

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

Amir Shahmoradi,<sup>1,2</sup> Dariya K. Sydykova,<sup>2</sup> Stephanie J. Spielman,<sup>2</sup>  
Eleisha L. Jackson,<sup>2</sup> Eric T. Dawson,<sup>2</sup> Austin G. Meyer,<sup>2</sup> Claus O. Wilke<sup>2</sup>

<sup>1</sup> Department of Physics, The University of Texas at Austin, TX, 78712

<sup>2</sup> 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

## Abstract

Several recent works have shown that site-specific evolutionary variation in proteins can be predicted from protein structure. Most prominently, sites that are buried and/or have many contacts with other sites in a structure evolve more slowly than surface sites with few contacts. Here, we present a comprehensive study of numerous different structural properties that may constrain sequence variation, including measures of buriedness (relative solvent accessibility, contact number), measures of structural fluctuations (B factors, root-mean-square fluctuations, variation in dihedral angles), and variability in designed structures. Structural fluctuation measures were obtained from molecular dynamics simulations performed on 9 non-homologous viral protein structures, and from variation in homologous variants of these proteins where available. Variability in designed structures was obtained from flexible-backbone design via Rosetta. 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). We further found that measures of buriedness tend to be better predictors of evolutionary variation than measures of structural fluctuations. Variability in designed structures was a weaker predictor of evolutionary variability than both measures of buriedness and measures of fluctuations. We conclude that simple measures of buriedness are better predictors of evolutionary variation than more complicated predictors obtained from dynamic simulations, ensembles of homologous structures, or computational protein design.

## Introduction

Identification of the driving factors in protein evolution has been one of the important objectives in molecular biology and protein research (Marsh and Teichmann, 2014). It is already well-established and understood that highly conserved amino acid sites in the protein sequences often fall in hotspot regions responsible for the protein’s biophysical function or happen to be pivotal in maintaining the protein’s native conformation (??). Aside from biophysical constraints, several structural determinants of protein dynamics and flexibility have

been recently proposed to impose site-specific evolutionary pressure on the protein sequence. Examples include residue-level solvent exposure (Ramsey et al., 2011; Scherrer et al., 2012; Meyer and Wilke, 2013; ?), local protein density (Liao et al., 2005; ?; Franzosa and Xia, 2009; Yeh et al., 2014), and measures of residue-level flexibility of the protein backbone (Liu and Bahar, 2012; Nevin Gerek et al., 2013).

Along with RSA, other measures of residue buriedness, such as residue contact number (Liao et al., 2005; Franzosa and Xia, 2009; Yeh et al., 2014) have been proposed and shown to correlate with sequence variability, or even argued to serve as better predictors than RSA. Based on physical arguments and experimental evidence (xx what kind of arguments?), Halle (2002) argued that the local residue packing density is a direct proxy measure of residue and site-specific backbone flexibility, in particular the Debye-Waller factors, commonly known as B factors. Therefore, given the strong observational evidence for the significant positive correlation of residue density and packing with sequence evolution (Yeh et al., 2014), one may also expect to observe a positive trend between local flexibility and sequence variability. Indeed, several authors have argued for the potential role of protein dynamics on sequence variability (Liu and Bahar, 2012; Nevin Gerek et al., 2013).

Although multitude of structural variables have been shown or predicted to influence residue-level sequence variability, there is currently no consensus on which variable and to what extent has the dominant role in regulating sequence variability, independently of other structural determinants. So far, a comprehensive study of all potential structural determinants of protein sequence evolution has been missing in the literature, with the existing work mainly focusing on individual variables. In particular, measures of residue spatial fluctuations and protein dynamics have only received marginal attention and consideration as potential contributing factors to sequence evolution.

