Proportion of Solvent-Exposed Amino Acids in a Protein and Rate of Protein Evolution

Yeong-Shin Lin,† Wei-Lun Hsu,‡ Jenn-Kang Hwang,*‡ and Wen-Hsiung Li†*

*Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan; †Department of Ecology and Evolution, University of Chicago; and ‡Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan

Translational selection, including gene expression, protein abundance, and codon usage bias, has been suggested as the single dominant determinant of protein evolutionary rate in yeast. Here, we show that protein structure is also an important determinant. Buried residues, which are responsible for maintaining protein structure or are located on a stable interaction surface between 2 subunits, are usually under stronger evolutionary constraints than solvent-exposed residues. Our partial correlation analysis shows that, when whole proteins are included, the variance of evolutionary rate explained by the proportion of solvent-exposed residues (P_{exposed}) can reach two-thirds of that explained by translational selection, indicating that P_{exposed} is the most important determinant of protein evolutionary rate next only to translational selection. Our result suggests that proteins with many residues under selective constraint (e.g., maintaining structure or intermolecular interaction) tend to evolve slowly, supporting the "fitness (functional) density" hypothesis.

Introduction

The issue of what factors determine the rate of protein evolution has drawn much attention in recent years (for review, see McInerney 2006; Pal et al. 2006; Rocha 2006). Three major hypotheses have been proposed to explain large variation in protein evolutionary rate. One is that functionally less important proteins evolve faster than more important ones (Ohta 1973; Kimura and Ohta 1974; Wilson et al. 1977). This hypothesis was claimed to be supported by a weak but significant correlation between gene dispensability and protein evolutionary rate (Hirsh and Fraser 2001; Yang et al. 2003; Wall et al. 2005; Zhang and He 2005), but it is still controversial (Pal et al. 2003). The second hypothesis is that the rate is primarily determined by the proportion of residues involved in specific functions, that is, the "functional density" hypothesis (Zuckerkandl 1976) or the "fitness density" hypothesis (Drummond et al. 2005; Pal et al. 2006). In this vein, Fraser et al. (2002) claimed that proteins with more interaction partners evolve more slowly because they have a higher functional density. However, this claim was questioned because the level of gene expression was not controlled (Bloom and Adami 2003), but gene expression level has been found to be a major determinant of protein evolutionary rate (Pal et al. 2001; Rocha and Danchin 2004; Wall et al. 2005). Recently, a third hypothesis was proposed by Drummond et al. (2006) that translational selection is the single dominant determinant, that is, the number of translation events a gene experiences determines its evolutionary rate. An explanation for why gene or protein expression level governs the evolutionary rate was provided by Drummond et al. (2005).

Here, instead of considering the protein as a whole as in the studies reviewed above, we look into differences in evolutionary constraints among residues to examine the fitness (functional) density hypothesis. Dickerson (1971) found that surface residues that interact with other proteins tend to be highly conserved. Later, Kimura and Ohta (1973) found that the rate of amino acid substitution at surface residues of the α and β globins evolve 10 times faster than residues in the heme pocket. Similarly, it has been found that residues in the interfaces of obligate protein complexes are more conserved than residues in transient interactions (Mintseris and Weng 2005) and that the solvent-inaccessible core of a protein is better conserved than solvent-accessible residues in a protein (Overington et al. 1992; Goldman et al. 1998; Bustamante et al. 2000). Moreover, residues in the buried core and residues on the solvent-exposed surfaces were shown to have different substitution patterns due to different selection pressures (Tseng and Liang 2006). From these findings, it is reasonable to speculate that a protein with a small proportion of solvent-exposed residues (P_{exposed}) should evolve slowly. However, a contradictory result was found recently (Bloom, Drummond, et al. 2006). It is therefore interesting to investigate whether the structure of a protein, especially the solvent accessibility of the residues, is an important determinant of protein evolutionary rate.

