Intermediate divergence levels maximize the strength of structure–sequence correlations in enzymes and viral proteins

Eleisha L. Jackson,[1,2,3] Amir Shahmoradi,[2,3,4] Stephanie J. Spielman,[1,2,3] Benjamin R. Jack,[1,2,3] and Claus O. Wilke[1,2,3]*

[1]Department of Integrative Biology, The University of Texas at Austin, Austin, Texas 78712

[2]Center for Computational Biology and Bioinformatics, The University of Texas at Austin, Austin, Texas 78712

[3]Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, Texas 78712

[4]Department of Physics, The University of Texas at Austin, Austin, Texas 78712

Received 7 November 2015; Accepted 4 March 2016
DOI: 10.1002/pro.2920
Published online 00 Month 2016 proteinscience.org

Abstract: Structural properties such as solvent accessibility and contact number predict site-specific sequence variability in many proteins. However, the strength and significance of these structure–sequence relationships vary widely among different proteins, with absolute correlation strengths ranging from 0 to 0.8. In particular, two recent works have made contradictory observations. Yeh *et al*. (Mol. Biol. Evol. 31:135–139, 2014) found that both relative solvent accessibility (RSA) and weighted contact number (WCN) are good predictors of sitewise evolutionary rate in enzymes, with WCN clearly out-performing RSA. Shahmoradi *et al*. (J. Mol. Evol. 79:130–142, 2014) considered these same predictors (as well as others) in viral proteins and found much weaker correlations and no clear advantage of WCN over RSA. Because these two studies had substantial methodological differences, however, a direct comparison of their results is not possible. Here, we reanalyze the datasets of the two studies with one uniform analysis pipeline, and we find that many apparent discrepancies between the two analyses can be attributed to the extent of sequence divergence in individual alignments. Specifically, the alignments of the enzyme dataset are much more diverged than those of the virus dataset, and proteins with higher divergence exhibit, on average, stronger structure–sequence correlations. However, the highest structure–sequence correlations are observed at intermediate divergence levels, where both highly conserved and highly variable sites are present in the same alignment.

Keywords: protein evolution; protein design; relative solvent accessibility; site variability; packing density

Introduction

Proteins are subject to a number of biophysical and functional constraints that influence their evolutionary trajectories.[1–4] These constraints contribute to observed patterns in both whole-gene evolutionary rate variation[5–9] and evolutionary rate variation among sites within individual proteins.[10–14] Such evolutionary rate variation in turn contributes to heterogeneity in site-specific sequence variability.

A number of studies have sought to understand the roles that biophysical constraints, particularly structural constraints, play in this observed site-specific variability within proteins. Structural

Additional Supporting Information may be found in the online version of this article.

Disclosure: The authors declare no conflict of interest.

Grant sponsor: NIH grant; Grant number: R01 GM088344; Grant sponsor: NSF Cooperative Agreement; Grant number: DBI-0939454 (BEACON Center); Grant sponsor: National Institutes of Health (www.nih.gov/); Grant number: F31GM113622; Grant sponsor: National Science Graduate Research Fellowship; Grant number: DGE-1110007.

*Correspondence to: C. O. Wilke; Department of Integrative Biology, College of Natural Sciences, The University of Texas at Austin, 205 W. 24th St. Stop C0930, Austin, TX 78712. E-mail: wilke@austin.utexas.edu

properties such as solvent exposure and packing density have emerged as strong predictors of sitewise evolutionary rates.[11,13,15,16] Solvent exposure is typically measured with the metric relative solvent accessibility (RSA), which indicates the extent to which a given residue comes into contact with solvent (i.e., water).[17] Residues that are exposed on the surface of the protein have high RSA, with complete exposure indicated with an RSA of one. Residues that are buried and/or in the protein core have low RSA, with completely buried residues having an RSA of zero. RSA has a significant, positive relationship with evolutionary rate, such that more buried residues tend to evolve more slowly than exposed residues do.[10,16,18–23]

Alternatively, packing density indicates how tightly packed a given residue is by neighboring amino acids in a protein's tertiary structure. A residue's packing density is commonly measured as weighted contact number (WCN), which is defined as the sum of the inverse square distance of all residues in the protein to the focal amino acid.[24,25] Recent work has suggested that WCN is a strong determinant of site-specific variability in proteins, and that residues with high WCN evolve more slowly than do residues with low WCN.[8,11,12,15]

However, some studies have yielded apparently contradictory results regarding the extent of the predictive power that these structural properties have on sitewise evolutionary rate (ER). For example, Yeh *et al.*[11] investigated structure–sequence relationships in a dataset of 216 monomeric enzymes, finding that WCN is a stronger determinant of sitewise ER than RSA, although RSA was still a significant predictor. Importantly, Yeh *et al.*[11] recovered strong correlations between structure and ER, with WCN and RSA explaining up to ~41% of the variance in site-specific ER. By contrast, Shahmoradi *et al.*[13] examined the structure–sequence relationship on a set of 9 viral proteins. While Shahmoradi *et al.*[13] similarly found that both RSA and WCN are significant predictors of rate in proteins, the correlations Shahmoradi *et al.*[13] observed were much smaller in magnitude.[13] Specifically, they found that at best, structural predictors could explain only ~15% of the variance in ER. Given these disparate findings, it remains unclear which of the two studies is the more representative one.

Although both Yeh *et al.*[11] and Shahmoradi *et al.*[13] examined the relationship between sequence and structural properties, they used different methods and datasets. First, Yeh *et al.*[11] measured ER using the method Rate4Site,[26,27] whereas Shahmoradi *et al.*[13] focused on sequence entropy, which is not a rate. Second, Yeh *et al.*[11] used a much more comprehensive dataset of monomeric enzymes, and Shahmoradi *et al.*[13] analyzed a comparatively smaller set of viral proteins, which are subject to an additional layer of selective forces imposed by the host immune system. Finally, Shahmoradi *et al.*[13] considered additional structural predictors, namely protein design and flexibility, while Yeh *et al.*[11] focused on RSA and WCN alone. This use of different methods makes it difficult to directly compare results from the two studies.

Here, we attempt to reconcile these two studies, by reanalyzing both the enzyme dataset from Yeh *et al.*[11] and the virus dataset from Shahmoradi *et al.*[13] in one consistent analysis pipeline. We focus on three structural predictors from the two studies: WCN, RSA, and variability in designed proteins. We confirm that, indeed, correlations between rate and structural predictors are much smaller for the viral proteins compared to the enzymes. However, differences in structural characteristics do not appear to drive the low predictive power in the viral protein dataset. Instead, we find that the enzyme and viral protein datasets primarily differ in the extent of sequence variability in the multiple-sequence alignments (MSAs) used to infer evolutionary rates. Using evolutionary models, we quantify sequence divergence for all individual MSAs, and we find that the enzyme dataset displays very high levels of divergence while the viral protein dataset has experienced minimal evolutionary divergence. Across both datasets, we observe that the strongest structure–sequence correlations are observed at intermediate divergence levels. We conclude that the strength of the structure–structure relationship in proteins is, in part, determined by the extent of sequence variability in the datasets analyzed.

