol 1:18–25.
Overington J, Donnelly D, Johnson MS, Šali A, Blundell TL. 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci 1:216–26.
Pakula AA, Young VB, Sauer RT. 1986. Bacteriophage λ *cro* mutations: effects on activity and intracellular degradation. Proc Natl Acad Sci USA 83:8829–33.
Pal C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927–31.

Pal C, Papp B, Hurst LD. 2003. Rate of evolution and gene dispensability. Nature 421:496–7.
R Development Core Team. 2005. R: a language and environment for statistical computing. Vienna, Austria: R foundation for statistical computing.
Shakhnovich BE, Deeds E, Delisi C, Shakhnovich E. 2005. Protein structure and evolutionary history determine sequence space topology. Genome Res 15:385–92.
Shakhnovich EI. 1998. Protein design: a perspective from simple tractable models. Fold Des 3:R45–58.
Sharp P, Li W. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–95.
Shortle D, Lin B. 1985. Genetic analysis of staphylococcal nuclease: identification of three intragenic "global" suppressors of nuclease-minus mutations. Genetics 110:539–55.
Sokal RR, Rohlf FJ. 1994. Biometry. 3rd ed. New York: Freeman.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–80.
Thorne JL, Goldman N, Jones DT. 1996. Combining protein evolution and secondary structure. Mol Biol Evol 13:666–73.
Tiana G, Broglia RA, Shakhnovich EI. 2000. Hiking in the energy landscape in sequence space: a bumpy road to good folders. Proteins 39:244–51.
Tiana G, Shakhnovich BE, Dokholyan NV, Shakhnovich EI. 2004. Imprint of evolution on protein structure. Proc Natl Acad Sci USA 101:2846–51.
Voigt CA, Mayo SL, Arnold FH, Wang ZG. 2001. Computational method to reduce the search space for directed protein evolution. Proc Natl Acad Sci USA 98:3778–83.
Wall DP, Fraser HB, Hirsh AE. 2003. Detecting putative orthologs. Bioinformatics 19:1710–1.
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 102:5483–8.
Wingreen NS, Li H, Tang C. 2004. Designability and thermal stability of protein structures. Polymer 45:699–705.
Wolynes PG. 1996. Symmetry and the energy landscapes of biomolecules. Proc Natl Acad Sci USA 93:14249–55.
Wroe R, Bornberg-Bauer E, Chan HS. 2005. Comparing folding codes in simple heteropolymer models of protein evolutionary landscape: robustness of the superfunnel paradigm. Biophys J 88:118–31.
Yang ZH. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–6.
Zhang J, He X. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol 22:1147–55.
Zhang W, Sun ZB, Zou XW. 2005. Designability of protein structures on the hexagonal lattice model. Chin Phys Lett 22: 2133–6.
Zou JM, Saven JG. 2000. Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. J Mol Biol 296:281–94.
Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. Evolving genes and proteins. New York: Academic Press. p 97–166.

Kenneth Wolfe, Associate Editor

Accepted June 13, 2006