Materials and Methods

Genomic Data

We studied genes in the *Saccharomyces cerevisiae* genome and obtained nonsynonymous rates (K_A) from Wall et al. (2005), protein interaction modules from Han et al. (2004), mRNA expression level data from Holstege et al. (1998), protein abundance data from Ghaemmaghami et al. (2003), and codon adaptation index (CAI) values from Drummond et al. (2006); CAI indicates the strength of codon usage bias (Sharp and Li 1987). We obtained protein subunit data and gene dispensability data following Lin et al. (2007). Those open reading frames (ORF) without gene names were excluded. Principal component regression was performed using R with the package "pls" (Ihaka and Gentleman 1996). Protein abundance and mRNA expression level were log transformed.

Solvent Accessibility Prediction Using the Homology Model

Although using the three-dimensional (3D) structures of yeast proteins to estimate the proportion of exposed amino acids in a protein (P_{exposed}) is the best choice, completely determined 3D structures are available for only about 100 yeast proteins. For yeast proteins without 3D

Key words: evolutionary rate, protein structure, fitness density, functional density, solvent accessibility, disordered.

E-mail: whli@uchicago.edu.

Mol. Biol. Evol. 24(4):1005–1011. 2007
doi:10.1093/molbev/msm019
Advance Access publication January 29, 2007

structure, we have therefore used the (Protein Data Bank) homologues for yeast ORFs in the *Saccharomyces* Genome Database (SGD; http://www.yeastgenome.org/). The PDB homologues are protein structures from various species (including *S. cerevisiae*) homologous to yeast ORFs, and we used them to estimate P_{exposed}, assuming that the 3D structures for the homologues are identical.

For each yeast ORF, the PDB homologue with the lowest divergence to the yeast ORF sequence was chosen. The solvent accessible surface areas (ACCs) for each residue of the PDB homologue were obtained from DSSP (database of secondary structure assignments for all protein entries in the PDB; http://swift.cmbi.ru.nl/gv/dssp/) (Kabsch and Sander 1983). Both the core residues of a protein, which are important in maintaining protein structure, and the residues on the stable interaction surface between 2 subunits can be regarded as buried because in the native 3D structure of a protein complex they are indeed not solvent exposed. Therefore, for protein structures including more than 1 chain, the interchain contacts were included to calculate the ACC values for these residues. Relative solvent accessibility (RelACC) was the ACC value for each residue divided by the maximum value of ACC for the amino acid (represented in percentage), which is estimated from a Gly-X-Gly extended tripeptide conformation. We define residues with RelACC higher than 25% as exposed residues and the others as buried. The exposed and buried residues defined using this threshold have approximately equal numbers. The P_{exposed} value for each PDB homologue was thus calculated. We have also tried 2 other RelACC values (0% and 50%) as the threshold, but the conclusions were essentially the same.

Note that there are a number of problems with the PDB data. First, even for a yeast protein with the 3D structure completely determined, P_{exposed} is not known but must be estimated. Second, for a yeast protein without the 3D structure, we have to use the structure of a protein homologous to the yeast protein. Moreover, PDB homologues are not available for many yeast ORFs. Third, the structural similarity between the yeast protein and its distant PDB homologue may hold only for the well-folded, conserved domains, but not for the other regions. Fourth, many PDB structures are only partially determined, and most of them are restricted to the well-folded regions in the proteins. Only the structures of well-folded proteins may be completely determined. The alignment length between the PDB homologue and the yeast ORF sequence can approximately represent the proportion of the structure determined. The unaligned regions are either not structurally determined or too diverged between the yeast protein and its PDB homologue. They are often disordered regions in 3D structures. Fifth, some PDB structures only include one or a few subunits, not the entire protein complex. In this case, residues on the stable interaction surface, which should be buried in vivo, are mistakenly treated as exposed residues in the PDB structure data. Therefore, P_{exposed} estimated from PDB homologues is only applicable for a limited number of proteins. To overcome these problems, we also used support vector machine (SVM) to predict P_{exposed} for each ORF directly from the amino acid sequence.

