Solvent Exposure Imparts Similar Selective Pressures across a Range of Yeast Proteins

Gavin C. Conant and Peter F. Stadler†‡§||*

*Division of Animal Sciences and Informatics Institute, University of Missouri-Columbia; †Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany; ‡RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany; §Institute for Theoretical Chemistry, University of Vienna, Vienna, Austria; and ||Santa Fe Institute

We study how an amino acid residue's solvent exposure influences its propensity for substitution by analyzing multiple alignments of 61 yeast genes for which the crystal structure is known. We find that the selective constraint on the interior residues is on average 10 times that of residues on the surface. Surprisingly, there is no correlation between the overall selective constraint observed for a protein alignment and the ratio of constraints on interior and surface residues. By modeling the selective constraint on several amino acid properties, we show that although residue volume and hydropathy are strongly conserved across most alignments, there is little variation in interior versus surface conservation for these two properties. By contrast, residue charge (isoelectric point) is less generally conserved when considering the protein as a whole but shows a strong constraint against the introduction of charged residues into the protein interior.

Introduction

Site-to-site variation in substitution rates in protein-coding genes has been both a technical challenge (e.g., in phylogenetic inference; Reeves 1992; Sidow and Steel 1992; Yang 1993) and a source of biological insight (e.g., when site-to-site variation in the frequency of nonsynonymous substitutions is used to identify locations of positive selection; Nielsen and Yang 1998; Yang, Nielsen et al. 2000).

One useful way to think of such site-to-site variation is in terms of differing selective constraints (Yang, Swanson, and Vacquier 2000) and the underlying biochemical and biophysical sources of these differences. An obvious source of this variation is a residue's location in the protein's 3D structure. Both an amino acid residue's exposure to solvent and its secondary structure environment are known to affect its substitution rate (Thorne et al. 1996; Goldman et al. 1998; Bustamante et al. 2000; Mintseris and Weng 2005; Bloom et al. 2006). Higher solvent accessibility appears to not only be associated with higher substitution rates (Goldman et al. 1998) but also specifically with reduced selective constraint on nonsynonymous substitutions (Bustamante et al. 2000; Bloom et al. 2006).

In a previous analysis, we studied how the range of permissible amino acid substitutions varied among different types of proteins (Conant et al. 2007). As others had also found (Tourasse and Li 2000), we concluded that proteins differed in their relative frequencies of amino acid substitutions. Here, we extend this analysis to determine whether there are significant differences in substitution patterns among sites within the same protein due to their varying exposures to the solvent. We examine both the difference in overall selective constraint between interior and surface residues and whether the physical and chemical properties of the amino acid residues can help to explain this variation in substitution rates due to solvent exposure.

Key words: amino acid substitution, evolutionary models, protein structure.

E-mail: conantg@missouri.edu.

Mol. Biol. Evol. 26(5):1155–1161. 2009
doi:10.1093/molbev/msp031
Advance Access publication February 20, 2009

Methods

Protein Structures

The Protein Databank (PDB; Berman et al. 2000) was queried for sequences derived from the bakers' yeast *Saccharomyces cerevisiae*. The sequences from the resulting structures were then compared with the yeast genome (Goffeau et al. 1996), and exact substring matches (allowing for gaps in the determined structure) were retained. The exposed surface area of each residue was inferred using Analytic Surface Calculation (ASC; Eisenhaber and Argos 1993; Eisenhaber et al. 1995). The proportion of total surface exposed (x) was calculated by dividing the above values by the surface area of the residue in question (measured in a Gly-X-Gly chain; Chothia 1976).

Sequence Data

Sequence data for homologous genes from eight genomes related to *S. cerevisiae* were obtained from the Yeast Genome Order Browser (YGOB; Byrne and Wolfe 2005). Data from five other genomes were also added: *Candida albicans* (downloaded from the Candida Genome database; Jones et al. 2004; Arnaud et al. 2005), *Debaryomyces hansenii* (Dujon et al. 2004), *Saccharomyces mikatae* (Kellis et al. 2003), *Saccharomyces paradoxus* (Kellis et al. 2003), and *Yarrowia lipolytica* (Dujon et al. 2004). Because of the presence of a whole-genome duplication (WGD) in *S. cerevisiae* (Wolfe and Shields 1997; Dietrich et al. 2004; Kellis et al. 2004), it is overly restrictive to require reciprocal best Blast hits to identify orthologs in these five genomes. Instead, we slightly relax our criterion in cases where *S. cerevisiae* has a duplicate from WGD (identified using YOGB; Byrne and Wolfe 2005) and identify an ortholog as the gene giving the best hit in a given genome, which itself hits either the *S. cerevisiae* gene in question or that gene's paralog from the WGD.