Proteins are intrinsically dynamic entities in vivo, far from their perceived rigidities in crystal structures and their dynamic behavior is expected to influence their sequence evolution (Marsh and Teichmann, 2014). However, contrary to RSA and residue contact number, an accurate determination of the protein’s dynamical behavior and residue fluctuations—solely based on the set of 3-dimensional atomic coordinates in crystal structures—remains a challenging task. B factors are generally considered as an attractive proxy to local flexibility, though the atomistic definition of B factor may not be appropriate for the study of side-chain flexibility and fluctuations. Experimental studies of protein dynamics in vivo has also proven extremely difficult if not impossible (Vendruscolo 2007).

Alternatively, Molecular Dynamics (hereafter, MD) simulations provide an ideally suited method of studying protein dynamics and its potential role in driving sequence evolution. Here in this work, we attempt to present a comprehensive study of several potential structural determinants of sequence variability from both protein crystallography and Molecular Dynamics perspectives. In addition to the factors already discussed in previous works, such as RSA and residue contact number, we also consider new dynamical measures of structural variability in the study, such as variance of the backbone and residue dihedral angles and residue spatial fluctuations from MD simulations and discuss their potential influence on sequence variability. The extent and breadth of such analysis however, limits diversity of the input data used in our work to highly evolving proteins that have multiple high-resolution homologous crystal structures in Protein Data Bank (Berman, 2008). The availability of multiple structures is required for the calculation of site-specific spatial fluctuations and

comparison of the results with the same variability measures based on MD simulations. In addition, the selected proteins should also have ample sequence data to ensure good statistic for sequence alignment and the calculation of sequence variability and evolutionary rates.

In the following sections, first we briefly present the methodology employed for data selection, sequence alignments, data analysis and the procedure for obtaining the relevant structural variables from Molecular Dynamics simulations and homologous structures, followed by the results and a discussion of the potential contributors to sequence variability or evolution. We show that measures of residue buriedness such as RSA and contact number outperform site-specific measures of residue fluctuations and discuss the potential underlying biases and reasons contributing to this observation.

## 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. We only used sequences from after 1998 but 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  $\omega$  rate categories as free parameters. We employed an Empirical Bayes approach (Yang, 2000) to infer  $\omega$  values for each position in the alignment. These  $\omega$  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  $\geq 10\%$ .

## Measures of buriedness and of structural flexibility

As 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  $C\alpha$  atoms surrounding the  $C\alpha$  atom of the focal residue within a spherical neighborhood of a predefined radius  $r_0$ . Following Yeh et al. (2014), we used  $r_0 = 13\text{\AA}$ . We calculated WCN as the total number of surrounding  $C\alpha$  atoms for each focal residue, weighted by the inverse square separation between the  $C\alpha$  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.

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

$$\text{RMSF}_j = \left[ \sum_i (\mathbf{r}_i^{(j)} - \mathbf{r}_0^{(j)})^2 \right]^{1/2} \quad (2)$$

where  $\text{RMSF}_j$  is the root-mean-square fluctuation at site  $j$ ,  $\mathbf{r}_i^{(j)}$  is the position of the  $C\alpha$  atom of residue  $j$  at MD frame  $i$ , and  $\mathbf{r}_0^{(j)}$  is the position of the  $C\alpha$  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

$$\text{RMSF}_j = \left[ \sum_i w_i (\mathbf{r}_i^{(j)} - \langle \mathbf{r}^{(j)} \rangle)^2 \right]^{1/2}, \quad (3)$$

where  $\mathbf{r}_i^{(j)}$  now stands for the position of the  $C\alpha$  atom of residue  $j$  in structure  $i$ ,  $\langle \mathbf{r}^{(j)} \rangle$  is the mean position of that  $C\alpha$  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  $\text{Var}(\phi)$ ,  $\text{Var}(\psi)$ , and  $\text{Var}(\chi_1)$ . 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  $\alpha_i$  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

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

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  $\text{Var}(\alpha) = 0$  corresponding to the minimum variability of the dihedral angle and  $\text{Var}(\alpha) = 1$  to the maximum, respectively (Berens, 2009). Since the  $\chi_1$  angle is undefined for Ala and Gly we excluded all sides with these residues in analyses involving  $\chi_1$ .

