Preprint

Predicting evolutionary site variability from structure in viral proteins: buriedness, flexibility, and design

Amir Shahmoradi · Dariya K. Sydykova · Stephanie J. Spielman · Eleisha L. Jackson · Eric T. Dawson · Austin G. Meyer · Claus O. Wilke

Abstract Several recent works have shown that protein structure can predict site-specific evolutionary sequence variation. In particular, sites that are buried and/or have many contacts with other sites in a structure have been shown to evolve more slowly, on average, than surface sites with few contacts. Here, we present a comprehensive study of the extent to which numerous structural properties can predict sequence variation. The structural properties we considered include buriedness (relative solvent accessibility and contact number), structural flexibility (B factors, root-mean-square fluctuations, and variation in dihedral angles), and variability in designed structures. We obtained structural flexibility measures both from molecular dynamics simulations performed on 9 non-homologous viral protein structures and from variation in homologous variants of those proteins, where available. We obtained measures of variability in designed structures from flexible-backbone design in the Rosetta software. We found that most of the structural properties correlate with site variation in the majority of structures, though the correlations are generally weak (correlation coefficients of 0.1 to 0.4). Moreover, we found that measures of buriedness tend to be better predictors of evolutionary variation than are measures of structural fluctuations. Finally, variability in designed structures was a weaker predictor of evolutionary variability than were both buriedness and fluctuation measures. We conclude that simple measures of buriedness are better predictors of evolutionary variation than are more complicated predictors obtained from dynamic simulations, ensembles of homologous structures, or computational protein design.

Introduction

Patterns of amino-acid sequence variation in protein-coding genes are shaped by the biophysical structure and function of the expressed proteins (Wilke and Drummond, 2010; Marsh and Teichmann, 2014). As the most basic reflection of this relationship, buried residues in proteins tend to be more evolutionarily conserved than exposed residues (Overington et al, 1992; Goldman et al, 1998; Mirny and Shakhnovich, 1999; Dean et al, 2002). More specifically, when evolutionary variation is plotted as a function of Relative Solvent Accessibility (RSA, a measure of residue buriedness), the relationship falls, on average, onto a straight line with a positive slope (Franzosa and Xia, 2009; Ramsey et al, 2011; Franzosa and Xia, 2012; Scherrer et al, 2012). Importantly, however, this relationship represents an average over many sites and many proteins. At the level of individual sites in individual proteins, RSA is often only weakly correlated with evolutionary variation (Meyer and Wilke, 2013; Meyer et al, 2013; Yeh et al, 2014).

Measures of residue buriedness other than RSA, such as residue contact number (CN) have also been shown to correlate with sequence variability (Liao et al, 2005; Franzosa and Xia, 2009; Yeh et al, 2014), and some have argued that CN predicts evolutionary variation better than RSA (Yeh et al, 2014). Because CN may be a proxy for residue and site-specific backbone flexibility (Halle, 2002), a positive trend between local structural variability and sequence variability may also exist (Yeh et al, 2014). Indeed, several authors have suggested that such protein dynamics may play a role in sequence variability (Liu

Amir Shahmoradi
Department of Physics, The University of Texas at Austin, TX, 78712.

Amir Shahmoradi · Dariya K. Sydykova · Stephanie J. Spielman · Eleisha L. Jackson · Eric T. Dawson · Austin G. Meyer · Claus O. Wilke
Department of Integrative Biology, Center for Computational Biology and Bioinformatics, and Institute for Cellular and Molecular Biology, The University of Texas at Austin, TX, 78712. E-mail: wilke@austin.utexas.edu

and Bahar, 2012; Nevin Gerek et al, 2013; Marsh and Teichmann, 2014). However, a recent paper argued against the flexibility model, on the grounds that it could introduce certain non-linearities that are not observed in the data (Huang et al, 2014).

While RSA and CN can be calculated in a straightforward manner from individual crystal structures, measures of structural flexibility, either at the side-chain or the backbone level, are more difficult to obtain. Two strong candidates for measuring structural flexibility are examining existing structural data or simulating protein dynamics. First, NMR ensembles and/or PDB structures may approximate *in vivo* physiologically relevant structural fluctuations well. Indeed, the thermal motion of atoms in a crystal are recorded in B factors, which are available for every atom in every PDB structure. It is also possible to align homologous PDB protein structures and subsequently assess the extent of fluctuations. Second, one could measure protein fluctuations using a simulation approach, either using coarse-grained modeling, e.g. via Elastic Network Models (Sanejouand, 2013), or using atom-level modeling, e.g. via molecular dynamics (MD) (Karplus and McCammon, 2002). However, it is not well understood which, if any, of the measures of structural flexibility discussed above provide insight into the evolutionary process, in particular residue-specific evolutionary variation.

Here, we provide a comprehensive analysis of the extent to which numerous different structural quantities predict evolutionary sequence (amino-acid) variation. We considered two measures of evolutionary sequence variation: site entropy, as calculated from homologous protein alignments, and evolutionary rate. For structural predictors, we included both measures of buriedness, including RSA and CN, and measures of structural flexibility, including B factors, several measures of backbone and side-chain variability obtained from MD simulations, and backbone variability obtained from alignments of homologous crystal structures. We additionally considered site variability, as predicted from computational protein design with Rosetta.

On a set of nine viral proteins, measures of buriedness generally performed better at predicting evolutionary site variation than did either measures of structural flexibility or computational protein design. Among the measures of structural flexibility, measures of side-chain variability performed better than do measures of backbone variability, possibly because the former are more tightly correlated with residue buriedness. Finally, site variability predicted from computational protein design performed worse than the best-performing measures of structural fluctuations.

Materials and Methods

Sequence data, alignments, and evolutionary rates

All viral sequences except influenza sequences were retrieved from `http://hfv.lanl.gov/components/sequence/HCV/search/searchi.html`. The sequences were truncated to the desired genomic region but not in any other way restricted. Influenza sequences were downloaded from `http://www.fludb.org/brc/home.spg?decorator=influenza`. We only considered human influenza A, H1N1, excluding H1N1 sequences derived from the 2009 Swine Flu outbreak or any sequence from before 1998, but we did not place any geographic restrictions.

For all viral sequences, we removed any sequence that was not in reading frame, any sequence which was shorter than 80% of the longest sequence for a given viral protein (so as to remove all partial sequences), and any sequence containing any ambiguous characters. Alignments were constructed using amino-acid sequences with MAFFT (Katoh et al, 2002, 2005), specifying the `--auto` flag to select the optimal algorithm for the given data set, and then back-translated to a codon alignment using the original nucleotide sequence data.

To assess site-specific sequence variability in amino-acid alignments, we calculated the Shannon entropy (H_i) at each alignment column i:

$$H_i = -\sum_j P_{ij} \ln P_{ij}, \tag{1}$$

where P_{ij} is relative frequency of amino acid j at position i in the alignment.

For each alignment, we also calculated evolutionary rates, as described (Spielman and Wilke, 2013). In brief, we generated a phylogeny for each codon alignment in RAxML (Stamatakis, 2006) using the GTRGAMMA model. Using the codon alignment and phylogeny, we inferred evolutionary rates with a Random Effects Likelihood (REL) model, using the HyPhy software (Kosakovsky Pond et al, 2005). The REL model was a variant of the GY94 evolutionary model (Goldman and Yang, 1994) with five ω rate

categories as free parameters. We employed an Empirical Bayes approach (Yang, 2000) to infer ω values for each position in the alignment. These ω values represent the evolutionary-rate ratio dN/dS at each site.

Protein crystal structures

A total of 9 viral protein structures were selected for analysis, as tabulated in Table 1. Sites in the PDB structures were mapped to sites in the viral sequence alignments via a custom-built python script that creates a consensus map between a PDB sequence and all sequences in an alignment.

For each of the viral proteins, homologous structures were identified using the `blast.pdb` function of the R package Bio3D (Grant et al, 2006). BLAST hits were retained if they had $\geq 35\%$ sequence identity and $\geq 90\%$ alignment length. Among the retained hits, we subsequently identified sets of homologous structures with unique sequences and with mutual pairwise sequence divergences of $\geq 2\%$, $\geq 5\%$, and $\geq 10\%$.

Molecular Dynamics Simulations

Molecular dynamics (MD) simulations were carried out using the GPU implementation of the *Amber12* simulation package (Salomon-Ferrer et al, 2013) with the most recent release of the Amber fixed-charge force field (ff12SB; c.f., AmberTools13 Manual). Prior to MD production runs, all PDB structures were first solvated in a box of TIP3P water molecules (Jorgensen et al, 1983) such that the structures were at least $10\mathring{A}$ away from the box walls. Each individual system was then energy minimized using the steepest descent method for 1000 steps, followed by conjugate gradient for another 1000 steps. Then, the structures were constantly heated from 0K to 300K for 0.1ns, followed by 0.1ns constant pressure simulations with positional harmonic restraints on all atoms to avoid instabilities during the equilibration process. The systems were then equilibrated for another 5ns without positional restraints, each followed by 15ns of production simulations for subsequent post-processing and analyses. All equilibration and production simulations were run using the SHAKE algorithm (Ryckaert et al, 1977). Langevin dynamics were used for temperature control.