Solvent Accessibility Prediction Using SVM

We used the same training data set as Kim and Park (2004), which includes 480 proteins all with known 3D structures and with less than 25% sequence similarity between sequences. The ACC for each residue of these 480 proteins were obtained from DSSP. Residues on the stable interaction surface between 2 subunits were also regarded as buried residues. We used position-specific scoring matrices (PSSM), secondary structure profiles, and hydropathy indexes (Kyte and Doolittle 1982) as feature factors. A 15-amino acid–sliding window was used to represent the local environment of the protein sequences. We used 5iterations of PSI-Blast (Altschul et al. 1997) against the nonredundant protein sequence database to produce PSSM. The secondary structure profiles describing the occurring probabilities for helix, sheet, and coil were generated using the PSIPRED secondary structure prediction method (Jones 1999). SVM prediction was performed using a library for SVM version 2.6 (Chang and Lin 2001). A 7-fold cross validation test yields 78% accuracy. Note that P_{exposed} (PDB) is only calculated for the determined, well-folded regions, whereas P_{exposed} (SVM) can be estimated for the whole protein.

Results and Discussion

Proportion of Exposed Residues in a Protein

To investigate how well P_{exposed} (SVM) represents the proportion of exposed residues, we compared it with P_{exposed} (PDB) for ORFs with their PDB homologues for the cases with alignment length >98% and sequence identity >98% (i.e., completely determined PDB structures from *S. cerevisiae*, data set I in table 1). Because the buried/exposed state of residues predicted by SVM have 22% inaccuracy, proteins with high (low) P_{exposed} would tend to have their P_{exposed} (SVM) underpredicted (overpredicted); in other words, the predicted values would have a smaller variance. We indeed found 9/9 ORFs with P_{exposed} (PDB) >0.6 have their P_{exposed} (SVM) slightly underpredicted, whereas 30/31 ORFs with P_{exposed} (PDB) <0.4 have their P_{exposed} (SVM) slightly overpredicted. Compared with P_{exposed} (PDB), P_{exposed} (SVM) has, as expected, a slightly higher mean (closer to 0.5) and a smaller variance (table 1). The correlation coefficient between P_{exposed} (PDB) and P_{exposed} (SVM) is high but not perfect ($R = 0.72$, $n = 96$, $P = 1.3 \times 10^{-16}$). The reason might be that most ORFs (84/96) have their P_{exposed} (PDB) between 0.3 and 0.6, and not evenly distributed from 0 to 1; the noise introduced by SVM may therefore likely disturb the correlation. Nevertheless, the correlation between P_{exposed} (SVM) and K_A is very similar to the correlation between P_{exposed} (PDB) and K_A (0.450 vs. 0.465, first row in table 1). This result suggests that P_{exposed} (SVM) can be used to estimate the correlation between P_{exposed} and K_A.

All proteins in data set I have their 3D structures well determined and this fact implies that they are all well-folded proteins. Structurally less well-determined proteins usually contain disordered regions, which contain mainly exposed residues and have been found to evolve rapidly (Brown et al. 1992). To study this problem, we subdivided our data set by the alignment length between the PDB homologue

Table 1
Comparison between P_{exposed} (PDB) Estimated from PDB Homologues and P_{exposed} (SVM) Estimated from SVM Predictions

			P_{exposed} (PDB)		P_{exposed} (SVM)	
The alignment length, l	Number of Proteins Used (n)	Correlation between P_{exposed} (PDB) and P_{exposed} (SVM)	Mean ± Standard Deviation	Correlation with K_A	Mean ± Standard Deviation	Correlation with K_A
Data set I	96	0.721***	0.450 ± 0.109	−0.093	0.465 ± 0.072	−0.106
$l > 98\%$	151	0.665***	0.446 ± 0.108	−0.049	0.456 ± 0.076	−0.066
$98\% \geq l > 90\%$	146	0.404**	0.436 ± 0.094	0.006	0.431 ± 0.073	−0.013
$90\% \geq l > 70\%$	225	0.229*	0.452 ± 0.091	−0.089	0.462 ± 0.090	0.177*
$70\% \geq l > 50\%$	205	0.292*	0.468 ± 0.089	−0.022	0.510 ± 0.110	0.201*
$50\% \geq l > 30\%$	218	—	—	—	0.561 ± 0.104	0.266*
$30\% \geq l$	218	—	—	—	0.611 ± 0.142	0.411***

NOTE.—The length of alignment is between the ORF sequence and the PDB homologue. Data set I: alignment length >98%, and sequence identity >98%. P_{exposed} (PDB) is only represented for proteins with the alignment length >50%, and omitted for others (—).