Results

We analyzed two distinct datasets. One was a set of 208 diverse enzyme monomers selected from the prior analysis by Yeh *et al.*[11] The other dataset was a smaller set of nine viral proteins from Shahmoradi *et al.*[13] Note that while the viral dataset from Shahmoradi *et al.*[13] includes some viral enzymes, in the following we will use the term "enzymes" to refer specifically to the proteins from the Yeh *et al.*[11] dataset.

Homologous sequences for each protein were taken from Yeh *et al.*[11] and Shahmoradi *et al.*[13] For each protein we made a multiple–sequence alignment using MAFFT[28,29] on amino-acid sequences. From these alignments we calculated site-specific evolutionary rates using Rate4Site.[26] We measured solvent accessibility for a given residue by its relative solvent accessibility (RSA) [Fig. 1(A)]. We measured packing density in the protein structures using side chain WCN [Fig. 1(B)]. Previous studies have used C_α WCN when correlating WCN with ER.[11,13,15] However, a recent study[30] has shown that calculating WCN using the center of mass of the side chain results in stronger WCN–ER correlations.

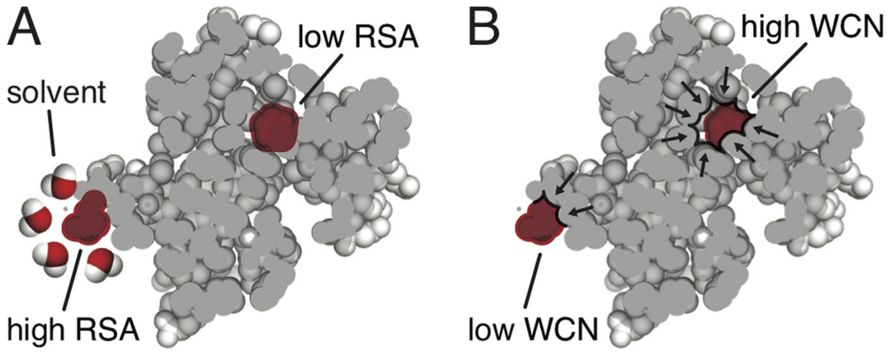

Figure 1. Description of structural properties. (A) Visualization of solvent accessibility. (B) Visualization of local packing density. Each colored red particle represents a residue in the protein. In A, the lower red particle represents a surface residue. The red and white molecules indicate solvent molecules (e.g., water) that are contacting the red amino acid. This residue has a larger solvent accessibility because there is a larger proportion of the residue surface exposed to solvent. The upper red particle represents a core residue. This residue is not in contact with any solvent molecules and thus has low solvent accessibility. Relative solvent accessibility is obtained by normalizing the solvent accessibility of a given residue by the maximum amount of solvent accessibility for that amino acid. In B, the arrows pointing towards each residue indicate contacts between the red focal residue and its neighboring residues. The upper red residue represents a residue that has many neighbors (represented by the arrows) and thus has a high weighted contact number. The lower red residue is a surface amino acid with few neighbors and thus has a lower weighted contact number.

Therefore, here we used side chain WCN throughout. We also measured the variability in designed sequences. For each protein in the viral dataset and for each enzyme less than 200 residues in length we computationally designed 500 sequences using the respective structure as a template. From these sequences we inferred a "design rate" (DR) at each site, calculated as the expected steady-state evolutionary rate for an alignment with the given amino-acid frequencies.

Structural predictors of evolutionary rate

To quantify the strength of structure–rate relationships in proteins, we correlated, separately for each protein, structural properties at individual sites with site-specific ER. Unless otherwise noted, we used Spearman's correlations throughout. The first structural property that we examined was relative solvent accessibility (RSA). Prior work has shown that RSA has a positive relationship with evolutionary rate.[8,10,11,13,15] This positive relationship between solvent accessibility and ER was verified in our analysis on the two datasets. Within both datasets, residues that have high RSA evolved faster on average. However, the strength of the relationship between RSA and ER varied between the enzyme and viral protein datasets. The enzymes, on average, had larger RSA–ER correlations with a mean correlation coefficient of 0.55 compared to 0.18 for viral proteins (t test: $P = 3.324 \times 10^{-5}$) [Fig. 2(A) and Table I].

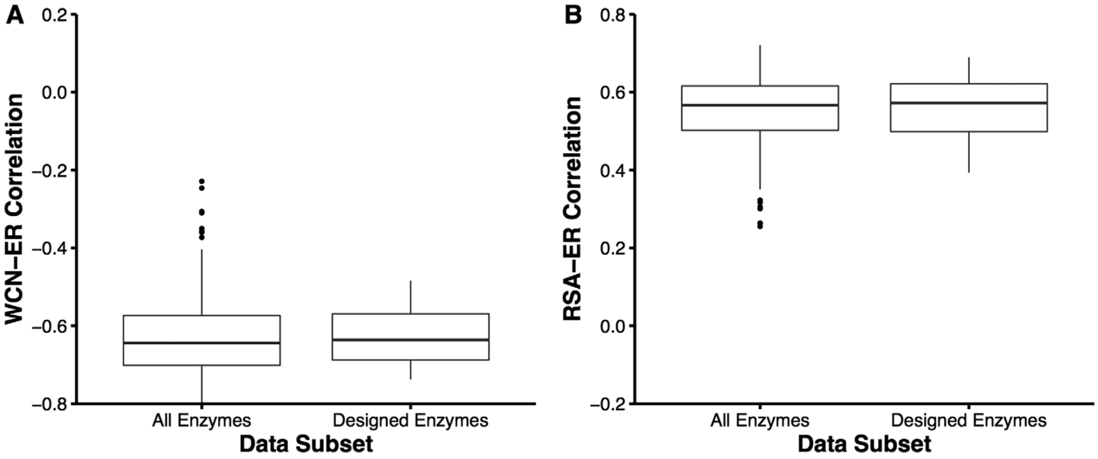

Figure 2. Distribution of correlation coefficients between structural properties and evolutionary rate (ER). (A) Spearman correlation coefficients between RSA and ER for the two datasets (t test: $P = 3.324 \times 10^{-5}$). (B) Spearman correlation coefficients between WCN and ER for the two datasets. For all structural properties, on average, viral proteins show weaker correlations than do enzymes (t test: $P = 2.454 \times 10^{-5}$).

Table I. *Averages of Spearman Correlation Coefficients Between Structural Properties and Evolutionary Rate (ER)*

Dataset	$\langle\rho_{\text{ER-WCN}}\rangle$	$\langle\rho_{\text{ER-RSA}}\rangle$	$\langle\rho_{\text{ER-DR}}\rangle^{a}$	$\langle\rho_{\text{ER-WCN}}\rangle^{a}$	$\langle\rho_{\text{ER-RSA}}\rangle^{a}$
Enzyme	−0.626	0.549	0.240	−0.625	0.561
Virus	−0.207	0.184	−0.022	−0.207	0.184

The structural properties analyzed are RSA, WCN, and predicted rate of designed proteins (DR). The analysis was performed on two datasets, one comprises 208 enzyme monomers and comprises nine viral proteins. Structure–ER correlations are higher in absolute magnitude in enzymes.

[a] Correlation coefficients calculated using the 32 enzymes and nine viral proteins for which there were designed sequences.