Measures of buriedness and of structural flexibility

For measures of residue buriedness, we calculated Relative Solvent Accessibility (RSA), contact number (CN), and weighted contact number (WCN). To calculate RSA, we first calculated the Accessible Surface Area (ASA) for each residue in each protein, using the DSSP software (Kabsch and Sander, 1983). We then normalized ASA values by the theoretical maximum ASA of each residue (Tien et al, 2013) to obtain RSA. We calculated CN for each residue as the total number of Cα atoms surrounding the Cα atom of the focal residue within a spherical neighborhood of a predefined radius r_0. Following Yeh et al (2014), we used $r_0 = 13\mathring{A}$. We calculated WCN as the total number of surrounding Cα atoms for each focal residue, weighted by the inverse square separation between the Cα atoms of the focal residue and the contacting residue, respectively (Shih et al, 2012).

In most analyses, we actually used the inverse of CN and/or WCN, iCN = 1/CN and iWCN = 1/WCN. Note that for Spearman correlations, which we use throughout here, replacing a variable by its inverse changes the sign of the correlation coefficient but not the magnitude.

For measures of structural flexibility, we considered RMSF, variability in backbone and side-chain dihedral angles, and B factors. We calculated RMSF for backbone Cα atoms based on both MD trajectories and homologous crystal structures. For MD trajectories, we calculated RMSF as

$$\mathrm{RMSF}_j = \Big[\sum_i \big(\mathbf{r}_i^{(j)} - \mathbf{r}_0^{(j)} \big)^2 \Big]^{1/2} \tag{2}$$

where RMSF_j is the root-mean-square fluctuation at site j, $\mathbf{r}_i^{(j)}$ is the position of the Cα atom of residue j at MD frame i, and $\mathbf{r}_0^{(j)}$ is the position of the Cα atom of residue j in the original crystal structure.

To calculate RMSF from homologous structures, we first aligned the structures using the Bio3D package (Grant et al, 2006), and then we calculated

$$\mathrm{RMSF}_j = \Big[\sum_i w_i \big(\mathbf{r}_i^{(j)} - \langle \mathbf{r}^{(j)} \rangle\big)^2\Big]^{1/2}, \tag{3}$$

where $\mathbf{r}_i^{(j)}$ now stands for the position of the Cα atom of residue j in structure i, $\langle \mathbf{r}^{(j)} \rangle$ is the mean position of that Cα atom over all aligned structures, and w_i is a weight to correct for potential phylogenetic relationship among the aligned structures. The weights w_i were calculated using BranchManager (Stone and Sidow, 2007), based on phylogenies built with RAxML as before.

To assess variability in backbone and side-chain dihedral angles, we calculated $\mathrm{Var}(\phi)$, $\mathrm{Var}(\psi)$, and $\mathrm{Var}(\chi_1)$. The variance of a dihedral angle was defined according to the most common definition in directional statistics: First, a unit vector $\mathbf{x}_i$ is assigned to each dihedral angle α_i in the sample. The unit vector is defined as $\mathbf{x}_i = (\cos(\alpha_i), \sin(\alpha_i))$. The variance of the dihedral angle is then defined as

$$\mathrm{Var}(\alpha) = 1 - ||\langle \mathbf{x} \rangle||, \tag{4}$$

where $||\langle \mathbf{x} \rangle||$ represents the length of the mean $\langle \mathbf{x} \rangle$, calculated as $\langle \mathbf{x} \rangle = \sum_i \mathbf{x}_i / n$. Here, n is the sample size. The variance of a dihedral angle is, by definition, a real number in the range $[0, 1]$, with $\mathrm{Var}(\alpha) = 0$ corresponding to the minimum variability of the dihedral angle and $\mathrm{Var}(\alpha) = 1$ to the maximum, respectively (Berens, 2009). Since the χ_1 angle is undefined for Ala and Gly we excluded all sites with these residues in analyses involving χ_1.

B factors were extracted from the crystal structures. We only considered the B factors of the Cα atom of each residue.

Sequence Entropy from Designed Proteins

Designed entropy was calculated as described (Jackson et al, 2013). In brief, proteins were designed using RosettaDesign (Version 39284) (Leaver-Fay et al, 2011) using a flexible backbone approach. This was done for all PDB structures in Table 1 as initial template structures. For each template, we created a backbone ensemble using the Backrub method (Smith and Kortemme, 2008). The temperature parameter in Backrub was set to 0.6, allowing for an intermediate amount of flexibility. For each of the 9 template structures we designed 100 proteins.

Availability of data and methods

All details of simulations, input/output files, and scripts for subsequent analyses are available to view or download at `https://github.com/clauswilke/structural_prediction_of_ER`.

Results

Data set and structural variables considered

Our goal in this work was to determine which structural properties best predict amino-acid variability at individual sites in viral proteins. To this end, we selected 9 viral proteins for which we had both high-quality crystal structures and abundant sequences to assess evolutionary sequence variation (Table 1). We quantified evolutionary variability in two ways: by calculating sequence entropies for each alignment column, and by calculating site-specific evolutionary-rate ratios $\omega = dN/dS$ (see Methods for details). Throughout this paper, we primarily report results obtained for sequence entropy. Results for ω were largely comparable, with some specific caveats detailed below.

As predictors of evolutionary variability, we considered two broad classes of structural properties: residue buriedness and residue flexibility. We additionally considered the variation seen in computationally designed protein variants. Measures of buriedness quantify the extent to which a residue is protected from solvent. Here, we have considered the buriedness measures relative solvent accessibility (RSA), contact number (CN), and weighted contact number (WCN). Both CN and WCN assess the number of other residues a focal residue contacts, but CN simply counts the number of contacts within a sphere of a given radius around the α-carbon of the focal residue. WCN, on the other hand, weights contacts by

the distance between the two residues. Note that while residue buriedness decreases as RSA increases, buriedness increases as CN or WCN increased. To avoid this difference in directionality, we replaced CN and WCN with their inverse, iCN = 1/CN and iWCN = 1/WCN, in most analyses. Importantly, as Spearman rank correlations were used, this substitution only changed the sign of correlations but not the magnitude.

Measures of structural flexibility assess the extent to which a residue fluctuates in space as a protein undergoes thermodynamic fluctuations in solution. We quantified these fluctuations using several different measures. We considered B factors, which measure the spatial localization of individual atoms in a protein crystal, RMSF, the root mean-square fluctuation of the Cα atom over time, and variability in side-chain and backbone dihedral angles, including Var(χ_1), Var(ϕ), and Var(ψ). We employed two broad approaches, one using PDB crystal structures and one using molecular dynamics (MD) simulation, to retrieve these measurements. Crystal structures yielded measures for B factors and RMSF; we obtained B factors from individual protein crystal structures, given in Table 1, and we calculated RMSF from aligned homologous crystal structures for those proteins which had sufficient sequence variation among crystal structures (see Methods and Table 2 for details). MD simulations yielded measures for RMSF and variability in residue dihedral and side-chain angles. More specifically, we simulated MD trajectories for all crystal structures in Table 1. For each protein, we first equilibrated the structure and then simulated 15ns of chemical time and recorded snapshots of the simulated structure every 10ps (see Methods for details). We obtained RMSF and angle variabilities from these snapshots. Additionally, we calculated time-averaged values of the three measures of buriedness, RSA, CN, and WCN. We refer to these time-averaged buriedness measures as MD RSA, MD CN, and MD WCN, respectively.

To assess variability in designed proteins, we used the Rosetta protein-design platform to generate 100 designed variants of each protein structure listed in Table 1. We then calculated the sequence entropy at each position in alignments of the designed variants. We refer to the resulting quantity as the *designed entropy*. We chose to include this quantity because previous work had shown that designed entropy captures some amount of evolutionary sequence variation observed in natural alignments (Jackson et al, 2013).