*$P < 0.05$; **$P < 10^{-6}$; ***$P < 10^{-9}$.

and the ORF sequence (table 1). The alignment length reflects the extent that a protein structure has been determined and, to some degree, reflects how a protein is folded. Because disordered regions are not determined in 3D structure and usually are solvent exposed, P_{exposed} likely increases with the disordered regions in a protein. The positive correlation between P_{exposed} (SVM) and the proportion of unaligned regions ($R = 0.52$, $n = 1163$, $P = 6.9 \times 10^{-81}$) supports this view, even though, as noted above, SVM prediction tends to underestimate P_{exposed} for proteins with a high P_{exposed}. In contrast, there is no significant change for P_{exposed} (PDB) when the alignment length threshold decreases (table 1), probably because only the well-folded cores are represented for these PDB homologues, whereas disordered regions are not. We also found a positive correlation between K_A and the proportion of unaligned regions ($R = 0.38$, $n = 1163$, $P = 3.6 \times 10^{-42}$), consistent with the finding that disordered regions evolve fast (Brown et al. 1992). Our result suggests that the disordered regions in a protein may largely determine its P_{exposed} and evolutionary rate. Therefore, P_{exposed} (PDB) is only applicable for structurally well-determined, well-folded proteins, whereas P_{exposed} (SVM) is suitable for ORFs in general, especially for proteins with disordered regions.

P_{exposed} and Rate of Protein Evolution

England and Shakhnovich (2003) suggested that proteins with a higher contact density (fewer exposed residues) are more designable (a protein structure encoded by many sequences). Bloom, Drummond, et al. (2006) proposed that proteins with higher designable structures evolve more rapidly. They stated, "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 K_A for the exposed residues." Therefore, the reduction in P_{exposed} "is more than compensated for by the increased variability of exposed residues in proteins with high contact density." Interestingly, we found that when the threshold of the alignment length is high, that is, for proteins with fewer disordered residues, P_{exposed} is negatively correlated with K_A (estimated either by PDB homologues or SVM prediction; table 1), which is consistent with the observation of Bloom, Drummond, et al. (2006). They also used a stringent criterion to restrict their data set, that is, the number of identities in the total length of the alignment is >80%. However, we found that P_{exposed} (SVM) is positively correlated with K_A when the alignment threshold is decreased, especially when proteins with many disordered regions are included (table 1).

We next noted that the negative correlation between P_{exposed} and K_A found in this study (e.g., in data set I) is not as strong as that in Bloom, Drummond, et al. (2006). The major reason is that they did not consider interchain (intersubunit) contacts, whereas we did. Bloom, Drummond, et al. (2006) found that proteins with a smaller contact density evolve much slower at their exposed sites. We analyzed 100 proteins in their data set, for which the complex annotations are available (Lin et al. 2007). We found that all the 11 proteins with a contact density <6 are heterocomplexes and 9 out of these 11 proteins have at least 7 subunits. In contrast, only 26 out of the 89 proteins with a contact density >6 were annotated to have 7 or more subunits. We also noted that the number of complex subunits (k) is negatively correlated with K_A in the data set of Bloom, Drummond, et al. (2006) ($R = -0.32$, $n = 62$, $P = 1.2 \times 10^{-2}$), which is consistent with Teichmann's (2002) finding that stable complex proteins evolve more slowly. Therefore, we suggest that for these well-folded proteins, selection pressure on residues at the interchain interaction sites is as important as designability (inferred from contact density) for determining the evolutionary rate.