Next we investigated the relationship between ER and packing density. For both datasets, residues with more contacts evolved slower [Fig. 1(B) and Table I]. This trend was also stronger for enzymes than for viral proteins, with a mean correlation coefficient of −0.63 for enzymes and −0.21 for viral proteins (t test: $P = 2.454 \times 10^{-5}$).

Protein design as a structural predictor

Using protein design to search sequence space, Kuhlman and Baker[31] found that sequences are close to optimal for a given structure (i.e., residues found at a given site are limited for a given structure). This constraint is especially true for buried residues. Given this result, Shahmoradi *et al.*[13] attempted to use sitewise variability in designed proteins as an additional structural predictor of ER.[13] Likewise, here, we used protein design as a third predictor of ER. However, unlike in Shahmoradi *et al.*,[13] we did not use design entropy at sites but instead calculated a "design rate" (DR) as our predictor. We calculated this rate by calculating a predicted nonsynonymous substitution rate (dN) from amino acid frequencies at each site, as derived in Spielman and Wilke.[32] We found that this predicted rate makes similar predictions as does design entropy (not shown). We used design rate here because it is the more principled quantity to compare to ER. For computational feasibility, for the enzyme dataset we only designed proteins that were less than or equal to 200 residues in length. This encompassed 32 enzymes. We designed proteins for all the structures in the viral protein dataset. Before performing our analysis, we compared the distributions of the strength of structure–rate correlations from the full enzyme dataset with that of the distributions obtained from the 32 proteins. The differences between mean of the distributions were not significant (t test: $P = 0.419$ for RSA, $P = 0.947$ for WCN, Supporting Information Fig. S1).

In viral proteins, DR had a mean correlation coefficient of approximately −0.02, and in enzymes the mean coefficient of correlation was approximately 0.24 (Fig. 3 and Table I). However, for viral proteins this lower mean correlation was slightly misleading because some proteins had positive correlations while others had negative correlations, for a mean near zero (Fig. 3). In both datasets, design rate was a weaker predictor of evolutionary rates compared to WCN and RSA.

Even though DR did not correlate that strongly with ER, it is possible that it could explain variance in ER not explained by either RSA or WCN. To investigate this possibility, we used DR at sites as a predictor in linear models, either individually or in combination with the two other structural predictors, and calculated the percent variance explained for each model. In general, for both enzymes and viral proteins, design rate was not a good predictor of ER at sites. However, DR, just like RSA and WCN, was better at predicting ER in enzymes than in viral proteins. For a model with design rate as a single predictor, the average R^2 was 0.01 for viral proteins and ~0.07 for enzymes (Supporting Information Figs. S2 and S3). Including DR as an additional predictor along with RSA and WCN added some additional predictive power for ER in both datasets. For example, the average R^2 of a model with RSA and WCN as predictors for enzymes was approximately 0.37 (Supporting Information Fig. S2). When we added DR as an additional predictor, the average R^2 increased to 0.40 (Supporting Information Fig. S2). This increase in predictive power was observed in the viral dataset as well. In summary, although DR was poor predictor of evolutionary rate at sites, it

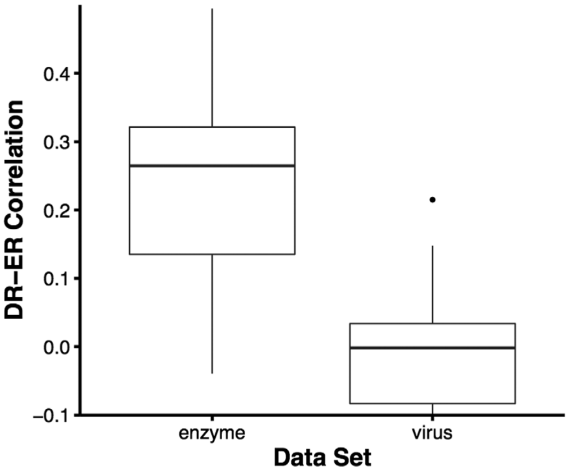

Figure 3. Correlation coefficients of design rate and evolutionary rate (ER). Distributions of Spearman's correlation coefficients between design rate (DR) and evolutionary rate (ER) for the two datasets. Enzyme proteins have higher correlations on average (t test: $P = 7.50 \times 10^{-4}$).

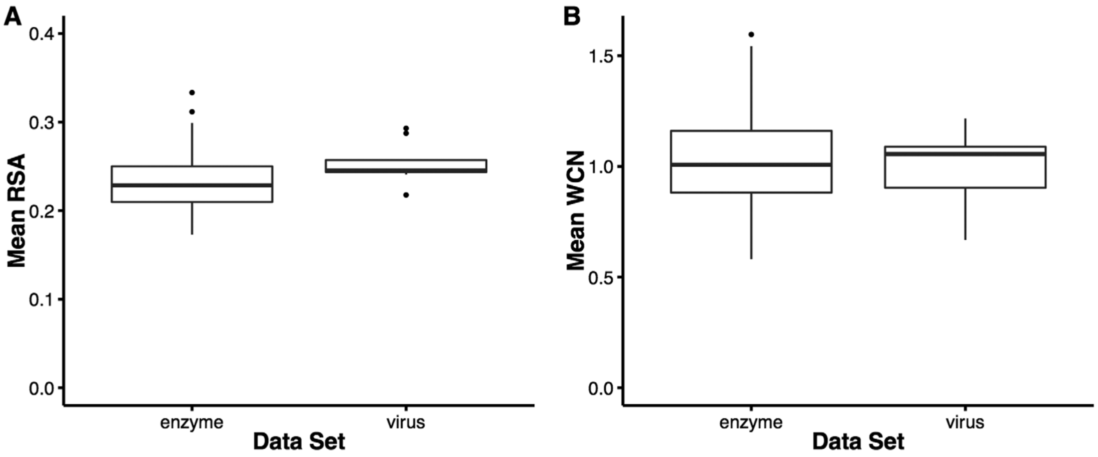

Figure 4. Distribution of average structural properties for each protein in the two datasets. (A) Distribution of average RSA. The distribution of average RSA different are very similar for both datasets (*t* test: $P = 0.027$). (B) Distribution of average WCN. The distribution of average WCN is the same for both datasets (*t* test: $P = 0.437$).

provided a small improvement in model performance, in particular for the enzyme dataset.

Effect of divergence of structure–rate relationships

We found WCN, RSA, and DR all to be poor predictors of ER in viral proteins. There could be at least two different explanations for this finding. First, there could be unique structural features found within the viral protein dataset that are not in the enzymes as indicated in Tokuriki *et al.*[33] Second, the viral proteins from Shahmoradi *et al.*[13] may have experienced unique selection pressures (such as immune escape) or different divergence times than the enzymes taken from Yeh *et al.*[11]

We found it unlikely that biophysical differences drove observed differences in the structure–rate correlations between the two datasets. First, any differences between the distributions for the mean WCN of the proteins within the datasets were not significant ($P = 0.437$ for WCN, Fig. 4). Differences in the mean RSA of the proteins were significant but the means were extremely similar (*t* test: $P = 0.027$ for RSA, Fig. 4). Second, the strength of structure–rate correlations was only weakly dependent on the mean WCN or mean RSA of a protein (Supporting Information Figs. S4 and S5). Proteins with larger mean RSA had only slightly larger RSA–ER correlations on average and the mean WCN was not related to the magnitude of structure–rate correlations (Supporting Information Figs. S4 and S5).