Evaluating structural predictors of evolutionary sequence variation

We began by comparing the Spearman correlations of sequence entropy with six different measures of local structural flexibility, including B factors, RMSF obtained from MD simulations (MD RMSF), and RMSF obtained from crystal structures (CS RMSF), and variability in backbone and side-chain dihedral angles (ϕ, ψ, and χ_1). The correlation strengths of these quantities with entropy are shown in Figure 1. Significant correlations ($P < 0.05$) are shown with filled symbols, and non-significant correlations are shown with empty symbols ($P \geq 0.05$). We found that the variability in backbone dihedral angles, Var(ϕ) and Var(ψ), explained the least variation in sequence entropy, while the variability in the side-chain dihedral angle Var(χ_1) explained, on average, more variation in sequence entropy than did any other measure of structural flexibility. B factors and the two measures of RMSF explained on average approximately the same amount of variation in entropy, even though the results for individual proteins were somewhat discordant (see also next sub-section).

Based on results from the above analysis, we proceeded to compare the relative explanatory power among the best-performing measures of structural flexibility (Var(χ_1), MD RMSF, and B factors), buriedness measures (RSA and iWCN), and designed entropy. Figure 2 shows the Spearman correlation coefficients between sequence entropy and each of the aforementioned quantities, for all proteins in our analysis. In this figure, several patterns emerge. First, nearly all correlations were positive and most were statistically significant, with the main exception of the Marburg virus RNA binding domain (PDB ID 4GHA). This protein only showed a single significant negative correlation between sequence entropy and χ_1 angle variability. Second, correlations were generally weak, such that no correlation coefficient exceeds 0.4. Third, on average, correlations were strongest for measures of residue buriedness, with RSA and iWCN yielding average correlations of $\rho = 0.23$ and $\rho = 0.22$, respectively. Fourth, designed entropy performed worse than measures of buriedness as a predictor of evolutionary sequence variability, but it performed roughly the same as the three flexibility measures in this figure; the values of designed entropy, Var(χ_1), MD RMSF, and B factors showed average correlations of $\rho = 0.13$, $\rho = 0.14$, $\rho = 0.11$, and $\rho = 0.12$, respectively.

MD time-averages vs. crystal-structure snapshots

Except for analyses involving B factors and CS RMSF, we obtained structural measures by averaging quantities over MD trajectories, comprising 15ns of chemical time each. This approach, however, did not reflect conventional practice for measuring buriedness quantities, which are typically measured from individual crystal structures. Therefore, we examined whether MD time-averages differed in any meaningful way from estimates obtained from crystal structures, and whether these estimates differed in their predictive power for evolutionary sequence variation.

As shown in Table 3, RSA, CN, and WCN from crystal structures were highly correlated with their corresponding MD trajectory time-averages, for all protein structures we examined (Spearman correlation coefficients of > 0.9 in all cases). Furthermore, the correlation coefficients we obtained when comparing the crystal structure buriedness measures to sequence entropy were virtually identical to coefficients obtained from the MC trajectory correlations (Figure 3A-C). Thus, in terms of predicting evolutionary variation, RSA, CN and WCN values obtained from static structures performed as well as their MD equivalents averaged over short time scales. By contrast, correlations between corresponding MD RMSF to CS RMSF measures were sometimes quite different, with correlation coefficients ranging from 0.218 to 0.723 (Table 3). Consequently, for the two proteins for which MD RMSF was the least correlated with crystal-structure RMSF (hepatitis C protease and Rift Valley fever nucleoprotein), the strength of correlation between site entropy and RMSF depended substantially on how RMSF was calculated (Figures 1 and 3D).

Finally, we examined whether correlations between sequence entropy and B factors or the two RMSF measures were comparable (Figure 4). Again, we found that correlations between sequence entropy and B factors were generally different from those obtained for both MD RMSF and CS RMSF. This result highlighted that, while B factors, MD RSMF, and MD RMSF all measure backbone flexibility, they each contain distinct information about evolutionary sequence variability in our data set.

Sequence entropy vs. evolutionary-rate ratio ω

In the previous subsections, we used sequence entropy as a measure of site-wise evolutionary variation. While sequence entropy is a simple and straightforward measure of site variability, it has two potential drawbacks. First, while measured from homologous protein alignments, sequence entropy doesn't correct for the phylogenetic relationship of those alignment sequences. Hence, entropy can be biased if some parts of the phylogeny are more densely sampled than others. Second, entropy does not take the actual substitution process into account. As a result, a single substitution near the root of the tree can result in a comparable entropy to a sequence of substitutions toggling back and forth between two amino acids.

To consider an alternative quantity of evolutionary variation that doesn't suffer from either of these drawbacks, we calculated the evolutionary-rate ratio $\omega = dN/dS$ for all proteins at all sites, and repeated all analyses with ω instead of entropy. We found that results generally carried over, but with somewhat weaker correlations. Figure 5 plots, for each protein, the Spearman correlations between ω and the various structural quantities and designed entropy versus the correlation between entropy and the same quantities. Most data points fall below the $x = y$ line and are shifted downwards by approximately 0.1. Thus, correlations of structural quantities and designed entropy with ω are, on average, approximately 0.1 smaller than correlations of the same quantities with sequence entropy.

Multi-variate analysis of structural predictors

The various structural quantities we have considered are by no means independent of each other. Measures of buriedness co-vary with each other, as do measures of structural flexibility. Further, the latter co-vary with the former, as does designed entropy. Therefore, we conducted a joint multivariate analysis, which included most structural quantities considered in this work. We employed this strategy to determine the extent to which these quantities contain independent information about sequence variability while additionally assessing whether combining multiple structural quantities yields improved predictive power. We employed a principal component (PC) regression approach, which has previously been used successfully to disentangle genomic predictors of whole-protein evolutionary rates (Drummond et al, 2006; Bloom et al, 2006). For each PC analysis described below, we first carried out a PC analysis of the predictor variables (i.e., the structural quantities such as RSA and RMSF), and we subsequently regressed the response (either sequence entropy or ω) against the individual components.

For a first PC analysis, we pooled all structural quantities and then regressed entropy against each PC separately, for all proteins in our data set. This strategy allowed us to analyze all proteins in our data set individually but in such a way that results were comparable from one protein to the next. Note that we excluded CS RMSF from this analysis in order to include results from all nine viral proteins. The results of this analysis are shown in Figure 6. The first component (PC1) explained, on average, the largest amount of variation in sequence entropy (see Figure 6A). PC3 yielded the second-highest r^2 values, on average, while all other components explained very little variation in sequence entropy. When looking at the composition of the components, we found that RSA, iWCN, RMSF, and Var(χ_1) all loaded strongly on PC1, while PC2 and PC3 where primarily represented by designed entropy and B factors (see Figure 6B and C). RMSF also had moderate loadings on PC3. Interestingly, designed entropy and B factors load with equal signs on PC2 but with opposite signs on PC3.

We interpret PC1 as a buriedness component. By definition, PC1 measures the largest amount of variation among the structural quantities, and all structural quantities reflect to some extent the buriedness of residues. PC2 and PC3 are more difficult to interpret. Since designed entropy and B factors load strongly on both but with two different combinations of signs, we think the most parsimonious interpretation is to consider PC2 as a component representing sites with high designed entropy and high spatial fluctuations (as measured by B factors) and PC3 representing sites with high designed entropy and low spatial fluctuations. Using these interpretations, our PC regression analysis suggests that of all the structural quantities, residue buriedness is the best predictor of evolutionary variation. Designed entropy is a useful predictor as well, but it tends to perform better at sites with low spatial fluctuations.

For a second PC analysis, we included the predictor CS RMSF, which therefore restricted the data set to include only six proteins (see Table 2). This analysis, which retained sequence entropy as the response variable, yielded comparable results to the first PC analysis. The main differences occurred in PC2 and PC3, where CS RMSF generally loaded in the opposite direction of B factor, and either in the same (PC2) or the opposite (PC3) direction of designed entropy (Figure S1).

Finally, we redid the two PC analyses described above but instead used ω as the response variable (Figures S2 and S3). Again, these results were largely comparable to results from PC analyses with sequence entropy as the response.

Discussion

We have carried out a comprehensive analysis of the extent to which different structural quantities predict sequence evolutionary variation in nine viral proteins. We found that measures of buriedness generally performed better than did measures of structural flexibility. Further, measures of buriedness also performed better than a computational protein-design approach that employed a sophisticated all-atom force field to determine allowed amino-acid distributions at each site. Finally, averaging structural quantities over 15ns of MD simulations provided no improvement in predictive power relative to taking the same quantities from individual crystal structures.

Our results are broadly in agreement with recent work by Echave and collaborators (Yeh et al, 2014; Huang et al, 2014). These authors found that RSA and CN showed comparable correlation strengths with evolutionary sequence variation (Yeh et al, 2014). Further, they demonstrated that the observed relationship between evolutionary variation and residue–residue contacts was not consistent with a flexibility model that puts evolutionary variability in proportion to structural flexibility (Huang et al, 2014). Instead, a mechanistic stress model, in which amino-acid substitutions cause physical stress in proportion to the number of residue–residue contacts affected, could explain all the observed data (Huang et al, 2014).