Multiple sequence alignments were generated using T-Coffee (Notredame et al. 2000) and visually inspected. After eliminating putative orthologs with poor alignment, we removed alignments in which less than seven homologous sequences were retained. The product of these filtering steps was the set of 61 structure–alignment pairs analyzed below. Nucleotide alignments were inferred from protein

sequence alignments and gaps excluded from further analysis.

The phylogenetic relationships of the yeast species studied here are reasonably well understood (Kurtzman and Robnett 2003; Diezmann et al. 2004). However, the relationship between the phylogeny of a given gene and the species phylogeny becomes complicated in the presence of WGD (Scannell et al. 2006, 2007; Conant and Wolfe 2008). For this reason, we chose to individually infer a phylogeny for each of the nucleotide alignments studied. These phylogenies were inferred by maximum likelihood using PAUP* 4.0b10 (Swofford 2002) under the Hasegawa, Kishino, Yano model (Hasegawa et al. 1985) with rate heterogeneity among sites modeled as following a discrete gamma distribution with four categories (Yang 1993, 1994). Both the gamma shape parameter α and the transition–transversion rate ratio parameter κ were estimated by the built-in numerical maximum likelihood optimization method of PAUP*.

To ascertain whether this approach to inferring gene phylogenies was likely to bias our analysis, we selected 14 alignments that contain no genes with surviving paralogs from the WGD. The WGD can cause differences between gene and species phylogenies even in this case, but it is less likely (though see Scannell et al. 2007). We repeated our analyses using the presumed species tree (Kurtzman and Robnett 2003; Diezmann et al. 2004). As can be seen in supplemental figures 1 and 2, Supplementary Material online, our conclusions are generally insensitive to the topology used.

Inferences Using Models of Codon Evolution

We fit the above alignments to three previously described models of codon evolution, the MG/GY94 model, the Linear Combination of Amino acid Properties (LCAP) model, and the Similarity Groups (SG) model. The first model was developed by Goldman and Yang (1994) and is similar to that of Muse and Gaut (1994) (see Conant et al. 2007 for discussion). The LCAP model makes nonsynonymous substitution rates dependent on weighted combinations of the differences between the two residues in question for five physical and chemical properties (Conant et al. 2007). Thus, the instantaneous substitution rate for nonsynonymous codons A and B is given by:

$$R_{A,B} = C \cdot \exp\left[\sum_{j=1}^{5} \alpha_j \cdot \Delta j_{A,B}\right] \cdot R_{\text{nucleotide}}(A,B). \quad (1)$$

Here $\Delta j_{A,B}$ gives the difference in the jth property between the amino acids in question, and $R_{\text{nucleotide}}$ gives the rate of the nucleotide substitution in question. Values for the α_js are estimated by maximum likelihood.

As discussed previously (Conant et al. 2007), the five properties of the LCAP model are not strictly independent. However, we can test for independent predictive power by excluding a given property and asking whether a likelihood ratio test (LRT, see below) indicates that the model fit has been significantly worsened. For example, although polarity and hydropathy are correlated (Conant et al. 2007), removal of hydropathy from the model while polarity remains generally worsens the fit of the LCAP model to these data, but the converse is not true (table 1). Thus, here we use three amino acid properties (volume, isoelectric point, and hydropathy). For most alignments considered, the removal of any one of these three properties significantly worsens the quality of the model fit (table 1).

The SG model is our parameterization (Conant et al. 2007) of a general class of models that divide the amino acid residues into classes with differing substitution rates between them (Sainudiin et al. 2005; Wong et al. 2006). We discuss the details of this model in the Results section.

Effects of Solvent Exposure

To study the effects of solvent exposure on evolutionary estimates, we implemented versions of all three models with independent values of the nonsynonymous substitutions parameters (ω for the MG/GY94 model, I_w and I_b for the SG model, and α_js and C for the LCAP model) for the surface and the interior residues. Sites in a sequence follow the first value with a probability that equals the relative amino acid exposure x described above and follow the other value with probability $1 - x$. Codons not included in the structural data were excluded.

Model Significance Tests

Nested models of evolution were compared using likelihood ratio tests (LRTs), assuming that twice the difference in ln-likelihood ($2\Delta \ln L$) follows a chi-square distribution with the degrees of freedom being equal to the number of extra parameters in the alterative model (Sokal and Rohlf 1995).