B factors were extracted from the crystal structures. We only considered the B factors of the C $\alpha$  atom of each residue.

## 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 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.

## 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](https://github.com/clauswilke/structural_prediction_of_ER).

## Results

### Data set and structural variables considered

Our goal in this work is to determine which structural properties best predict amino-acid variability at individual sites in proteins. To this end, we selected 9 viral proteins for which we had both high-quality crystal structures and abundant sequences to assess evolutionary variability (Table 1). We quantified evolutionary variability in two ways, by calculating entropies for each alignment column (sequence entropy) and by calculating the evolutionary-rate ratio  $\omega = dN/dS$  (see Methods for details). Throughout this paper, we primarily report results obtained for sequence entropy. Results for  $\omega$  are 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. Measures of buriedness quantify the extent to which a residue is protected from solvent. A commonly used measure of buriedness is solvent-accessible surface area (ASA) or its normalized variant relative solvent accessibility (RSA). Here we considered RSA, since it can be compared among residues of different sizes. We also considered contact number (CN) and weighted contact number (WCN). Both of these quantities assess the number of other residues a focal residue contacts. CN simply counts the number of contacts within a sphere of a given radius around the  $\alpha$ -carbon of the focal residue. WCN weights contacts by the distance between the two residues. Note that residue buriedness decreases as RSA increases, but it increases with increasing CN or WCN. To avoid this difference in directionality, in most analyses we replaced CN and WCN with their inverse,  $iCN = 1/CN$  and  $iWCN = 1/WCN$ . Because nearly all analyses we carried out were non-parametric and only depended on the rank order of the variables, this substitution only changed the sign of correlations but not the magnitude.

Measures of flexibility assess the extent to which a residue fluctuates in space as a protein undergoes thermodynamic fluctuations in solution. There are different ways to quantify these fluctuations. First, we can measure the root mean-square deviation of the  $C\alpha$  atom over time. This quantity is commonly called RMSF. Second, we can consider B factors, which measure the spatial localization of individual atoms in a protein crystal. Third, we can measure the variability in side-chain or backbone dihedral angles, such as  $\text{Var}(\chi_1)$ ,  $\text{Var}(\phi)$ , or  $\text{Var}(\psi)$ . Here, we considered all these measures of structural variability.

To calculate measures of flexibility, we generated molecular-dynamics (MD) trajectories for all crystal structures in Table 1. In all cases, we equilibrated the structure and then simulated 15ns of chemical time (see Methods). We recorded snapshots of the simulated structure every 10ps. From these snapshots, for each residue we calculated RMSF as well as the variability in dihedral and side-chain angles. We also calculated time-averaged values

of the three measures of buriedness RSA, CN, and WCN. We refer to these time-averaged values as MD RSA, MD CN, and MD WCN, respectively.

## Structural buriedness vs. structural flexibility as predictors of evolutionary variation

We first compared a subset of variables among the measures of buriedness and measures of fluctuations in their explanatory power for sequence variation. Figure 1 shows the Spearman correlation between sequence entropy and each of the quantities RSA, iWCN,  $\text{Var}(\chi_1)$ , RMSF, and B factor, for each protein. Significant correlations ( $P < 0.05$ ) are shown with filled symbols, and non-significant correlations are shown with empty symbols ( $P \geq 0.05$ ). In this figure, several patterns emerge. First, all correlations are generally positive and significant. The main exception is the RNA binding domain of Marburg virus, PDB ID 4GHA, which mostly shows no correlation but shows a negative correlation between sequence entropy and the variability in the  $\chi_1$  angle. Second, correlations are generally weak. No correlation coefficient exceeds 0.4. Third, on average the correlation strength decreases as we move from left to right in the figure. The two measures of buriedness, RSA and iWCN, show the strongest correlations. Their correlations are on average  $\rho = 0.23$  and  $\rho = 0.22$ , respectively. The three fluctuation measures  $\text{Var}(\chi_1)$ , RMSF, and B factor exhibit weaker correlations, on average  $\rho = 0.14$ ,  $\rho = 0.11$ , and  $\rho = 0.12$ , respectively.