The correlation strengths we observed were consistently lower than those observed previously (Jackson et al, 2013; Yeh et al, 2014). We believe that this result was due to our choice of analyzing viral proteins instead of the cellular proteins or enzymes used in prior works. First, while viral sequences are abundant, their alignments may not be as diverged as alignments that can be obtained for sequences from cellular organisms. For example, our influenza sequences spanned only approximately one decade. Despite the high mutation rates observed in RNA viruses, the evolutionary variation that can accumulate over this time span is limited. This relatively lower evolutionary divergence makes resolving differences between more and less conserved sites much more difficult. Second, many viral proteins experience a substantial amount of selection pressure to evade host immune responses. The resulting positive selection on viral sequences may mask evolutionary constraints imposed by structure. For example, influenza hemagglutinin displays positive selection throughout the entire sequence, regardless of the extent of residue burial (Meyer and Wilke, 2013; Meyer et al, 2013; Suzuki, 2006; Bush et al, 1999).

We have found here that correlations between sequence entropy and structural quantities were consistently higher than correlations between the evolutionary-rate ratio ω and structural quantities. One possible explanation is again our choice of viral sequences. These sequences almost certainly contain some polymorphisms, which may diminish the reliability of ω estimates (Kryazhimskiy and Plotkin, 2008). While the effect of polymorphisms on sequence entropy is not known, it seems plausible that entropy would be less sensitive to them. Alternatively, the difference between entropy and ω may reflect the distinct physical processes that entropy and ω measure. Entropy is a measure of the amino-acid diversity allowed at a given site, representing how many different amino acids are tolerated. By contrast, ω measures how rapidly amino-acid changes occur at a given site. While entropy and ω are generally correlated, a site can have high entropy and comparatively low ω, and vice versa. In particular, if substitutions rapidly toggle back and forth between two different amino acids at a site, then that site will have high ω and low entropy. By contrast, if a site diversified into a number of different alleles deep in the phylogeny but did not experience much subsequent evolution, that site will have comparatively high entropy but low ω. Since structural quantities, such as measures of buriedness, reflect the biophysical constraints imposed on sites, one could surmise that they would be better predictors of the allowed amino-acid diversity at a site than the speed of substitution at that site. By contrast, biological processes, rather than structural quantities, such as immune escape would likely be better predictors of substitution rates than of amino-acid diversity.

We found that simple measures of buriedness, such as RSA and CN, were better predictors of evolutionary variation than was sequence variability predicted from computational protein design. In other words, simple quantities that can be obtained trivially from PDB structures performed better than a sophisticated protein-design strategy that makes use of an all-atom energy function and requires thousands of CPU-hours to complete. This result highlights that, even though computational protein design has yielded impressive results in specific cases (Kuhlman et al, 2003; Röthlisberger et al, 2008; Fleishman et al, 2011), this approach remains limited in its ability to predict evolutionary variation. Similarly, we have previously found that flexible backbone design with Rosetta produced designs whose surface and core were too similar (Jackson et al, 2013). We attributed this discrepancy to either the solvation model or the model of backbone flexibility we used (Backrub, see Smith and Kortemme 2008). The results we found here suggest that the model of backbone flexibility may indeed be the cause of at least some of the discrepancies between predicted and observed site variability. In particular, in our PC regression analysis, the component in which designed entropy loaded opposite to B factor and MD RMSF generally had the second-highest predictive power for evolutionary variability, after the component representing buriedness. In sum, designed entropy was a better predictor for evolutionary sequence variability for sites with less structural flexibility compared to sites with more flexibility.

Even though RSA and CN remain the best currently known predictors of evolutionary variation, neither quantity has particularly high predictive power. One reason why predictive power may be low is that neither quantity accounts for correlated substitutions at interacting sites. Yet such correlated substitutions happen regularly. For example, covariation among sites encodes information about residue-residue contacts and 3D structure (Halabi et al, 2009; Burger and van Nimwegen, 2010; Marks et al, 2011; Jones et al, 2014), and evolutionary models that incorporate residue–residue interactions tend to perform better than models that do not (Rodrigue et al, 2005; Bordner and Mittelmann, 2014). An improved predictor of evolutionary variation would have to correctly predict this covariation from structure. In principle, computational protein design, which takes into consideration the atom-level details of the protein structure, should properly reproduce covariation among sites. However, a recent analysis showed that there are significant limitations to the covariation that is predicted (Ollikainen and Kortemme, 2013). In addition, covariation in designed proteins is quite sensitive to the type of backbone variation modeled during design, and improved models of backbone flexibility may be required for improved prediction of covariation among sites (Ollikainen and Kortemme, 2013).

Acknowledgments

This work was supported in part by NIH grant R01 GM088344, DTRA grant HDTRA1-12-C-0007, ARO grant W911NF-12-1-0390, and the BEACON Center for the Study of Evolution in Action (NSF Cooperative Agreement DBI-0939454). The Texas Advanced Computing Center at UT Austin provided high-performance computing resources.

References

Berens P (2009) CircStat: a MATLAB toolbox for circular statistics. J Stat Software 31:1–21

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

Bordner AJ, Mittelmann HD (2014) A new formulation of protein evolutionary models that account for structural constraints. Mol Biol Evol 31:736–749

Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol 6:e1000,633

Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999) Predicting the evolution of human influenza A. Science 286:1921–1925

Dean AM, Neuhauser C, Grenier E, Golding GB (2002) The pattern of amino acid replacements in α/β-barrels. Mol Biol Evol 19:1846–1864

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

Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, Baker D (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332:816–821

Franzosa EA, Xia Y (2009) Structural determinants of protein evolution are context-sensitive at the residue level. Mol Biol Evol 26:2387–2395

Franzosa EA, Xia Y (2012) Independent effects of protein core size and expression on residue-level structure-evolution relationships. PLoS ONE 7:e46,602

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

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

Grant BJ, Rodrigues APC, ElSawy KM, McCammon AJ, Caves LSD (2006) Bio3D: an R package for the comparative analysis of protein structures. Bioinformatics 22:2695–2696

Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: Evolutionary units of three-dimensional structure. Cell 138:774–786

Halle B (2002) Flexibility and packing in proteins. Proc Natl Acad Sci USA 99:1274–1279

Huang TT, del Valle Marcos ML, Hwang JK, 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

Jackson EL, Ollikainen N, III AWC, Kortemme T, Wilke CO (2013) Amino-acid site variability among natural and designed proteins. PeerJ 1:e211

Jones DT, Buchan DWA, Cozzetto D, Pontil M (2014) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Mol Biol Evol 31:736–749

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics 79(2):926–935, DOI 10.1063/1.445869, URL `http://scitation.aip.org/content/aip/journal/jcp/79/2/10.1063/1.445869`

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

Karplus M, McCammon A (2002) Molecular dynamics simulations of biomolecules. Nature Struct Biol 9:646–652

Katoh K, Misawa K, Kuma KI, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl Acids Res 30:3059–3066

Katoh K, Kuma KI, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucl Acids Res 33:511–518

Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenetics. Bioinformatics 21:676–679

Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. PLoS Genet 4:e1000,304

Kuhlman B, Dantas G, Ireton G, Gabriele V, Stoddard B (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302:1364–1368

Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew DP, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban YEA, Fleishman SJ, Corn

JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popović Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods in Enzymology 487:545–574

Liao H, Yeh W, Chiang D, Jernigan RL, Lustig B (2005) Protein sequence entropy is closely related to packing density and hydrophobicity. PEDS 18:59–64

Liu Y, Bahar I (2012) Sequence evolution correlates with structural dynamics. Mol Biol Evol 29:2253–2263

Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS ONE 6:e28,766

Marsh JA, Teichmann SA (2014) Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. BioEssays 36:209–218

Meyer AG, Wilke CO (2013) Integrating sequence variation and protein structure to identify sites under selection. Mol Biol Evol 30:36–44

Meyer AG, Dawson ET, Wilke CO (2013) Cross-species comparison of site-specific evolutionary-rate variation in influenza haemagglutinin. Phil Trans R Soc B 368:20120,334

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

Nevin Gerek Z, Kumar S, Banu Ozkan S (2013) Structural dynamics flexibility informs function and evolution at a proteome scale. Evolutionary Applications 6:423–433

Ollikainen N, Kortemme T (2013) Computational protein design quantifies structural constraints on amino acid covariation. PLoS Comput Biol 9:e1003,313

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

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

Rodrigue N, Lartillot N, Bryant D, Philippe H (2005) Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene 347:207–217

Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D (2008) Kemp elimination catalysts by computational enzyme design. Nature 453:190–195

Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comp Phys 23:327–341

Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC (2013) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J Chem Theory Comput 9:3878–3888

Sanejouand YH (2013) Elastic network models: theoretical and empirical foundations. Methods Mol Biol 924:601–616

Scherrer MP, Meyer AG, Wilke CO (2012) Modeling coding-sequence evolution within the context of residue solvent accessibility. BMC Evol Biol Submitted

Shih CH, Chang CM, Lin YS, Lo W, Hwang JK (2012) Evolutionary information hidden in a single protein structure. Proteins: Structure, Function, and Bioinformatics 80:1647–1657

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

Spielman SJ, Wilke CO (2013) Membrane environment imposes unique selection pressures on transmembrane domains of G protein-coupled receptors. J Mol Evol 76:172–182

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690

Stone EA, Sidow A (2007) Constructing a meaningful evolutionary average at the phylogenetic center of mass. BMC Bioinformatics 8:222

Suzuki Y (2006) Natural selection on the influenza virus genome. Mol Biol Evol 23:1902–1911

Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO (2013) Maximum allowed solvent accessibilites of residues in proteins. PLOS ONE 8:e80,635

Wilke CO, Drummond DA (2010) Signatures of protein biophysics in coding sequence evolution. Cur Opin Struct Biol 20:385–389

Yang Z (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J Mol Evol 51:423–432

Yeh SW, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J (2014) Site-specific structural constraints on protein sequence evolutionary divergence: Local packing density versus solvent exposure. Mol Biol Evol

31:135–139

Tables

Table 1 PDB structures considered in this study.

Viral Protein	PDB ID	Chain	Sequence Length	Number of Sequences
Hemagglutinin Precursor	1RD8	AB	503	1039
Dengue Protease Helicase	2JLY	A	451	2362
West Nile Protease	2FP7	B	147	237
Japanese Encephalitis Helicase	2Z83	A	426	145
Hepatitis C Protease	3GOL	A	557	1021
Rift Valley Fever Nucleoprotein	3LYF	A	244	95
Crimean Congo Nucleocapsid	4AQF	B	474	69
Marburg RNA Binding Domain	4GHA	A	122	42
Influenza Nucleoprotein	4IRY	A	404	943

Table 2 Availability of homologous crystal structures. Although most viral proteins have many PDB structures available, the sequence divergence among these structures is low. Therefore, when calculating RMSF from crystal structures, we considered only those proteins with at least five homologous structures at 5% pairwise sequence divergence (highlighted in bold).

Viral Protein	BLAST hits[a]	Unique sequences			
		all	$\geq 2\%$[b]	$\geq 5\%$[b]	$\geq 10\%$[b]
Hemagglutinin Precursor	63	17	10	**9**	7
Dengue Protease Helicase	31	13	7	**7**	7
West Nile Protease	21	16	10	**7**	6
Japanese Encephalitis Helicase	31	12	7	**7**	7
Hepatitis C Protease	302	33	10	**5**	4
Rift Valley Fever Nucleoprotein	95	9	5	**5**	5
Crimean Congo Nucleocapsid	7	4	3	2	2
Marburg RNA Binding Domain	63	9	5	3	3
Influenza Nucleoprotein	69	15	4	4	2

[a] BLAST hits against all sequences in the PDB, excluding hits with $< 35\%$ sequence identity and $< 90\%$ alignment length
[b] Unique sequences at indicated minimum pairwise sequence divergence

Table 3 Correlations between quantities obtained from MD trajectories and from crystal structures. For each quantity and each protein, we calculated the Spearman correlation ρ between the values obtained from MD time averages and the values obtained from viral protein crystal structures. Note that crystal structures for all nine proteins were used for RSA, CN, and WCN calculations, but only the six proteins for which we had sufficient crystal structure variability were used for CS RMSF. We then calculated the minimum, maximum, mean, and standard deviation of these correlations.

Quantity	min ρ	max ρ	$\langle\rho\rangle$	SD(ρ)
RSA	0.937	0.981	0.948	0.012
CN	0.964	0.993	0.976	0.008
WCN	0.973	0.991	0.984	0.006
RMSF	0.218	0.723	0.502	0.181

Figures

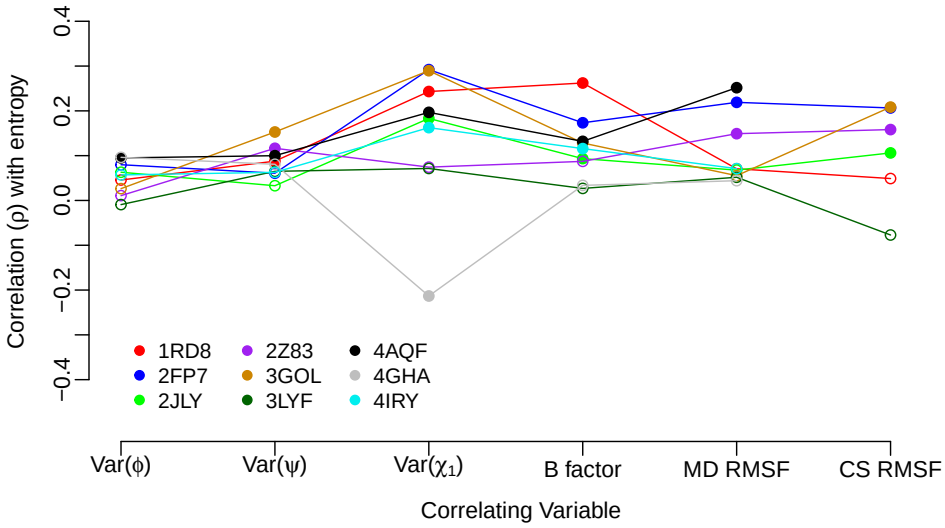

Fig. 1 Spearman correlation of sequence entropy with measures of structural variability. Each symbol represents one correlation coefficient for one protein structure. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. The quantities Var(ψ), Var(ϕ), Var(χ_1), and MD RMSF were obtained as time-averages over 15ns of MD simulations. B factors were obtained from individual crystal structures. CS RMSF values were obtained from alignments of homologous crystal structures when available. Almost all structural measures of variability correlate weakly, but significantly, with sequence entropy.

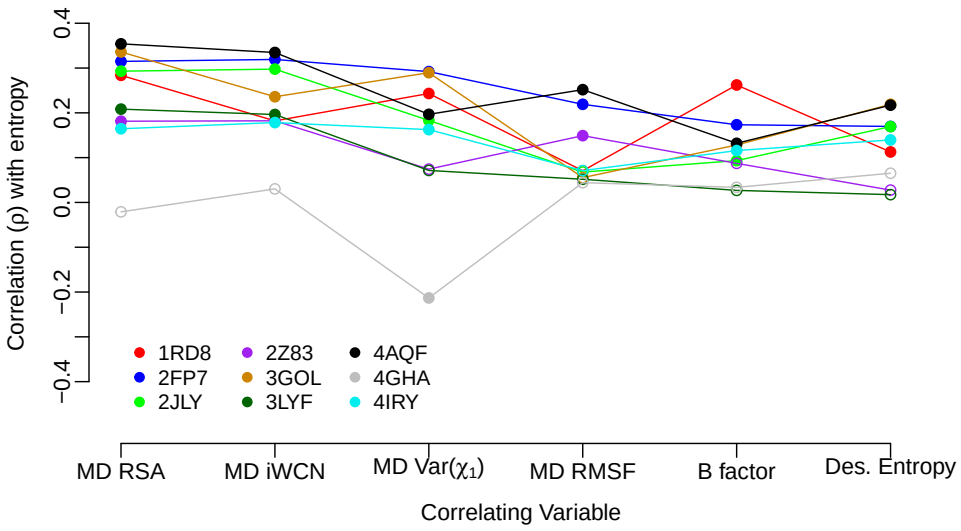

Fig. 2 Spearman correlation of sequence entropy with measures of buriedness and of structural variability, as well as designed entropy. Each symbol represents one correlation coefficient for one protein structure. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. The quantities RSA, iWCN, Var(χ_1), and MD RMSF were obtained as time-averages over 15ns of MD simulations. B factors were obtained from crystal structures. Compared to the measures of structural variability and to designed entropy, the buriedness measures consistently show stronger correlations with sequence entropy. Note that results for iWCN are largely identical to those for iCN, so only iWCN was included here.