We next investigated the possibility that differences in the multiple-sequence alignments for the two datasets were driving the differences in predictive power of RSA, WCN, and DR. On average the enzymes have more sequences in their representative alignments. We examined whether this difference was causing the difference in structure–rate

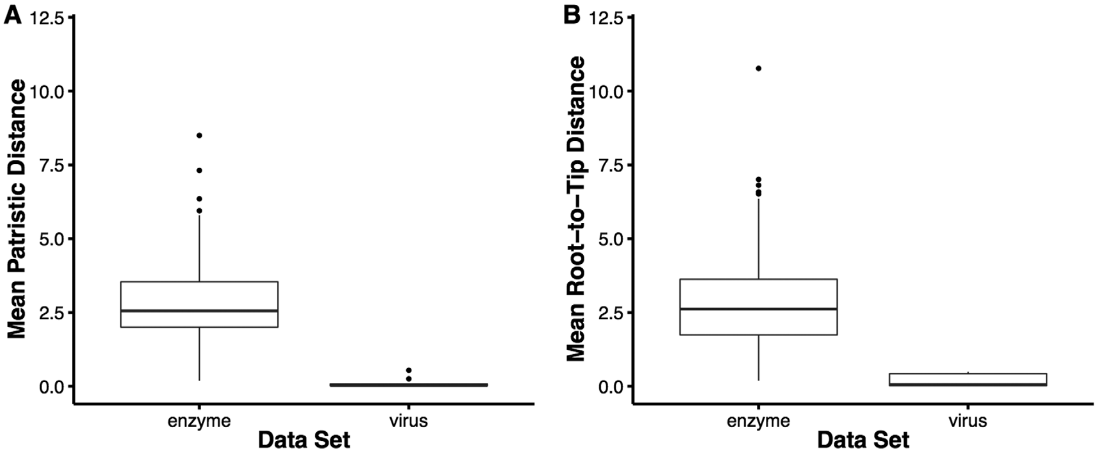

Figure 5. Divergence of sequences within the datasets. (A) Distributions of mean patristic distances for sequences in each protein alignment. Enzymes have larger mean patristic distances (*t* test: $P < 2.2 \times 10^{-16}$). (B) Distributions of mean root-to-tip distances for sequences in each protein alignment. Enzymes have larger mean root-to-tip distances (*t* test: $P < 2.2 \times 10^{-16}$). For both measures of divergence, the proteins within the enzyme dataset are more diverged. Divergence is relatively low between the viral proteins.

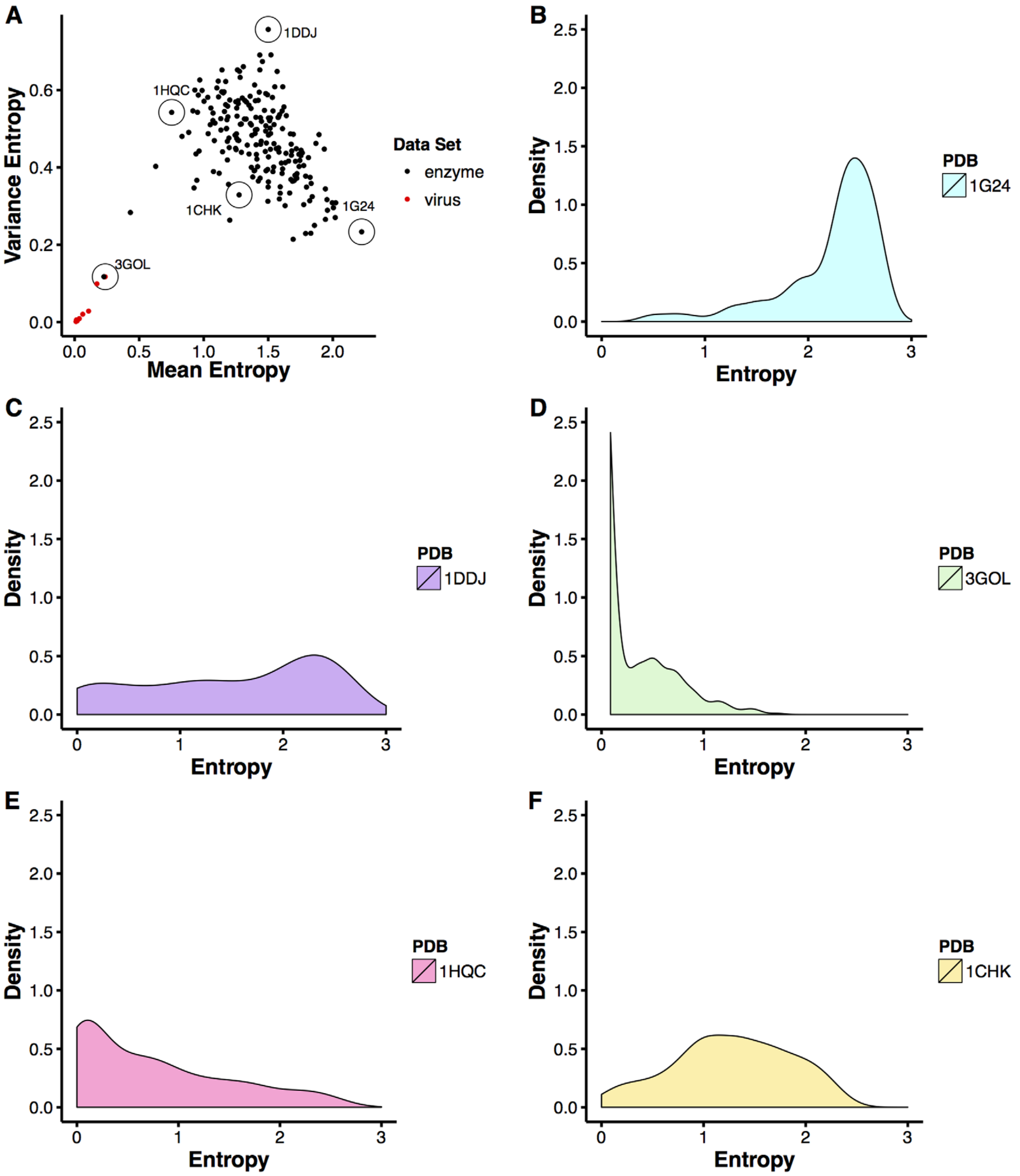

Figure 6. Comparison of the mean of entropy and the variance of entropy for individual proteins. (A) Variance in entropy at sites compared against overall mean entropy for each protein. Five different enzymes are highlighted, spanning the range of different combinations of high and low mean entropy and entropy variance. The enzymes are colored in black and the virus proteins are colored red. (B–F) Distributions of sitewise entropy values for the five proteins highlighted in (A). There are a variety of distributions in site entropy for different proteins. Note: The protein denoted by the PDB ID 3GOL is a viral protein.

correlation strength. We did observe a relationship between the number of sequences and the structure–rate strength. However, the strength of this relationship was modest for enzymes ($\rho = -0.185$, $P = 7.403 \times 10^{-3}$ for WCN–ER, and $\rho = 0.060$, $P = 0.390$ for RSA–ER) and was nonsignificant for viral proteins ($\rho = -0.433$, $P = 0.250$ for WCN–ER and $\rho = 0.633$, $P = 0.076$ for RSA–ER).