For example, to test if the α_js for the interior and surface residues ($\alpha_j^i$and α_j^s, respectively) in the LCAP model were significantly different, we compared an eight-parameter model where $\alpha_j^i \neq \alpha_j^s$ for all j to a seven-parameter model where one property k was given the same surface and interior constraint ($\alpha_k^i = \alpha_k^s$). The significance of the improvement in likelihood ($2\Delta \ln L_{\text{struct-prop}}$) for the eight-parameter model was inferred with an LRT (see table 1 for full details of the LRTs conducted).

Results

Comparison of Proteins with Different Relative Exposures

One method to detect differences in selective constraints between interior and surface residues is to calculate the correlation between the whole-gene average selective constraint ω_g and a gene's average residue exposure to the solvent E. Note that ω_g is the maximum likelihood estimate of the ratio of nonsynonymous-to-synonymous substitutions per site (K_a/K_s) and is commonly used as a measure of selective constraint (Yang and Nielsen 2000). For these data, we observe a weakly significant negative correlation (Pearson's $r = -0.29$, $P = 0.026$) between ω_g and E, despite previous work (Bustamante et al. 2000) and naïve expectation that would both predict that

Table 1
Tests of Model Significance

LRT	Null Model	#AA Sub. Cats.[a]	Alternative Model	#AA Sub. Cats.[a]	df	# Sig. Results[b]
$2\Delta\ln L_{\text{MG/GY-struct}}$	MG/GY 94	1	MG/GY 94	2	1	61**
$2\Delta\ln L_{\text{prop}}$	LCAP	1	LCAP	1	1	
Chem-comp	$\alpha_{\text{c-c}} = 0$		$\alpha_{\text{c-c}} \neq 0$			1*
Polarity	$\alpha_{\text{pol}} = 0$		$\alpha_{\text{pol}} \neq 0$			4*
Volume	$\alpha_{\text{vol}} = 0$		$\alpha_{\text{vol}} \neq 0$			59*
Isoelec	$\alpha_{\text{iso-e}} = 0$		$\alpha_{\text{isoe}} \neq 0$			25*
Hydropathy	$\alpha_{\text{hydro}} = 0$		$\alpha_{\text{hydro}} \neq 0$			44*
$2\Delta\ln L_{\text{struct-prop}}$	LCAP	2	LCAP	2	1	
Chem-comp	$\alpha^{\text{i}}_{\text{c-c}} = \alpha^{\text{s}}_{\text{c-c}}$		$\alpha^{\text{i}}_{\text{c-c}} \neq \alpha^{\text{s}}_{\text{c-c}}$			5*
Polarity	$\alpha^{\text{i}}_{\text{pol}} = \alpha^{\text{s}}_{\text{pol}}$		$\alpha^{\text{i}}_{\text{pol}} \neq \alpha^{\text{s}}_{\text{pol}}$			5*
Volume	$\alpha^{\text{i}}_{\text{vol}} = \alpha^{\text{s}}_{\text{vol}}$		$\alpha^{\text{i}}_{\text{vol}} \neq \alpha^{\text{s}}_{\text{vol}}$			9*; 15***
Isoelec	$\alpha^{\text{i}}_{\text{isoe}} = \alpha^{\text{s}}_{\text{iso-e}}$		$\alpha^{\text{i}}_{\text{isoe}} \neq \alpha^{\text{s}}_{\text{iso-e}}$			39*; 40***
Hydropathy	$\alpha^{\text{i}}_{\text{hydro}} = \alpha^{\text{s}}_{\text{hydro}}$		$\alpha^{\text{i}}_{\text{hydro}} \neq \alpha^{\text{s}}_{\text{hydro}}$			14*; 15***
$2\Delta\ln L_{\text{MG/GY-LCAP}}$	MG/GY 94	1	LCAP[c]	1	3	61**
$2\Delta\ln L_{\text{LCAP-struct}}$	LCAP[c]	1	LCAP[d]	2	4	61**
$2\Delta\ln L_{\text{MG/GY-LCAP-struct}}$	MG/GY 94	2	LCAP[d]	2	6	61**
$2\Delta\ln L_{\text{MG/GY-SG}}$	MG/GY 94	1	SG	1	1	
Volume[e]						32**
Charge[f]						22**
Polarity[g]						44**
$2\Delta\ln L_{\text{SG-struct}}$	SG	1	SG	2	2	
Volume[e]						60**
Charge[f]						61**
Polarity[g]						60**
$2\Delta\ln L_{\text{MG/GY-SG-struct}}$	MG/GY 94	2	SG	2	2	
Volume[e]						30**
Charge[f]						53**
Polarity[g]						50**

[a] Number of amino acid substitution rate categories. A single substitution category was used for models that did not incorporate differences between surface and interior residues. Such differences were modeled by including a second category of nonsynonymous rates (see text).