Note that the correlation between P_{exposed} and evolutionary rate may reflect 2 contradictory effects, that is, fitness (functional) density and designability. The switch from negative to positive correlations between P_{exposed} (SVM) and K_A as P_{exposed} (SVM) increases indicates that the effect of designability (inferred from contact density) on evolutionary rate (Bloom, Drummond, et al. 2006) might be restricted to the well-folded proteins (or cores). This also explains the slightly negative correlation between P_{exposed} (PDB) and K_A when the alignment length is not long (table 1), that is, only the designability of the well-folded cores is inferred by P_{exposed} (PDB). The significant positive correlation between P_{exposed} (SVM) and K_A for proteins containing large disordered regions (table 1) suggests that for these

Table 2
Partial Correlation Analyses between one of 5 Variables and K_A

Variable	Number of Proteins Used (n)	Correlation with K_A	Partial Correlation with K_A		
			R^2	P	The Factor Controlled
P_{exposed} (SVM)	2153	0.376	10.9%	$<10^{-15}$	mRNA expression
P_{exposed} (SVM)	1602	0.399	12.5%	$<10^{-15}$	Protein abundance
P_{exposed} (SVM)	2267	0.373	13.2%	$<10^{-15}$	CAI
mRNA expression	2153	−0.452	17.4%	$<10^{-15}$	P_{exposed} (SVM)
Protein abundance	1602	−0.416	13.9%	$<10^{-15}$	P_{exposed} (SVM)
CAI	2267	−0.353	11.8%	$<10^{-15}$	P_{exposed} (SVM)
P_{exposed} (SVM)	1568	0.397	11.3%	$<10^{-15}$	Translational selection
Translational selection	1568	−0.473	18.3%	$<10^{-15}$	P_{exposed} (SVM)

NOTE.—Translational selection is represented as the first component in the principal component analysis for mRNA expression, protein abundance, and CAI values.

proteins, fitness (functional) density can explain the variance of evolutionary rate much better than designability.

When all proteins are included, partial correlation analysis shows that P_{exposed} (SVM) still significantly positively correlates with evolutionary rate even when the translational selection predictors (mRNA expression, protein abundance, and codon usage bias measured by CAI) are controlled (table 2). The variance of evolutionary rates explained by P_{exposed} (SVM) is more than half of that explained by mRNA expression or protein abundance, and even slightly more than that explained by CAI. This result suggests that in general, fitness (functional) density has much higher impact than designability on protein evolutionary rate. It is likely that for well-folded proteins the variances of P_{exposed} and evolutionary rate are small, so that the differences in selection pressure between exposed and buried residues are almost compensated by the effect of designability in these proteins, but this is not true for other proteins (fig. 1).

Principal Component Regression Analysis

Note that although partial correlation analysis may be unreliable when data are noisy and the correlation is weak (Drummond et al. 2006), the correlation we found is not weak ($R^2 > 10\%$, table 2). On the other hand, the results of principal component regression (PCR) analysis can be misleading due to 2 problems (see Box 1). We use our data in table 3 as an example. Three translational selection–related predictors are used to perform the analysis, so they contribute to more than 90% of CP_1 (the first component), which explains 26.5% of the variance of K_A, whereas P_{exposed} (the major component of CP_2) seems to explain only ~5% of the variance. This inference is misleading because the 3 translational selection–related predictors are not mutually independent, and this decides the order of the components. To demonstrate this problem, let us consider only 1 translational selection–related predictor, say CAI, and the 2 P_{exposed} values obtained from PDB homologues and SVM prediction. Now the 2 P_{exposed} variables contribute to ~90% of CP_1, whereas CAI contributes only 11% (table 4). CP_1 explains almost 20% of the variance of K_A. We cannot conclude that all this 20% variance is contributed by P_{exposed} because CAI is not really controlled. This example shows that PCR analysis tends to overestimate the contribution of correlated predictors to the variation of a response variable but underestimate the contributions of other predictors. In the presence of nonindependent factors, if the purpose is to obtain 1 representative component and to see the correlation between this component and the response variable, PCR analysis is very useful. However, if the purpose is to see the correlation between one factor and the response variable whereas controlling other factors, PCR analysis can be misleading.