The two datasets showed significantly different levels of evolutionary divergence (Fig. 5). We calculated the divergence for each dataset using two quantities: mean root-to-tip distance and mean patristic distance. Root-to-tip distance represents the extent of evolutionary divergence from the dataset's common ancestor to a given sequence. The mean root-to-tip distance for each dataset was calculated as the average branch length, which indicates the number of substitutions, from the root in the tree to each terminal edge (tip) in the tree. Patristic, or pairwise, distance is the sum of branch lengths between two tips in a tree, and indicates how distantly related two sequences are to one another. As

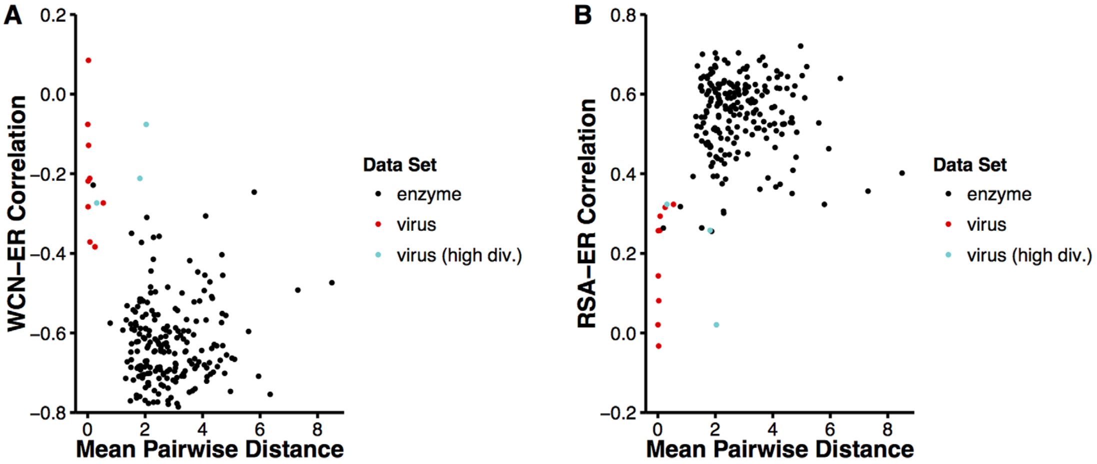

Figure 7. Comparison of structure–rate correlations with variance of entropy at sites. (A) Comparison of Spearman's correlation coefficients of WCN–ER and variance of entropy for proteins. (Spearman's correlation test: $\rho = -0.321$, $P = 1.526 \times 10^{-6}$ using only the original protein datasets), (B) correlations of RSA–ER and variance of entropy for proteins ($\rho = 0.236$, $P = 4.756 \times 10^{-4}$ using only the original protein datasets). Enzymes are black, the viral proteins with the original alignments are in red, and the viral proteins with the newly collected sequences are in turquoise. Enzymes have more variance in entropy across proteins and have larger structure–rate correlations in magnitude for both RSA and WCN. Virus proteins represented by the newly curated, more diverged alignments (see "Methods") have similar structure–rate correlations to the original viral protein dataset.

with mean root-to-tip-distance, a higher mean patristic distance indicated more evolutionary divergence. The enzyme alignments were much more diverged than the viral protein alignments (t test: $P < 2.20 \times 10^{-16}$ for mean root-to-tip distance and $P < 2.20 \times 10^{-16}$ for mean patristic distance).

Supporting Information Figure S6 shows structure–rate correlation strengths as a function of divergence (here measured as mean patristic distance). For both RSA–ER and WCN–ER correlations, proteins with MSAs that had higher levels of divergence tended to have higher structure–rate correlations in magnitude. However, the trend between RSA–ER and WCN–ER correlations and mean patristic distance was not very strong ($\rho = 0.161$, $P = 0.017$ for RSA–ER and $\rho = -0.117$, $P = 0.086$ for WCN–ER).

Because divergence correlated only weakly with the structure–rate correlations, we hypothesized that overall divergence in an alignment mattered less than did variability in divergence among sites in an alignment. To obtain strong correlations with structural quantities, we need both highly conserved and highly variable sites. To assess the variability in the alignment at each site, we next calculated Shannon entropies at each site. By plotting the variance in entropy among sites against the mean [Fig. 6(A)], we found that indeed some alignments had overall high divergence but low variability among sites while other alignments were less diverged on average but had higher variability among sites. Figure 6(B–F) shows specific examples of entropy distributions among sites for individual proteins. For example, consider the protein identified by PDB ID 1G24 [Fig. 6(B)]. This protein had high mean entropy while maintaining a relatively low variance of entropy. Thus, sites in this protein were uniformly highly variable. Note that the distributions of entropy varied greatly between proteins even when they were from the same dataset [Fig. 6(B–F)].

We next plotted structure–rate correlations against the variance in entropy and found strong correlations (Fig. 7, Spearman's correlation test: $\rho = -0.321$, $P = 1.526 \times 10^{-6}$ for WCN–ER, $\rho = 0.236$, $P = 4.746 \times 10^{-4}$ for RSA–ER). Proteins that had more variance in entropy across sites had larger structure–rate correlations in magnitude. Overall, enzymes were more diverged which in turn resulted, on average, in larger variances in entropy across proteins. The viral proteins were less diverged and as such had lower variances in site variability. However, even for the highly diverged enzymes, correlations with structural quantities were low unless the alignments showed high variation in site variability. Thus, structure–rate correlations are maximized at intermediate levels of divergence, where alignments are sufficiently diverged for a high dynamic range (both highly conserved and highly variable sites are present in the same alignment) but not overly saturated with divergence (so that all sites are highly diverged).

We also investigated the effect of alignment quality on the observed patterns. Highly diverged sequences are more difficult to align, and errors in multiple sequence alignments may propagate to yield spurious rate inferences at some sites. Such inferences may be partially responsible for the low structure–rate correlations for some proteins. To assess average alignment reliability, we calculated a reliability score using guidance[34,35] for each multiple sequence alignment. For each alignment, we calculated a column score (CS) at each site. CS scores range from 0, indicating an unreliably-aligned site, to 1, indicating a highly reliable alignment. We averaged the guidance CS for each multiple sequence alignment to obtain a mean guidance score representing the overall quality of an alignment. All of the viral proteins had scores greater than 0.98, indicating that these alignments had low uncertainty. The enzyme proteins had scores that span a very wide spectrum of quality, from 0 to 1. However, in enzymes, we found that the strength of structure–rate correlations was not correlated with alignment quality (Supporting Information Fig. S7, Spearman's correlation test: $\rho = -0.022$, $P = 0.746$ for WCN–ER, $\rho = -0.132$, $P = 0.057$ for RSA–ER). This finding suggests that alignment quality is not a significant factor in the observed strength of structure–rate correlations.