[b] **Significance level ($\alpha = 0.0008$), *significance level ($\alpha = 0.05$), and *** significance level ($\alpha = 0.05$) with $\alpha^{\text{i}}_{\text{c-c}} = \alpha^{\text{s}}_{\text{c-c}} = \alpha^{\text{i}}_{\text{pol}} = \alpha^{\text{s}}_{\text{pol}} = 0$.

[c] For this model, we required that $\alpha_{\text{c-c}} = \alpha_{\text{pol}} = 0$ (see text).

[d] For this model, we required that, $\alpha^{\text{i}}_{\text{c-c}} = \alpha^{\text{s}}_{\text{c-c}} = \alpha^{\text{i}}_{\text{pol}} = \alpha^{\text{s}}_{\text{pol}} = 0$ (see text).

[e] Groups used: {E, F, H, I, K, L, M, Q, R, Y, W}; {A, C, D, G, N, P, S, T, V}.

[f] Groups used: {H, K, R}; {D, E}; {A, C, F, G, I, L, M, N, P, Q, S, T, V, W, Y}.

[g] Groups used: {C, D, E, H, K, N, Q, R, S, T, W, Y }; {A, F, G, I, L, M, P, V}.

genes with proportionally more exposed residues (higher E) would have a higher ω_{g}. However, an important determinant of E is the length of the protein (larger proteins have proportionally smaller surface areas). Indeed E and length are more strongly (negatively) correlated than E and ω_{g} (Pearson's $r = -0.40$, $P = 0.001$). Because protein length and selective constraint are known to be correlated (Bloom et al. 2006), it is possible that the original correlation results from the covariation of these variables. And in fact the correlation between length and ω_{g} is nearly as large as that between E and ω_{g} (Pearson's $r = 0.26$, $P = 0.04$).

Modeling Structural Differences in Selective Constraint

These results point out the difficulties of between-gene comparisons, namely, that selective constraint is known to covary with a number of factors including expression (Drummond et al. 2005, 2006) and gene essentiality (Jordan et al. 2002; Pál et al. 2003). Instead, for this analysis, it is more appropriate to consider differences in substitution rates among sites within a single protein. To do so, we employ the models described above, where a codon's substitution rates are allowed to vary according to the surface exposure of the residue at that codon.

In all cases, a version of the MG/GY 94 model which allows a different selective constraint for interior (ω_{i}) and surface (ω_{s}) residues fits these data significantly better than does a single rate model where the average constraint is given by ω_{g} ($P < 0.0008$, Bonferonni significance level $\alpha = 0.0008$; $2\Delta\ln L_{\text{MG/GY-struct}}$, table 1). Selective constraints for interior positions were found to be higher for all alignments (fig. 1). Perhaps surprisingly, we see relatively little variation in the ratio of $\omega_{\text{i}}/\omega_{\text{s}}$, with all but three observations falling between 0.035 and 0.145. One might suspect that the value of this ratio would also depend on ω_{g} (the average constraint). However, this does not appear to be the case: There is much more variation in ω_{g} and no significant correlation between $\omega_{\text{i}}/\omega_{\text{s}}$ and ω_{g} (Pearson's $r = -0.16$, $P = 0.21$; fig. 1).

Substitution Patterns and Differences among the Amino Acid Residues

To further study the dynamics of substitution rates between surface and interior residues, we applied a model that describes the probability of a substitution between two codons as a function of the differences in the physical and

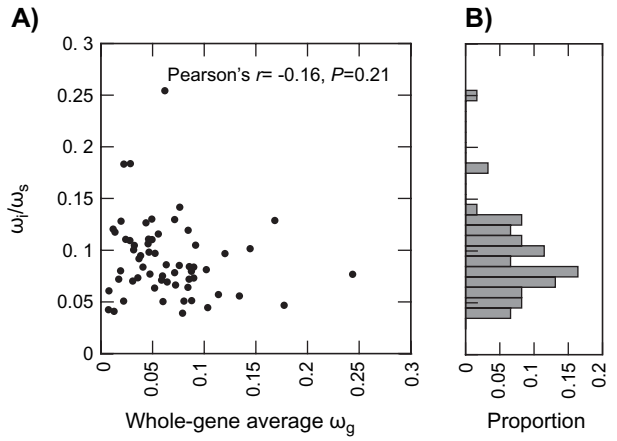

FIG. 1.—No association of average ω and the relative constraint of interior residues. (*A*) Scatter plot of the genewide average value of ω for the 61 genes studied (ω_g, *x* axis) versus the ratio of the interior average ω to the surface average ω (ω_i/ω_s, *y* axis). (*B*) Histogram showing the range of ω_i/ω_s seen along the *y* axis of *A*.