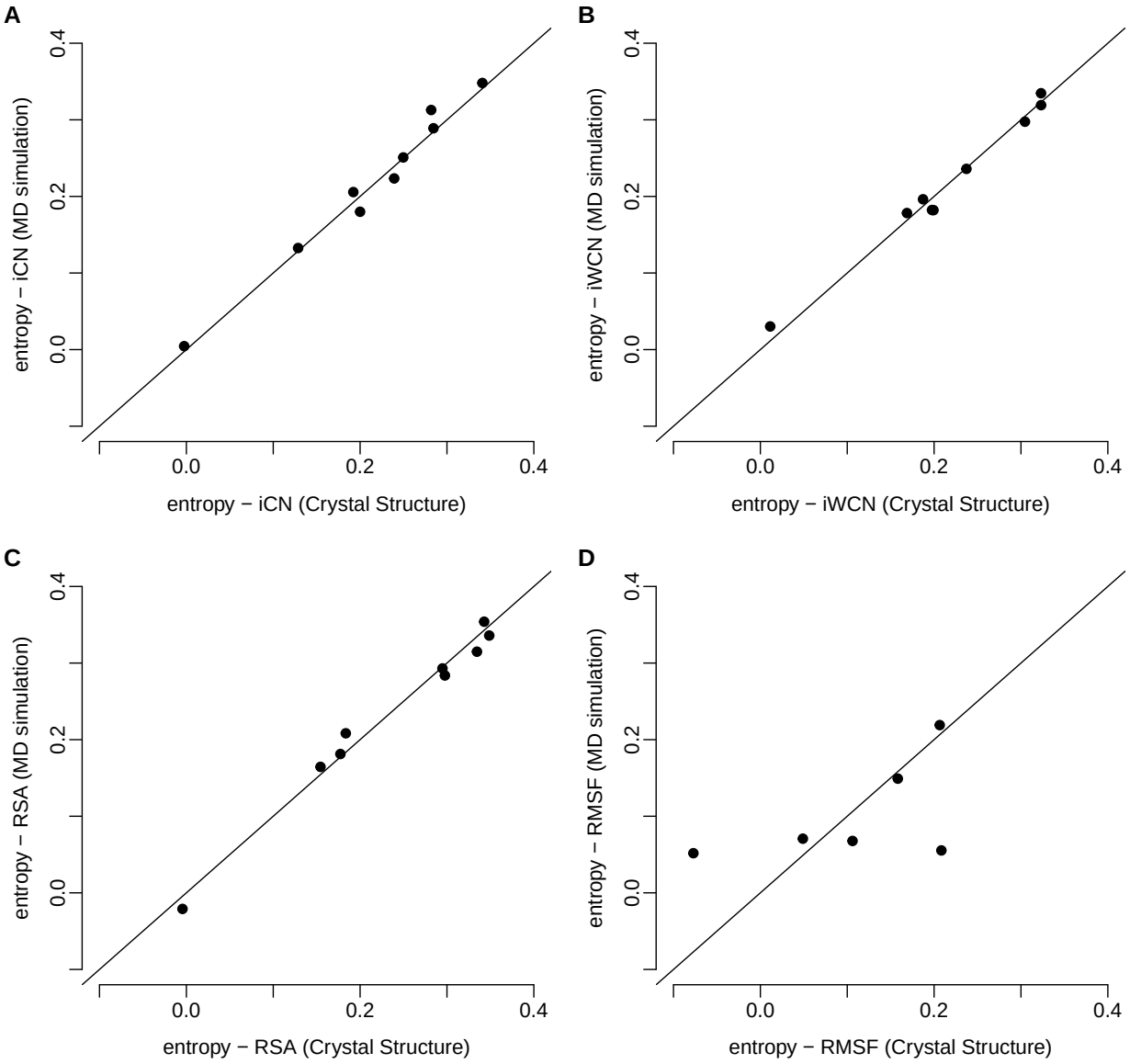

Fig. 3 Spearman correlations of sequence entropy with MD-derived and crystal-structure derived structural measures. The vertical axes in all plots represent the Spearman correlation of sequence entropy with one structural variable obtained from $15ns$ of molecular dynamics (MD) simulations. The horizontal axes represent the Spearman's rank correlation coefficient of sequence entropy with the same structural variable as in the vertical axes but measured from protein crystal structures. Each dot represents one correlation coefficient for one protein structure. The quantities iCN, iWCN, and RSA have nearly identical predictive power for sequence entropy regardless of whether they are derived from MD simulations or from crystal structures. By contrast, MD RMSF yielded very different correlations than did CS RMSF.

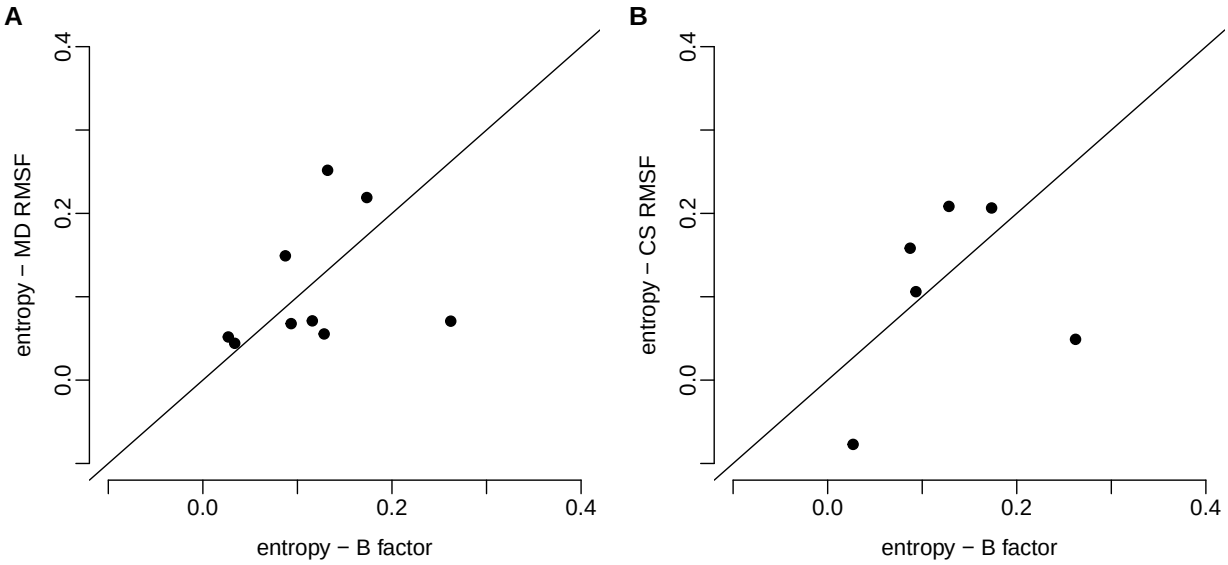

Fig. 4 Spearman correlations of sequence entropy with measures of structural variability. Vertical and horizontal axes represent Spearman correlations of the indicated quantities. Each dot represents one correlation coefficient for one protein structure. MD RMSF, CS RMSF, and B factors all explain different amounts of variance in sequence entropy for different proteins.

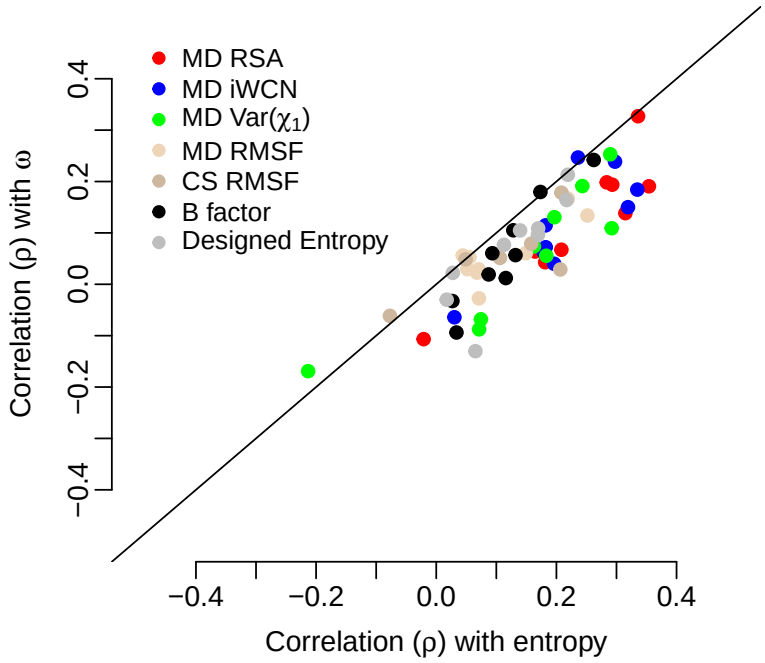

Fig. 5 Spearman correlations of structural quantities with sequence entropy and with the evolutionary rate ratio ω. Nearly all points fall below the $x = y$ line, indicating that structural quantities generally predict as much as or more variation in sequence entropy than in ω.