As a final test of the effect of divergence on structure–rate correlations, we obtained a series of more diverged viral alignments. Briefly, we used PSI–BLAST to obtain a set of homologous proteins for each of the viral proteins from Shahmoradi *et al*., using the UniProt90 database. This procedure was comparable to the procedure that had been used to assemble the enzyme alignments. Subsequently, we performed the same analysis using these alignments as we did on the other two datasets. Using this new methodology, we only managed to collect sufficient sequences to calculate meaningful evolutionary rates for three of the viral proteins (PDB IDs: 1RD8, 3GOL, and 3LYF). However, even though the dataset was small, we could compare it to the other two datasets for consistency. We found that the new viral dataset was more diverged than the original viral dataset but still less diverged than the enzyme dataset (Supporting Information Fig. S6). Despite this increased divergence in the new viral dataset, the strength of WCN–ER and RSA–ER correlations were similar to the original viral dataset. Additionally, the relationship between measures of divergence and the strength of structure–rate correlations was similar for both viral datasets (Fig. 7, Supporting Information Fig. S6). Even with the new approach it was difficult to obtain viral alignments with high divergence, which may be responsible for the lower structure–rate correlations still observed.

Discussion

The field of molecular evolution has a long history of attempting to identify the factors that affect the rate at which proteins evolve. At the level of whole-protein rates, some of the factors identified include expression level, interactions with other protein partners,[5,36–38] and selection for the costs of misfolding.[39] Recently, the emphasis has shifted towards explaining rate variation among sites within proteins, which seems to be driven primarily by biophysical, structural constraints.[10–15,22,40]

Among the structural constraints, packing density and relative solvent accessibility have emerged as the two best structural predictors of evolutionary rate.[10,11,13,15,20] Sites that are on the surface of the protein tend to evolve faster than sites in the protein interior. Similarly, sites that are densely packed and have more contacts tend to evolve slower and exhibit less sequence variability than sites with fewer contacts. However, how strongly these two structural quantities (solvent accessibility and local packing density) correlate with evolutionary rate at sites remains somewhat unclear.

Here we have examined the relationship between site variability and the strength of structure–rate relationships by performing a direct comparison of the enzyme dataset from Yeh *et al*.[11] and the viral proteins from Shahmoradi *et al*.[13] We have found that both WCN and RSA are significant predictors of ER in enzymes, with 37% of the variation in ER explained (on average) by WCN and 28% explained on average by RSA. In viral proteins, both quantities perform weaker, explaining on average 8 and 7% of variation in ER, respectively. Therefore, when analyzed using the same methods the datasets of Yeh *et al*.[11] and Shahmoradi *et al*.[13] both show that WCN performs better than RSA.

In addition to RSA and WCN, we have also considered a third predictor, protein design rate (DR). Protein design had previously been used in Shahmoradi *et al*.[13]. We have found that protein design rate is a much poorer predictor of rates at sites than RSA and WCN are. This result could represent a limitation in current methods of sequence space sampling techniques, limitations in the scoring function used in this study, or it could be that protein design rate does not capture biophysical forces that are predictive of evolutionary rates. For example, Ollikainen and Kortemme[41] published a study that examined the ability of protein design to capture naturally occurring covariation of amino acids at sites. Although flexible-backbone design was able to recapitulate some covariation from natural sequences, not all covariation could be explained by design, indicating that other forces besides structure could be involved in natural patterns of sequence covariation. Additionally, Jackson *et al*.[42] found that protein design did not recapture some important structure–sequence patterns observed in yeast proteins. Notably, in that study, designed proteins did not exhibit the same relationship between solvent accessibility and site variability observed in natural proteins and hydrophobic residues were often underrepresented in the protein core. These studies underscore the possibility that either current protein design methods are imperfect at mimicking natural structural constraints or that structural constraints do not capture all of the biophysical effects on sequence evolution.

In contrast to the rate predictors in the enzyme dataset, for the viral dataset, the structural predictors (RSA, WCN, or DR) all performed poorly. We have found that neither differences in structural features (WCN, RSA, or DR) nor differences in evolutionary rates are likely a driving factor in the difference in correlation strength. Therefore, we have investigated the possibility that there are fundamental differences in the two datasets themselves.

We have found that the lack of divergence within the viral proteins of the dataset taken from Shahmoradi *et al.*[13] is primarily responsible for the observed low structure–rate correlations. For a protein to have a high structure–rate correlation, there needs to be a high level of variability in divergence among the sites in the multiple–sequence alignment. In other words, a protein must have a combination of sites that are highly conserved and sites that are highly variable. If all sites in a protein are conserved or all sites are saturated with many substitutions, so that there is no variability within the multiple–sequence alignment, then structure–rate correlations will be low. This combination of highly conserved and highly variable sites will only occur when there is an intermediate level of divergence. This is also why absolute divergence has a much weaker relationship with the strength of structure–rate correlations as compared to variance of entropy. Although it is critical for a dataset to have sufficient divergence, it is only a necessary and not a sufficient requirement for strong structure–rate correlations. The enzyme dataset of Yeh *et al.*[11] has a variety of proteins with differing levels of divergence and, on average, has MSAs that are more diverged. The intermediate level of divergence in these enzymes results in larger structure–rate correlations.

In addition, variation in selection at sites within a protein can affect the strength of observed structure–rate correlations. Across a protein, structure may differentially affect site variability and hence the strength of structure–rate correlation strength varies. Selection against misfolding can constrain residues within the protein core while selection for key protein–protein interactions[43,44] and/or against nonspecific protein–protein interactions[45] may impact the variability seen on the protein surface. For example, important binding sites on the surface of the protein might be constrained decreasing the overall variability in variance of site variability. This would result in lower observed structure–rate correlations.

Although proteins as a whole exhibit common selective pressures, depending on the type of protein there might be additional factors that affect rate. Both viral proteins and enzymes exhibit some of the same selective pressures such as selection for stability and pressure to fold and adopt the correct native conformation. Enzymes are used to catalyze chemical reactions and as such have additional constraints such as structural constraints for a proper active site for catalytic function. On the other hand, viruses use their proteins to infect and replicate within their hosts. These proteins are utilized to perform a variety of necessary functions for viral replication such as host cellular entry[46,47] and nuclear importation.[48] As host immune systems attack these viruses, they evolve to escape from these host mechanisms resulting in signatures of positive selection within these proteins. Because of the differences in selective pressures facing these two protein types there might be different structural constraints on sequence variability and evolutionary rate.

We would like to emphasize that even though the distributions of average WCN and average RSA among proteins are similar for both datasets, there could be other structural differences among the proteins in the two datasets that might affect structure–rate correlations. Our purpose here was not to provide a rigorous, detailed analysis of structural differences among the two datasets. We only examined two obvious structural features (i.e., average packing of residues and average residue solvent accessibility) and showed that they are likely not the cause for the major discrepancy in correlation strengths among the two datasets. More sophisticated structural analyses may identify unique structural features among viral proteins,[33] and future research will have to determine whether these features have a measurable impact on structure–rate relationships. Furthermore, our results only apply to the two datasets discussed. Any additional general conclusions about the impact of divergence on observed structure–rate correlations in other systems would need further study.