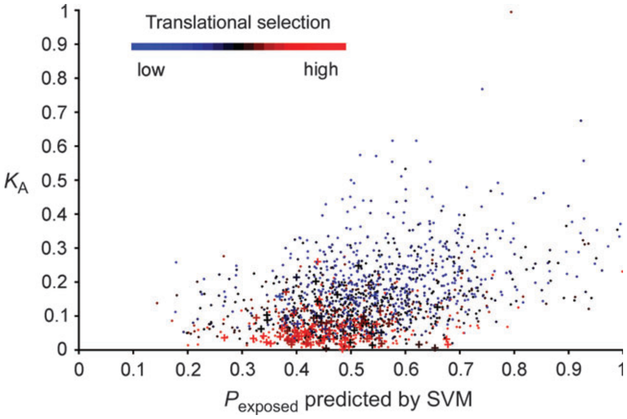

FIG. 1.—A positive correlation between nonsynonymous substitution rate (K_A) and P_{exposed} predicted by SVM. Proteins with their alignment length between the PDB homologue and the ORF sequence larger than 98% are indicated by crosses, whereas others are represented by circles. Red indicates highly expressed genes, whereas blue indicates lowly expressed ones. The strength of translational selection is based on the first component in the principal component analysis for mRNA expression, protein abundance, and CAI values.

Contribution of P_{exposed} to Variation in Rate of Protein Evolution

To conduct the analysis more appropriately, we used suitable controls. We defined CP_1 in the principal component analysis (PCA) for the 3 predictors, mRNA expression, protein abundance, and CAI, as translational selection, and we controlled it to calculate partial correlation between P_{exposed} (SVM) and K_A. (Because mRNA expression, protein abundance, and CAI together represent translational selection + noise and because CP_1 is the best variable to represent these 3 predictors at the same time, controlling translational selection would be better than controlling the 3 predictors individually). As seen near the bottom of table 2, the contribution of P_{exposed} (SVM) to the variance of K_A is 11.3% when the translational selection is

Table 3
Results of Principal Component Regression Analysis of mRNA Expression, Protein Abundance, CAI, and P_{exposed} (SVM) on K_A for 1,568 Genes

	Principal Components			
	CP$_1$	CP$_2$	CP$_3$	CP$_4$
Percent variance in predictors	58.3	23.0	10.0	8.7
Percent variance explained (R^2) in K_A	26.5***	4.9***	0.1	0.1
Percent contributions of a predictor				
mRNA expression	**32.6**	0.4	6.3	**60.7**
Protein abundance	**30.8**	1.4	**65.4**	2.4
CAI	**30.4**	6.1	**28.0**	**35.5**
P_{exposed} (SVM)	6.2	**92.1**	0.3	1.4

NOTE.—Boldface indicates that the indicated predictor contributes at least 20% to the indicated component.
***$P < 10^{-9}$.

Table 4
Results of Principal Component Regression Analysis of 2 P_{exposed} Values (from SVM Prediction and PDB Homologues), Gene Dispensability, Protein Length, and Translational Selection–Related Predictor (CAI) on K_A for 1,163 Genes

	Principal Components				
	CP$_1$	CP$_2$	CP$_3$	CP$_4$	CP$_5$
Percent variance in predictors	30.1	22.7	21.1	16.6	9.5
Percent variance explained (R^2) in K_A	19.3***	12.1***	0.6	0.0	1.7*
Percent contributions of a predictor					
P_{exposed} (SVM)	**46.2**	0.6	6.9	1.6	**44.8**
P_{exposed} (PDB)	**41.9**	10.4	3.0	2.0	**42.7**
Dispensability	0.7	**30.6**	**33.1**	**34.7**	0.8
Protein length	0.4	**28.6**	**53.2**	6.4	11.4
CAI	10.7	**29.8**	3.9	**55.4**	0.3

NOTE.—Boldface indicates that the indicated predictor contributes at least 20% to the indicated component. The observation that P_{exposed} (SVM) and P_{exposed} (PDB) contribute almost equally to the first component does not imply that they contribute equally to K_A. They contribute equally to CP$_1$ because in this way CP$_1$ can include maximal information (see Box 1).
*$P < 0.05$; ***$P < 10^{-9}$.

controlled, which is about two-thirds of the contribution by translational selection (18.3%) to the variance of K_A when P_{exposed} (SVM) is controlled. This analysis suggests that P_{exposed} contributes ~10% to variation in K_A and is the most important known determinant next to translational selection. Note that this might be an underestimate because of the considerable uncertainty involved in the estimation of P_{exposed}.