chemical properties of corresponding amino acids. Previously, we used five such properties: the volume of the residue side chain, the isoelectric point, the residue's polarity, the chemical composition of the side chain, and the residue's hydropathy (Conant et al. 2007). To ascertain whether all five properties were required to adequately model the variation in the sequences used here, we applied another LRT. For each property *j*, we compared the likelihood of observing the 61 sequence alignments when that property was included in the model ($\alpha_j \neq 0$) with the likelihood when it was excluded ($\alpha_j = 0$). Of these 61 alignments, only one showed a nominally significant constraint in chemical composition and only four in polarity ($P < 0.05$; $2\Delta\ln L_{\text{prop}}$; table 1). These two properties also only rarely showed significant differences after separation based on solvent accessibility ($P < 0.05$; $2\Delta\ln L_{\text{struct-prop}}$; table 1). These numbers are not significantly different from what would be expected under a 5% false-positive error rate ($P \geq 0.19$), leading us to exclude these two properties from further analysis. Note that when discussing individual property results, we have used the nominal significance cutoff of $P \leq 0.05$ for illustrative purposes: Our conclusions do not rest on this judgment of significance, and thus we do not control the inherent multiple testing issues.

In all cases, the LCAP model fits these data better than does the MG/GY 94 model ($P < 0.0003$; $2\Delta\ln L_{\text{MG/GY-LCAP}}$; table 1). To allow for differences in surface and interior substitution patterns, we allow $\alpha_j^i \neq \alpha_j^s$ and $C^i \neq C^s$. Doing so significantly improves the fit of the data to the model ($P < 0.0001$; $2\Delta\ln L_{\text{LCAP-struct}}$; table 1). For each sequence alignment and each of the three properties, we tested whether we could reject the null hypothesis $\alpha_j^i = \alpha_j^s$ ($2\Delta\ln L_{\text{struct-prop}}$; table 1). We plot the ratio of the interior to surface weight (α_j^i/α_j^s) against $2\Delta\ln L_{\text{struct-prop}}$ in figure 2. Strikingly, the constraint on isoelectric point differs in nominal significance between the interior and surface in 40 sequence alignments, whereas volume and hydropathy are only significantly differently constrained in 15 alignments each ($P < 0.05$; $2\Delta\ln L_{\text{struct-prop}}$; table 1). This difference is likely a correlate of the fact that volume and hydropathy are significantly constrained on average throughout these sequence alignments: Fifty-nine and 44 cases show significant improvement in fit when a single α_{vol} or α_{hydro} is added ($P < 0.05$; $2\Delta\ln L_{\text{prop}}$; table 1), whereas isoelectric point constraint is significant in only 25 comparisons.