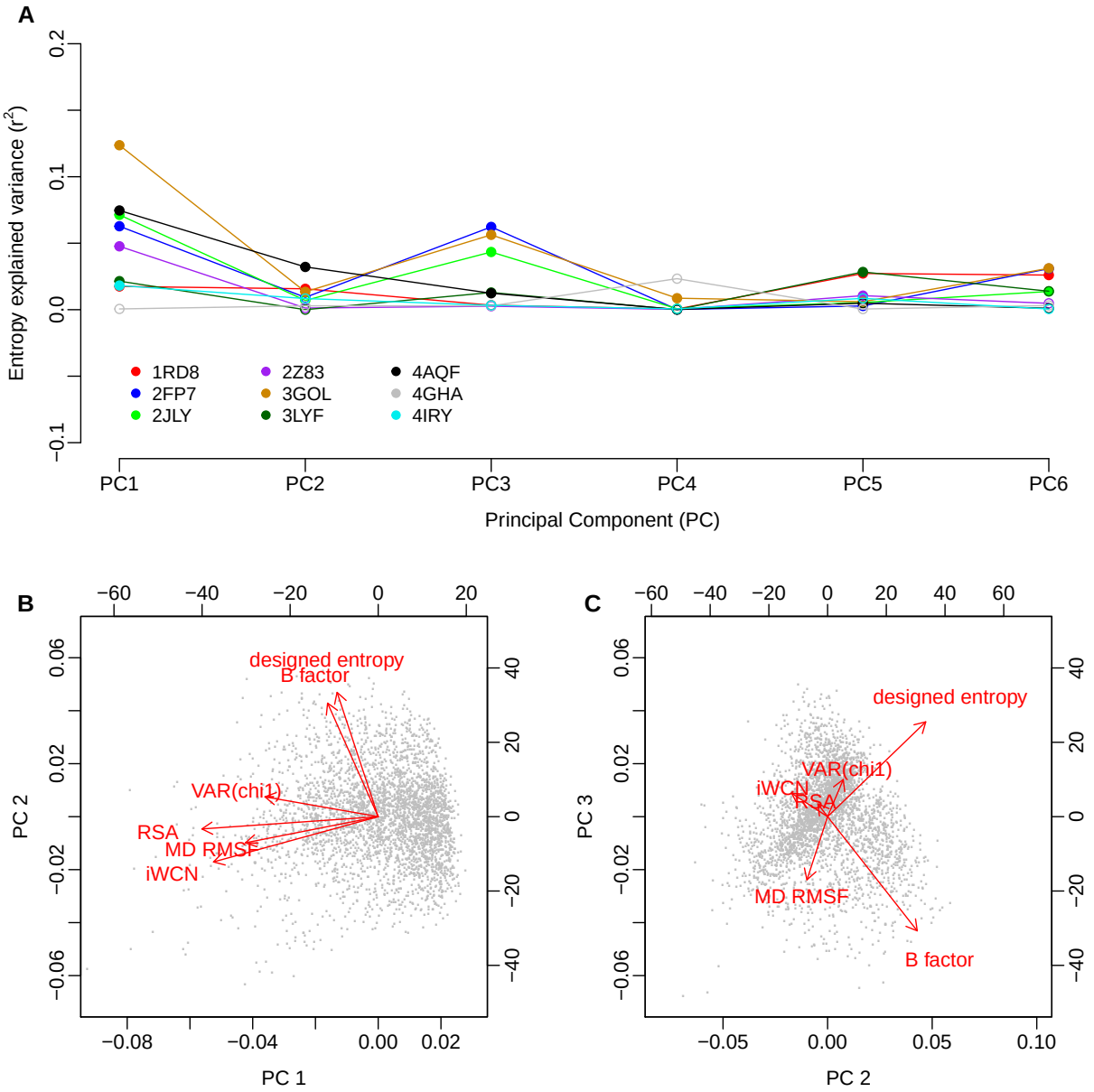

Fig. 6 Principal Component (PC) Regression of sequence entropy against the structural variables. **(A)** Variance in entropy explained by each principal component. For most proteins, PC1 and PC3 show the strongest correlations with sequence entropy. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. **(B)** and **(C)** Composition of the three leading components. Red arrows represent the loadings of each of the structural variables on the principal components; black dots represent the amino acid sites in the PC coordinate system. The variables RSA, iWCN, MD RMSF, and Var(χ_1) load strongly on PC1 and weakly on PC2, while B factor and designed entropy load strongly on PC2 and weakly on PC1. Interestingly, B factor and designed entropy also load strongly on PC3, but in opposite directions.

Supplementary Figures

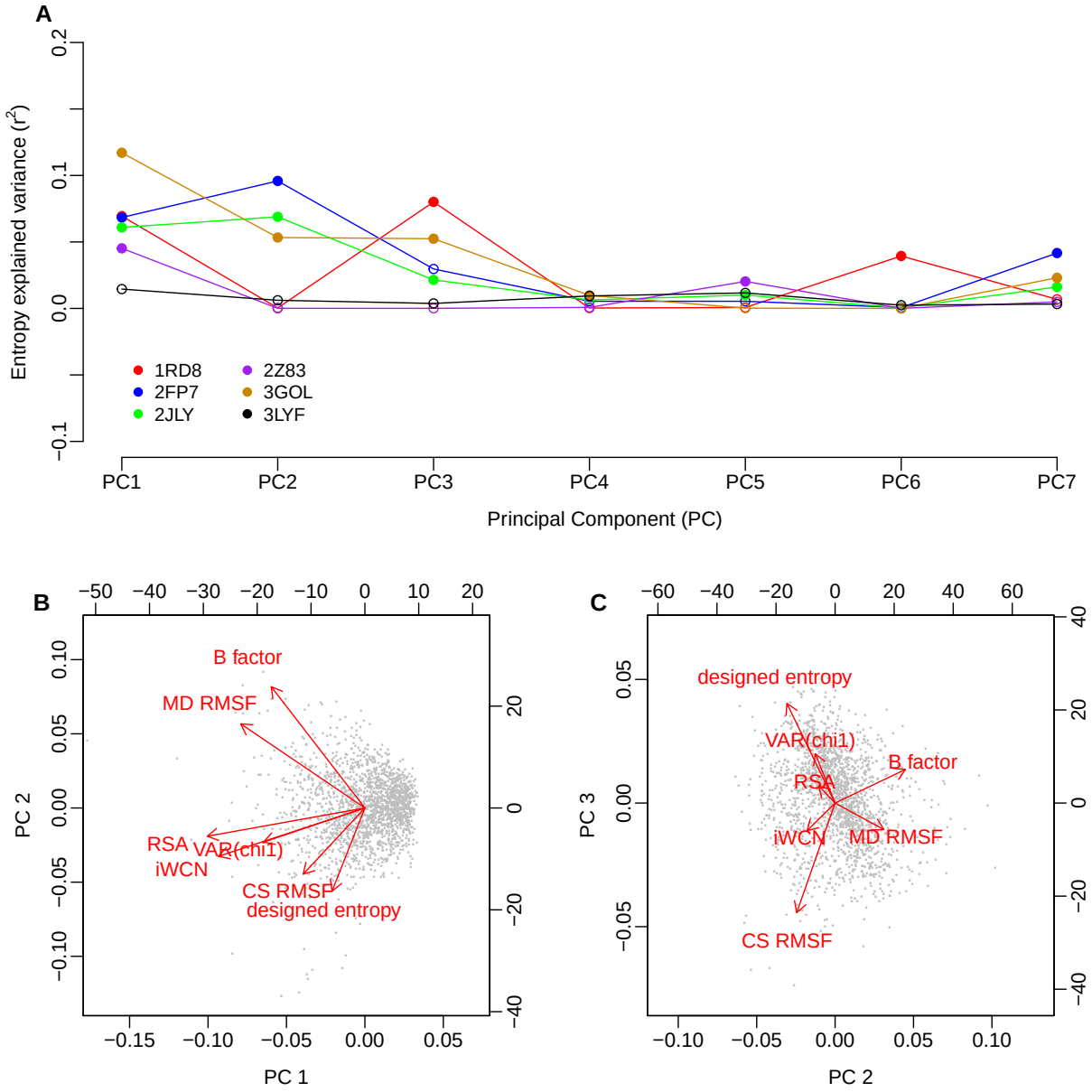

Fig. S1 Principal Component (PC) Regression of sequence entropy against the structural variables, including CS RMSF. **(A)** Variance in entropy explained by each principal component. For most proteins, PC1 and either PC2 or PC3 show the strongest correlations with sequence entropy. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. **(B)** and **(C)** Composition of the three leading components. Red arrows represent the loadings of each of the structural variables on the principal components; black dots represent the amino acid sites in the PC coordinate system.