Fraser (2005) showed that party hubs (proteins interacting with most of their partners simultaneously; Han et al. 2004) evolve slower than date hubs (proteins interacting with different partners at different times). Because most party hubs are protein complexes, whereas date hubs are not (Han et al. 2004), we compared P_{exposed} (SVM) values between them. Not surprisingly, party hubs, on average, have a smaller P_{exposed} (49.7%) compared with date hubs (56.9%, *t*-test $P = 5.0 \times 10^{-5}$). It is therefore reasonable to speculate that P_{exposed} should also explain part of the difference in evolutionary rate between party and date hubs. Similarly, subunits of a large heterocomplex should have more protein interactions and should be less dispensable (Lin et al. 2007). P_{exposed} may therefore underlie the correlations between these 2 factors and evolutionary rate (Hirsh and Fraser 2001; Fraser et al. 2002; Yang et al. 2003; Wall et al. 2005; Zhang and He 2005).

It is worth noting that proteins with a high P_{exposed} may evolve slowly or fast, whereas proteins with a low P_{exposed} almost always have a low evolutionary rate (fig. 1). This result suggests that protein 3D structure provides only a general index, that is, buried resides cannot evolve freely. Some exposed residues (e.g., residues at active sites or ligand-binding sites) may be functionally important and are thus conserved. Protein mutagenesis experiments have shown that increasing a protein's thermodynamic stability dramatically increases its tolerance to mutations that suggests deleterious mutations usually act by hindering the formation of a properly folded protein rather than altering a protein's function (Bloom et al. 2005; Bloom, Labthavikul, et al. 2006). The evolutionary constraint of highly expressed proteins was suggested to reduce the burden of protein misfolding (Drummond et al. 2005). Similarly, it is likely that buried residues are conserved because they are important to make proteins fold or interact correctly among subunits or proteins. Although translational selection largely governs the rate of evolution for the whole protein (Drummond et al. 2006), our study shows that fitness (functional) density negatively correlates with protein evolutionary rate, that is, a protein with more residues under selective constraint tends to evolve more slowly. We expect that an even better correlation will be found when the fitness (functional) density can be appropriately defined rather than estimated as buried residues.

Box 1

PCA transforms *n* factors (which may not be mutually independent) to *n* independent components by rotating the axes such that the first component has the largest variance by any projection of the data, and the second component has the second largest variance, and so on. Given a data set $\{x_1, x_2, x_3\}$, we can obtain the first component, $\text{CP}_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3$, where a_{11}^2, a_{12}^2, and a_{13}^2 indicate the contributions of x_1, x_2, and x_3 to CP$_1$ and they are summed to 1. We can then use CP$_1$ to correlate with a response, *y*, and calculate R^2, the variance of *y* explained by CP$_1$.

However, although a_{11}^2 indicates the contributions of x_1 to CP$_1$, using $a_{11}^2 \times R^2$ to represent the variance of *y* explained by x_1 is invalid. This problem can be demonstrated by a simple example with a data set $\{x_1, x_2\}$, where x_1 and x_2 have the same variance and are correlated. We can then obtain $\text{CP}_1 = a_{11}x_1 + a_{12}x_2$ and $\text{CP}_2 = a_{21}x_1 + a_{22}x_2$, where $a_{11} = a_{12}$ and $a_{21} = -a_{22}$, so that x_1 and x_2 contribute equally to both $\text{CP}_1(a_{11}^2 = a_{12}^2)$ and $\text{CP}_2(a_{21}^2 = a_{22}^2)$. We can thus correlate CP$_1$ and CP$_2$ with *y* and calculate the variance of *y* explained by CP$_1$ and CP$_2$, respectively. When $y = x_1$, it is obvious that x_1 can completely explain the variance of *y*, whereas only a proportion of the variance of *y* can be explained by x_2 (x_1 and x_2 are correlated). However, the fact that the variances of *y* explained by x_1 and x_2 are different cannot be revealed by PCR analysis (x_1 and x_2 contribute equally to both CP$_1$ and CP$_2$, i.e., $a_{11}^2 = a_{12}^2$ and $a_{21}^2 = a_{22}^2$.)