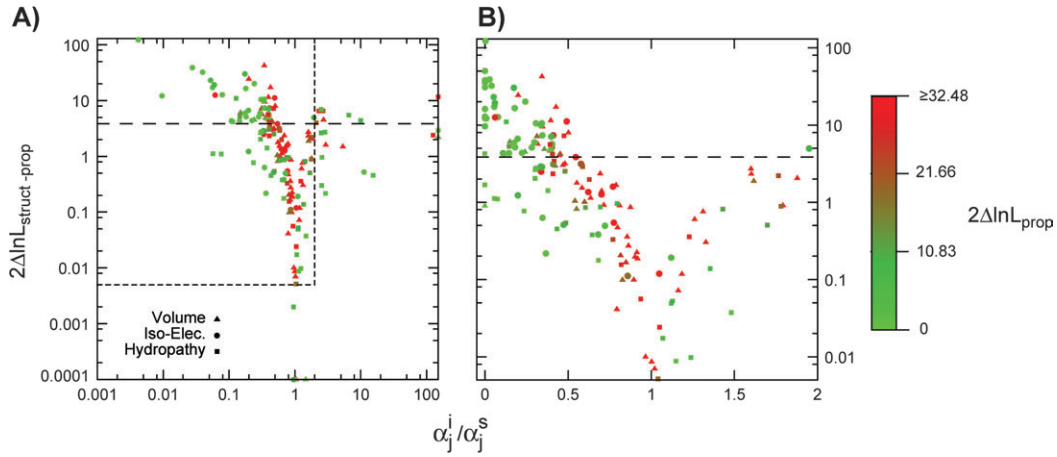

FIG. 2.—Relative interior to surface constraint for three amino acid properties. Plotted on the *x* axis is the ratio of the weight given to that property for interior residues to the weight given to surface residues (α_j^i/α_j^s). When this ratio is equal to one, the constraint is the same for interior and surface residues. Values less than one indicate greater constraint in the interior and values greater than one less constraint in the interior. On the *y* axis is shown twice the difference in ln-likelihood between a model where only one weight is given to the property in question and one where the interior and surface residues have differing weights ($2\Delta\ln L_{\text{struct-prop}}$; table 1). Large values of this statistic are associated with significant differences in constraint between the surface and interior. The horizontal dashed line in the two panels indicates a value of $2\Delta\ln L_{\text{struct-prop}} = 3.85$, corresponding to $P = 0.05$. Points are colored based on twice the difference in ln-likelihood for comparing a "single" rate model where $\alpha_j = 0$ to one where $\alpha_j \neq 0$ ($2\Delta\ln L_{\text{prop}}$; table 1). Large values of this statistic (red) indicate significant evidence for an overall constraint on that property, ignoring structural features. The scale bar on the right indicates three values of $2\Delta\ln L_{\text{prop}}$ (= 10.83, $P = 0.001$, = 21.66, $P = 3.3 \times 10^{-16}$, and ≥32.48, $P \leq 1.2 \times 10^{-8}$). Note that values of $2\Delta\ln L_{\text{struct-prop}}$ less than 10^{-4} in panel *A* are indicated with this value (four cases), whereas cases where $\alpha_j^s = 0 = 0$ (i.e., α_j^i/α_j^s is undefined) are indicated with arbitrary values of 150 (three cases). (*A*) Overview of the range of constraint across the 61 alignments for the three properties. Observe the log scale on the *x* axis (meaning that values where $\alpha_j^i/\alpha_j^s = 0$ are not shown). (*B*) Enlargement of the area in the dashed box in panel *A* with the log scale omitted for the *x* axis and thus showing cases of $\alpha_j^i/\alpha_j^s = 0$.

Solvent Exposure and Amino Acid Residue Classification

The LCAP model is not the only model allowing for variation in amino acid substitution rates. Another approach is to divide the amino acids into groups such that group members are more similar to each other in some property than they are to members of the other group(s). One can then allow one substitution rate for two residues that are members of the same group (I_w) and a second (generally lower) rate (I_b) when the two residues are in different groups (Sainudiin et al. 2005; Wong et al. 2006). We have previously implemented this model and refer to it as the SG model (Conant et al. 2007).

We thus asked if the above results were supported by analyses using the SG model. We used three groupings of the amino acid residues presented by Sainudiin et al. (2005): those by volume, charge, and polarity (table 1). These groups roughly reflect the volume, isoelectric point, and hydropathy properties used in the LCAP model above. We can model differences in surface and interior substitution rates by allowing a different I_b and I_w value for interior and surface residues (I_b^i, I_b^s, and I_w^i, I_w^s, respectively).

In general, the results from the SG model are similar to those from the LCAP model. Thus, volume and polarity are more likely than charge to show an overall constraint in a protein when solvent exposure is disregarded ($2\Delta\ln L_{\text{MG/GY-SG}}$; table 1). Likewise, charge more commonly improves the fit of the model when surface and interior residues are distinguished ($2\Delta\ln L_{\text{MG/GY-SG-struct}}$; table 1).

The LCAP and SG models are not nested with respect to each other and should not be compared with an LRT. We nonetheless note that the LCAP model in all cases constitutes an improved fit over the corresponding MG/GY 94 model ($2\Delta\ln L_{\text{MG/GY-LCAP}}$ and $2\Delta\ln L_{\text{MG/GY-LCAP-struct}}$; table 1). The SG model on the other hand is often not an improvement over the corresponding MG/GY 94 model ($2\Delta\ln L_{\text{MG/GY-SG}}$ and $2\Delta\ln L_{\text{MG/GY-SG-struct}}$; table 1). These results illustrate a point we have made previously: Differing models of codon substitution are appropriate for different problems. The LCAP model is helpful in this analysis because it can account for multiple amino acid properties simultaneously. In other applications, the SG model may be preferable, both because its parameters have straightforward interpretations and because the groups used can be chosen to be appropriate for the question at hand.

Discussion

Our results show that there are systematic differences in the evolution of interiors and surfaces of proteins. Indeed, these two regions not only evolve differently, but these differences also are similar across proteins and correlate with the physical properties of the amino acids.