Materials and Methods

Structures, sequences, and measures of sequence properties

The results presented in this work were based on two datasets. The first was a dataset of 208 monomeric enzymes, taken from Echave *et al.*[14] who reanalyzed the structures originally studied by Yeh *et al.*[11] The Echave *et al.*[14] dataset was slightly smaller than the original dataset because Echave *et al.*[14] removed proteins that had missing data at insertion sites. The dataset from Echave *et al.*[14] was originally comprised of 209 proteins but we removed one additional protein, 1CQQ, during our analysis (see below for details). Thus, our final enzyme dataset had 208 proteins. In brief, these proteins were all enzyme monomers randomly picked from the Catalytic Site Atlas 2.2.11.[49] Proteins in this dataset varied from 95 to 1287 residues in length. Each structure was accompanied by a multiple-sequence alignment of 300 homologous sequences. The second dataset was taken from Shahmoradi *et al*.[13] and consisted of nine viral proteins. The viral proteins ranged from 122 to 557 residues in length and each structure was accompanied by a multiple–sequence alignment of up to 2362 homologous sequences. Although both datasets vary in the number of sequence alignments, we did not enforce a medium number sequences in the multiple-sequence alignments needed to be included in the study since all alignments had at least 95 sequences.

Sequence alignments for both datasets were constructed by aligning the amino-acid sequences using the alignment software MAFFT,[28,29] specifying the auto flag to select the optimal algorithm for the given dataset. The alignments were then used to calculate site-specific measures of evolutionary rate for each individual protein in both datasets. We calculated a measure of site-specific evolutionary rate for each protein using the software Rate4Site.[26] First, maximum likelihood phylogenetic trees were inferred with RAxML, using the LG substitution matrix and the CAT model of rate heterogeneity.[50,51] For each structure, we used the respective sequence alignment and phylogenetic tree to infer site-specific substitution rates with Rate4Site, using the empirical Bayesian method and the JTT model of sequence evolution.[26]

Using the alignments, we also calculated the Shannon entropy (H_i), at each alignment column i:

$$H_\mathrm{i} = -\sum_{j} P_{ij} \ln P_{ij},$$

where P_{ij} was the relative frequency of amino acid j at position i in the alignment. Sequence entropy is a measure of variability at each site.

Finally, we calculated the divergence of each multiple-sequence alignment, using two measures: mean root-to-tip distance and mean patristic distance. Mean root-to-tip distance counts the average number of substitutions that have occurred along the tree. The mean patristic distance of an alignment was the average patristic distance of a tree where patristic distance was defined as the sum of the branch lengths between two nodes (i.e., sequences) within the tree.[52] Both root-to-tip distance and patristic distance were calculated using DendroPy.[53]

For the viral proteins we collected a second dataset. Using the sequences from the nine viral proteins from Shahmoradi *et al*.[13] as queries, we used PSI-BLAST[54] against the Uniprot90 to obtained homologous sequences for each protein. We used MAFFT and RaxML to create alignments and build trees for each protein. Trees could not be created for three of the proteins because their alignments did not have a sufficient number of sequences. We also chose to discard proteins from the analysis that did not have at least 25 sequences. This was done to guard against inaccurate rates. We calculated evolutionary rates for the remaining three proteins (PDB IDs: 1RD8, 3GOL, and 3LYF) using Rate4Site.

We quantified MSA reliability using a reimplementation of the Guidance platform[34] introduced by Spielman *et al*.[35] Guidance quantifies how robust MSA columns are to the guide tree topology used in during a progressive alignment algorithm. For each MSA column, Guidance produces a column score ranging from 0, indicating that the column is highly unreliable, to 1, indicating that the column is highly reliable. Note that the implementation in Spielman *et al*.[35] differs from that in Penn *et al*.[34] through its use of FastTree[55] to construct perturbed guidetrees. Here, Guidance was run with 100 bootstrap replicates using the MAFFT[28,29] alignment software, specifying the "auto" flag. We derived an overall guidance score for each MSA by averaging its resulting Guidance column scores.

Protein design

Using Rosetta,[56] we computationally designed 500 structures for select proteins in each dataset. For the viral proteins, we designed 500 structures for each of the proteins taken from Shahmoradi *et al*.[13] For the enzymes, we designed structures for each protein that was at most 200 residues in length. For each protein, we first designed 500 flexible ensembles using Backrub.[57] Backrub generates a set of flexible backbone "ensembles" onto which side chains can then be designed.[57,58] The Backub method takes a temperature parameter, *T*, that determines the extent of backbone flexibility during design. Higher temperatures allow for more backbone flexibility. Previous work has shown that moderate temperature parameters result in designed structures more similar to natural proteins.[41,42] Therefore, we used 0.6 as our temperature parameter. We then used the fixed-backbone method[59] to design side chains on these ensembles.

All designs were generated with Rosetta 3.5, 2014 week five release. To generate the series of ensembles using flexible-backbone design we used the following Rosetta commands:

```
./backrub -database rosetta_database \
   -s input.pdb -resfile NATAA.res -ex1 -ex2\
   -extrachi_cutoff 0 -backrub:mc_kt 0.6\
   -backrub:ntrials 10000 -nstruct 1
      -backrub:initial_pack
```

For the fixed-backbone design we used the following Rosetta commands:

```
./fixbb -database rosetta_database \
   -s input.pdb -resfile ALLAA.res -ex1 -ex2 \
   -extrachi_cutoff 0 -nstruct 1 -overwrite \
   -minimize_sidechains -linmem_ig 10
```

After design, we removed proteins that did not map back properly to the alignments. This resulted in the removal of one structure, 1CQQ, completely from the study. This resulted in a total of 32 enzymes in addition to the viral proteins.

Using the sequence alignments of designed proteins we predicted a sitewise rate, using the expression for dN proposed by Spielman and Wilke[32] (as implemented in the software Pyvolve[60]). For this calculation, we assumed that the mutation rate at all sites was equal. We called this quantity the "design rate" (DR) at sites.

Calculation of structural properties

In our analysis, we used side chain weighted contact number (WCN) as proposed by Marcos and Echave.[30] This quantity is defined as

$$\mathrm{WCN}_i = \sum_{i \neq j}^{N} \frac{1}{r_{ij}^2},$$

where r_{ij} is the distance between the geometric center of the side chain atoms of residue i and the geometric center of the side chain atoms of residue j, and N is the length of the protein. For glycine residues the distance to the C_α atom was used in lieu of the geometric center of the side chain.

To calculate relative solvent accessibility (RSA), we first calculated the accessible surface area (ASA) for each site in each protein, via DSSP.[61] We then normalized the ASA values by the theoretical maximum ASA values found in Table I of Tien *et al.*[17] All WCN and RSA calculations were done on the individual, monomeric protein chain of interest.

All data and analysis scripts required to reproduce the work are publicly available to view and download at https://github.com/wilkelab/rate_variability_variation.

ACKNOWLEDGMENT

The Texas Advanced Computing Center (TACC) at The University of Texas at Austin provided high-performance computing resources.