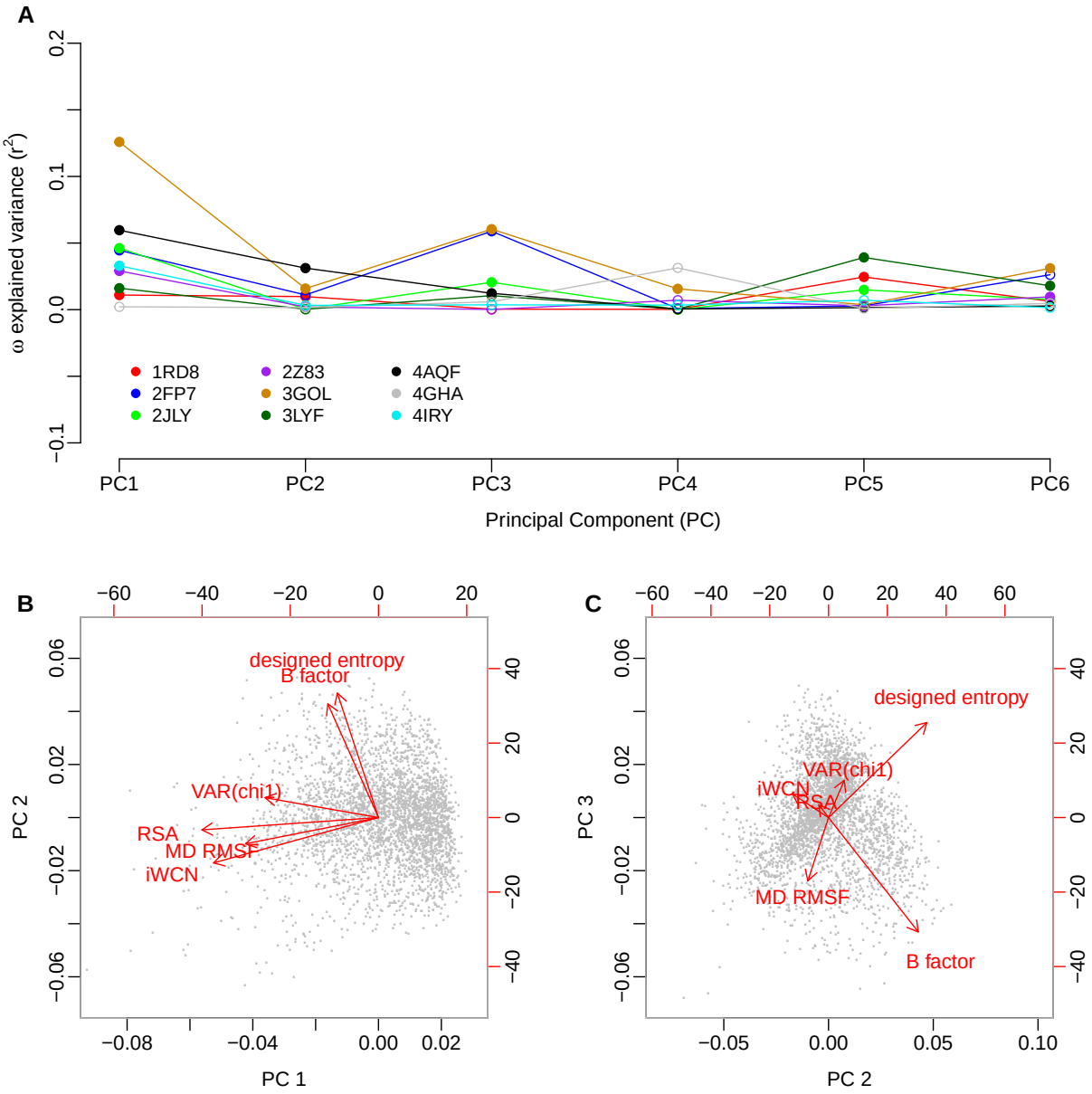

Fig. S2 Principal Component (PC) Regression of ω against the structural variables. **(A)** Variance in ω explained by each principal component. For most proteins, PC1 and PC3 show the strongest correlations with ω. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. **(B)** and **(C)** Composition of the three leading components. Red arrows represent the loadings of each of the structural variables on the principal components; black dots represent the amino acid sites in the PC coordinate system. Note that parts B and C are identical to those shown in Figure 6.

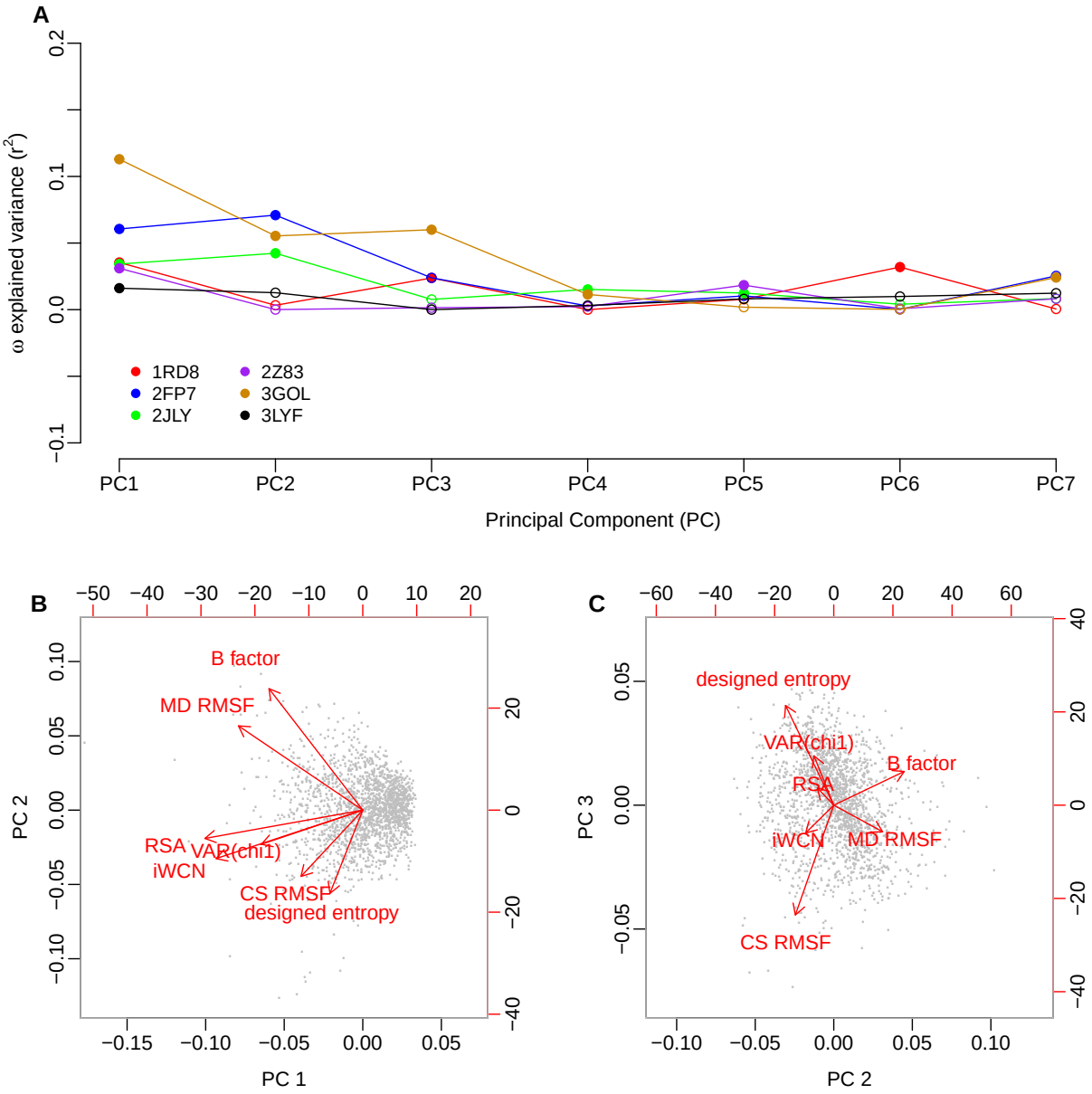

Fig. S3 Principal Component (PC) Regression of ω against the structural variables, including CS RMSF. **(A)** Variance in ω explained by each principal component. For most proteins, PC1 and either PC2 or PC3 show the strongest correlations with ω. Significant correlations ($P < 0.05$) are shown as filled symbols, and insignificant correlations ($P \geq 0.05$) are shown as open symbols. **(B)** and **(C)** Composition of the three leading components. Red arrows represent the loadings of each of the structural variables on the principal components; black dots represent the amino acid sites in the PC coordinate system. Note that parts B and C are identical to those shown in Figure S1.