Although proteins vary considerably in their overall selective constraint, the relative constraints on the interior versus the surface appear more uniform (fig. 1). This observation is somewhat surprising, and we do not have a simple explanation for the phenomenon. We speculate that it reflects the fact that substitutions in the interior are more likely to cause misfolding than a substitution on the surface (Chothia and Finkelstein 1990). The difference between surface and interior substitution patterns thus would result directly from the physics of protein folding and hence might well be at least approximately independent of the overall selective constraint. One issue that will bear investigation is the role of protein length, because larger proteins tend to have a greater proportion of buried residues. Interestingly, Bloom et al. (2006) have recently shown that proteins with many buried residues tend to evolve more rapidly, a result that may explain the association seen here between length and selective constraint. These authors attribute this effect to what is essentially a larger space of neutral mutations in proteins with more buried residues. However, comparing these authors' results with our analyses of individual genes is complicated by the issues already mentioned with associating differences between proteins and patterns within a protein.

We also find that the most obvious difference in substitution patterns between protein surfaces and interiors is in whether changes in residue charge are allowed: Such changes are more strongly selected against in protein interiors (fig. 2). Because charged residues are generally not found in protein interiors, another way to view this observation is as a selective constraint against the introduction of charged residues into the protein interior. Interestingly, neither residue volume nor residue hydropathy appears to be differentially constrained to this same degree. Because volume in particular is nonetheless significantly constrained across the proteins as a whole, we suggest that, because almost all residues are in contact with at least one other residue, changes in residue volume can disrupt protein folding at most positions in the protein.

Our conclusions will be most informative with regards to the evolution of globular proteins, because such proteins are more tractable for crystallization and hence overrepresented in the data set used. It is less clear whether the same evolutionary rules govern other protein classes, in particular because we have previously seen that membrane-spanning proteins evolve differently than do other proteins (Conant et al. 2007). We also cannot be certain that these patterns will hold in other taxa, although given the functional range of genes surveyed, we would not be surprised if this were the case.

Nonetheless, the observations above are of interest both because they provide an improved understanding of the evolutionary process and because they may also help provide a statistical framework for the process of protein engineering. It is already well known that amino acid residue conservation can indicate mutations that are likely to increase protein stability (Ohage et al. 1997; Lehmann et al. 2000, 2002). We suggest that models such as those above may indicate which particular amino acid properties are likely to be relevant when performing such analyses. They may also help suggest pathways to move the engineering process from the modification of existing functions to the development of new ones (Ryu and Nam 2000).

Finally, this work illustrates a complementary mode of analysis for understanding the factors influencing gene evolution. Although evolution has provided innumerable natural experiments whereby two different genomes or sets of genes can be compared to study what factors influence

substitution rates (Jordan et al. 2002, 2003; Pál et al. 2003; Hahn et al. 2004; Bloom et al. 2006; Drummond et al. 2006), it is also possible to employ a modeling approach where a potentially important factor is included in the model. Thus, our analysis here makes solvent exposure a parameter in a model. A useful feature of this tactic is that it becomes less likely that the signal of interest will be confounded by some third variable, such as gene expression level. By using both within and between gene approaches, we can more clearly discern patterns in molecular evolution. And one final benefit of developing models of sequence evolution is that they require us to define many of our assumptions, allowing the option of testing these assumptions in the future.

Supplementary Material

Supplementary figures 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/). Sequence alignments, tree files, and proportional surface exposures are available from http://web.missouri.edu/~conantg/data/yeast_struct_evol/alignments_trees_rates.tar.gz.

Acknowledgments

We thank K. Byrne for assistance with the Yeast Gene Order Browser and B. Cusack, A. Mosig, D. Scannell, M. Sémon, G.P. Wagner, K.H. Wolfe, M. Woolfit, and two anonymous reviewers for helpful suggestions during the preparation of this manuscript. This work was supported by a Science Foundation Ireland grant to K.H. Wolfe and by start-up funds to G.C.C. from the Food for the 21st Century Program at the University of Missouri-Columbia.

Literature Cited

Arnaud MB, Costanzo MC, Skrzypek MS, Binkley G, Lane C, Miyasato SR, Sherlock G. 2005. The *Candida* genome database (CGD), a community resource for *Candida albicans* gene and protein information. Nucleic Acids Res. 33:D358–D363.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. Nucleic Acids Res. 28:235–242.

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.

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

Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. 15:1456–1461.

Chothia C. 1976. The nature of the accessible and buried surfaces in proteins. J Mol Biol. 105:1–12.

Chothia C, Finkelstein AV. 1990. The classification and origins of protein folding patterns. Annu Rev Biochem. 59:1007–1039.