REFERENCES

1. Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, de Koning APJ, Dokholyan NV, Echave J, Elofsson A, Gerloff DL, Goldstein RA, Grahnen JA, Holder MT, Lakner C, Lartillot N, Lovell SC, Naylor G, Perica T, Pollock DD, Pupko T, Regan L, Roger A, Rubinstein N, Shakhnovich E, Sj—lander K, Sunyaev S, Teufel AI, Thorne JL, Thornton JW, Weinreich DM, Whelan S (2012) The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci 21:769–785.
2. Wilke CO, Drummond DA (2010) Signatures of protein biophysics in coding sequence evolution. Curr Opin Struct Biol 20:385–389.
3. Sikosek T, Chan HS (2014) Biophysics of protein evolution and evolutionary protein biophysics. J Roy Soc Interface 11:20140419.
4. Zhang J, Yang J-R (2015) Determinants of the rate of protein sequence evolution. Nat Rev Genet 16:409–420.
5. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296:750–752.
6. Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. Proc Natl Acad Sci USA 103:5869–5874.
7. Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352.
8. Liao H, Yeh W, Chiang D, Jernigan RL, Lustig B (2005) Protein sequence entropy is closely related to packing density and hydrophobicity. Protein Eng Des Select 18:59–64.
9. Serohijos AWR, Rimas Z, Shakhnovich EI (2012) Protein biophysics explains why highly abundant proteins evolve slowly. Cell Rep 2:249–256.
10. Franzosa EA, Xia Y (2009) Structural determinants of protein evolution are context-sensitive at the residue level. Mol Biol Evol 26:2387–2395.
11. Yeh S-W, Huang T-T, Liu J-W, Yu S-H, Shih C-H, Hwang J-K, Echave J (2014) Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. BioMed Res Int 2014: e572409.
12. Huang T-T, Marcos ML, del V, Hwang J-K, Echave J (2014) A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. BMC Evol Biol 14:78.
13. Shahmoradi A, Sydykova DK, Spielman SJ, Jackson EL, Dawson ET, Meyer AG, Wilke CO (2014) Predicting evolutionary site variability from structure in viral proteins: buriedness, packing, flexibility, and design. J Mol Evol 79:130–142.
14. Echave J, Jackson EL, Wilke CO (2015) Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. Phys Biol 12: 025002.
15. Yeh S-W, Liu J-W, Yu S-H, Shih C-H, Hwang J-K, Echave J (2014) Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. Mol Biol Evol 31:135–139.
16. Scherrer MP, Meyer AG, Wilke CO (2012) Modeling coding-sequence evolution within the context of residue solvent accessibility. BMC Evol Biol 12:179.
17. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO (2013) Maximum allowed solvent accessibilites of residues in proteins. PLoS One 8:e80635.
18. Goldman N, Thorne JL, Jones DT (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149:445–458.
19. Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function1. J Mol Biol 291:177–196.

20. 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.
21. Ramsey DC, Scherrer MP, Zhou T, Wilke CO (2011) The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics 188:479–488.
22. Franzosa EA, Xia Y (2012) Independent effects of protein core size and expression on residue-level structure–evolution relationships. PLoS One 7:e46602.
23. 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 1:216–226.
24. Lin C-P, Huang S-W, Lai Y-L, Yen S-C, Shih C-H, Lu C-H, Huang C-C, Hwang J-K (2008) Deriving protein dynamical properties from weighted protein contact number. Proteins 72:929–935.
25. Shih C-H, Chang C-M, Lin Y-S, Lo W-C, Hwang J-K (2012) Evolutionary information hidden in a single protein structure. Proteins 80:1647–1657.
26. Mayrose I, Graur D, Ben-Tal N, Pupko T (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol 21:1781–1791.
27. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18:S71–S77.
28. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066.
29. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518.
30. Marcos ML, Echave J (2015) Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. Peer J 3:e911.
31. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 97:10383–10388.
32. Spielman SJ, Wilke CO (2015) The relationship between dN/dS and scaled selection coefficients. Mol Biol Evol 32:1097–1108.
33. Tokuriki N, Oldfield CJ, Uversky VN, Berezovsky IN, Tawfik DS (2009) Do viral proteins possess unique biophysical features? Trends Biochem Sci 34: 53–59.
34. Penn O, Privman E, Landan G, Graur D, Pupko T (2010) An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol 27:1759–1767.
35. Spielman SJ, Dawson ET, Wilke CO (2014) Limited utility of residue masking for positive-selection inference. Mol Biol Evol 31:2496–2500.
36. Yang J-R, Liao B-Y, Zhuang S-M, Zhang J (2012) Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci USA 109: E831–E840.
37. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z (2005) Protein–protein docking benchmark 2.0: an update. Proteins 60:214–216.
38. Pang K, Cheng C, Xuan Z, Sheng H, Ma X (2010) Understanding protein evolutionary rate by integrating gene co-expression with protein interactions. BMC Syst Biol 4:179.
39. 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.
40. Echave J, Spielman SJ, Wilke CO (2016) Causes of evolutionary rate variation among protein sites. Nat Rev Genet 17:109–121.
41. Ollikainen N, Kortemme T (2013) Computational protein design quantifies structural constraints on amino acid covariation. PLoS Comput Biol 9:e1003313.
42. Jackson EL, Ollikainen N, Covert AW, Kortemme T, Wilke CO (2013) Amino-acid site variability among natural and designed proteins. Peer J 1:e211.
43. Elcock AH, McCammon JA (2001) Identification of protein oligomerization states by analysis of interface conservation. Proc Natl Acad Sci USA 98:2990–2994.
44. Valdar WSJ, Thornton JM (2001) Conservation helps to identify biologically relevant crystal contacts1. J Mol Biol 313:399–416.
45. Levy ED, De S, Teichmann SA (2012) Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. Proc Natl Acad Sci USA 109: 20461–20466.
46. Radoshitzky SR, Abraham J, Spiropoulou CF, Kuhn JH, Nguyen D, Li W, Nagel J, Schmidt PJ, Nunberg JH, Andrews NC, Farzan M, Choe H. (2007) Transferrin receptor 1 is a cellular receptor for New World haemorrhagic fever arenaviruses. Nature 446:92–96.
47. Allison AB, Kohler DJ, Ortega A, Hoover EA, Grove DM, Holmes EC, Parrish CR (2014) Host-specific parvovirus evolution in nature is recapitulated by *in vitro* adaptation to different carnivore species. PLoS Pathog 10:e1004475.
48. Schaller T, Ocwieja KE, Rasaiyaah J, Price AJ, Brady TL, Roth SL, Hué S, Fletcher AJ, Lee K, Kewal Ramani VN, Noursadeghi M, Jenner RG, James LC, Bushman FD, Towers GJ (2011) HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. PLoS Pathog 7:e1002439.
49. Porter CT, Bartlett GJ, Thornton JM (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32:D129–D133.
50. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.
51. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
52. Fourment M, Gibbs MJ (2006) PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. BMC Evol Biol 6:1.
53. Sukumaran J, Holder MT (2010) DendroPy: a Python library for phylogenetic computing. Bioinformatics 26: 1569–1571.
54. Altschul SF, Madden TL, Schäffer 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.
55. Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. Plos One 5:e9490.
56. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman KW, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandel DJ, Richter F, Andrew Ban Y-E, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S,

PopovicHavranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kulhman B, Baker D, Bradley P (2011) Chapter nineteen—Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 487:545–574.

57. Smith CA, Kortemme T (2008) Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J Mol Biol 380:742–756.

58. Smith CA, Kortemme T (2010) Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. J Mol Biol 402:460–474.

59. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302:1364–1368.

60. Spielman SJ, Wilke CO (2015) Pyvolve: a flexible python module for simulating sequences along phylogenies. PLoS One 10:e0139047.

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