The second problem is that, when the inputs include many nonindependent factors, the first component can be highly correlated to these factors, so that it can include as much information as possible. After the first component is decided, the second component is determined by including as much of the remaining information as possible. Given a data set $\{x_1, x_2, x_3, x_4\}$ where the variables in the subset $\{x_1, x_2, x_3\}$ are highly correlated with each other, CP_1 will be mainly composed of x_1, x_2 and x_3. If the subset $\{x_1, x_2, x_3\}$ has covariance with x_4, this covariance will therefore be mainly included in CP_1 but not other components. In this case, x_4 contributes mainly to CP_2 but also weakly to CP_1. For CP_1, CP_2 is actually controlled because CP_1 and CP_2 are independent to each other. However, we cannot say that for CP_1, factor x_4 is controlled because CP_1 includes the covariance shared between the subset $\{x_1, x_2, x_3\}$ and x_4.

Acknowledgments

We thank Eduardo P.C. Rocha, Martin Lercher, Kevin Bullaughey, and Jeffrey Tseng for comments and the Structural Bioinformatics Core at the National Chiao Tung University for hardware and software support. The work was supported by grants from National Science Council (NSC 094-2917-I-009-015 to Y.-S.L. and NSC 093-3112-B-009-001 to J.-K.H.) and National Institutes of Health grants to W.-H.L.

Literature Cited

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

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, Labthavikul ST, Otey CR, Arnold FH. 2006. Protein stability promotes evolvability. Proc Natl Acad Sci USA. 103:5869–5874.

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.

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 1992. Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol. 55:104–110.

Bustamante CD, Townsend JP, Hartl DL. 2000. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. Mol Biol Evol. 17:301–308.

Chang C-C, Lin C-J. 2001. Training υ-support vector classifiers: theory and algorithms. Neural Comput. 13:2119–2147.

Dickerson RE. 1971. The structures of cytochrome *c* and the rates of molecular evolution. J Mol Evol. 1:26–45.

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.

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

Fraser HB. 2005. Modularity and evolutionary constraint on proteins. Nat 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.

Ghaemmaghami S, Huh W-K, Bower K, Howson RW, A Belle, Dephoure N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. Nature. 425:737–741.

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

Han J-DJ, Bertin N, Hao T, et al. (11 co-authors). 2004. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 430:88–93.

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

Holstege FCP, Jennings EG, 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.

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

Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 292: 195–202.

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

Kim H, Park H. 2004. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. Proteins. 54:557–562.

Kimura M, Ohta T. 1973. Mutation and evolution at the molecular level. Genetics. 73 (Suppl.):19–35.

Kimura M, Ohta T. 1974. On some principles governing molecular evolution. Proc Natl Acad Sci USA. 71:2848–2852.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 157:105–132.

Lin Y-S, Hwang J-K, Li W-H. 2007. Protein complexity, gene duplicability and gene dispensability in the yeast genome. Gene. 387:109–117.

McInerney JO. 2006. The causes of protein evolutionary rate variation. Trends Ecol Evol. 21:230–232.

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

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature. 246:96–98.

Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci. 1:216–226.

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.

Rocha EPC. 2006. The quest for the universals of protein evolution. Trends Genet. 22:412–416.

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

Sharp PM, Li W-H. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.

Teichmann SA. 2002. The constraints protein–protein interactions place on sequence divergence. J Mol Biol. 324:399–407.

Tseng YY, Liang J. 2006. Regions and application in protein function inference: a Bayesian Monte Carlo approach. Mol Biol Evol. 23:421–436.

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.

Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. Annu Rev Biochem. 46:573–639.

Yang J, Gu Z, Li W-H. 2003. Rate of protein evolution versus fitness effect of gene deletion. Mol Biol Evol. 20:772–774.

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

Zuckerkandl E. 1976. Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. J Mol Evol. 7:167–183.

William Martin, Associate Editor

Accepted January 22, 2007