Conant GC, Wagner GP, Stadler PF. 2007. Modeling amino acid substitution patterns in orthologous and paralogous genes. Mol Phylogenet Evol. 42:298–307.

Conant GC, Wolfe KH. 2008. Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. Genetics. 179:1681–1692.

Dietrich FS, Voegeli S, Brachat S. (13 co-authors). 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. Science. 304:304–307.

Diezmann S, Cox CJ, Schonian G, Vilgalys RJ, Mitchell TG. 2004. Phylogeny and evolution of medical species of *Candida* and related taxa: a multigenic analysis. J Clin Microbiol. 42:5624–5635.

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.

Dujon B, Sherman D, Fischer G, et al. (69 co-authors). 2004. Genome evolution in yeasts. Nature. 430:35–44.

Eisenhaber F, Argos P. 1993. Improved strategy in analytic surface calculation for molecular systems: handling of singularities and computational efficiency. J Comput Chem. 14:1272–1280.

Eisenhaber F, Lijnzaad P, Argos P, Sander C, Scharf M. 1995. The double cube lattice method: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. J Comput Chem. 16:273–284.

Goffeau A, Barrell BG, Bussey H, et al. (15 co-authors). 1996. Life with 6000 genes. Science. 274:546;563–567.

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

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736.

Hahn MW, Conant GC, Wagner A. 2004. Molecular evolution in large genetic networks: connectivity does not equal constraint. J Mol Evol. 58:203–211.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 22:160–174.

Jones T, Federspiel NA, Chibana H, et al. (11 co-authors). 2004. The diploid genome sequence of *Candida albicans*. Proc Natl Acad Sci USA. 101:7329–7334.

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

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.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature. 428:617–624.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 423:241–254.

Kurtzman CP, Robnett CJ. 2003. Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. FEMS Yeast Res. 3:417–432.

Lehmann M, Kostrewa D, Wyss M, Brugger R, D'Arcy A, Pasamontes L, van Loon AP. 2000. From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. Prot Eng. 13:49–57.

Lehmann M, Loch C, Middendorf A, Studer D, Lassen SF, Pasamontes L, van Loon AP, Wyss M. 2002. The consensus

concept for thermostability engineering of proteins: further proof of concept. Prot Eng. 15:403–411.
Mintseris J, Weng Z. 2005. Structure, function, and evolution of transient and obligate protein–protein interactions. Proc Natl Acad Sci USA. 102:10930–10935.
Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 11:715–724.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics. 148:929–936.
Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302:205–217.
Ohage EC, Graml W, Walter MM, Steinbacher S, Steipe B. 1997. β-Turn propensities as paradigms for the analysis of structural motifs to engineer protein stability. Prot Sci. 6:233–241.
Pál C, Papp B, Hurst LD. 2003. Rate of evolution and gene dispensability. Nature. 421:496497.
Reeves JH. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J Mol Evol. 35:17–31.
Ryu DDY, Nam D-H. 2000. Recent progress in biomolecular engineering. Biotechnol Prog. 16:2–16.
Sainudiin R, Wong WS, Yogeeswaran K, Nasrallah JB, Yang Z, Nielsen R. 2005. Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. J Mol Evol. 60:315–326.
Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature. 440:341–345.
Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci USA. 104:8397–8402.
Sidow A, Steel TP. 1992. Estimating the fraction of invariable codons with a capture–recapture method. J Mol Evol. 35: 253–260.
Sokal RR, Rohlf FJ. 1995. Biometry, 3rd ed. New York: W.H. Freeman and Company.
Swofford DL. 2002. PAUP*. Sunderland (MA): Sinauer.
Thorne JL, Goldman N, Jones DT. 1996. Combining protein evolution and secondary structure. Mol Biol Evol. 13: 666–673.
Tourasse NJ, Li W-H. 2000. Selective constraints, amino acid composition, and the rate of protein evolution. Mol Biol Evol. 17:656–664.
Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature. 387:708–713.
Wong WS, Sainudiin R, Nielsen R. 2006. Identification of physicochemical selective pressure on protein encoding nucleotide sequences. BMC Bioinform. 7:148.
Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol. 10:1396–1401.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 39:306–314.
Yang Z, Nielsen R. 2000. Estimating synonymous and non-synonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17:32–43.
Yang Z, Nielsen R, Goldman N, Pedersen A-MK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431–449.
Yang Z, Swanson WJ, Vacquier VD. 2000. Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol Biol Evol. 17:1446–1455.

Jeffrey Thorne, Associate Editor

Accepted February 16, 2009