We next analyzed structural fluctuations more carefully, by comparing the correlations of entropy with six different measures of local structural flexibility. We considered variations in the backbone and side-chain dihedral angles ( $\phi$ ,  $\psi$ , and  $\chi_1$ ), B factors, RMSF obtained from MD simulations, and RMSF obtained from crystal structures. For the latter quantity, we aligned homologous structures with distinct sequences, obtained from the PDB (see Methods and Table 2 for details). The correlation strengths of these quantities with entropy are shown in Figure 2. We found that the variation in backbone dihedral angles,  $\text{Var}(\phi)$  and  $\text{Var}(\psi)$ , explained the least variation in sequence entropy, while the variation in the side-chain dihedral angle  $\text{Var}(\chi_1)$  explained, on average, more variation in sequence entropy than 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).

## MD time-averages vs. crystal-structure snapshots

For most analyses discussed so far, except analyses involving B factors or crystal RMSF, we averaged structural quantities over MD trajectories comprising 15ns of chemical time. This is not conventional practice for quantities such as RSA or contact numbers. Instead, most authors simply obtain these quantities from individual crystal structures. We therefore asked whether MD averages differed in any meaningful way from estimates obtained from crystal structures, and whether estimates from MD and from crystal structures differed in their predictive power for sequence variability.

We found that RSA, CN, and WCN from crystal structures were highly correlated with their averages over MD trajectories, for all protein structures we examined (Spearman correlation coefficients of  $> 0.9$  in all cases, Table 3). Further, when we correlated these quantities

with sequence entropy, we found that the correlation coefficients we obtained for each protein were virtually identical (Figure 3A-C). Thus, in terms of predicting evolutionary variation, RSA and contact numbers obtained from the static structures performed as well as their dynamic equivalents averaged over short time scales.

By contrast, we found more variation when we considered backbone flexibility as measured by root mean fluctuations (RMSF). RMSF cannot be obtained from a single crystal structure, but we can calculate it from an alignment of multiple crystal structures were available. When we compared RMSF from MD to RMSF from crystal structures, we found that they 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 the correlation between site entropy and RMSF depended substantially on whether RMSF was obtained from MD or from crystal structures (Figures 2 and 3D).

To further investigate the relationship between backbone fluctuations and sequence variability, we also compared the RMSF correlations to the correlations between sequence entropy and B factors. Again, we found that these correlations were generally different from the ones found for either the MD RMSF or the crystal structure RMSF (Figure 4). Thus, B factors, MD RMSF, and crystal RMSF, though all measures of backbone fluctuations, contained distinct information about sequence variability in our data set.

## Sequence entropy vs. evolutionary-rate ratio $\omega$

In the previous subsections, we have used sequence entropy as a measure of amino-acid variability at individual sites. While sequence entropy is a simple and straightforward measure of site variability, it has two potential drawbacks: First, entropy doesn't correct for the phylogenetic relationship of sequences in the alignment, and hence it can be biased if some parts of the phylogeny are more densely sampled than others. Second, entropy does not take into account the actual substitution process. 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 sequence variability 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  $\omega$  instead of entropy. We found that all results generally carried over, but the correlations tended to be somewhat weaker. Figure 5 plots, for each protein, the correlation between  $\omega$  and the various structural quantities versus the correlation between entropy and the same structural quantities. We see that all data points generally fall below the  $x = y$  line, and are shifted downwards by approximately 0.1. Thus, correlations of structural quantities with  $\omega$  are, on average, approximately 0.1 smaller than correlations of the same quantities with entropy.

Besides the measures of buriedness and measures of structural flexibility discussed in the previous subsections, we present in Figure 5 one additional structural quantity, *designed entropy*. Designed entropy is obtained by using computational protein design to generate artificial alignments of designed sequences, and then measuring sequence entropy in these artificial alignments. We had previously shown that designed entropy captures some of the variation observed in natural alignments (Jackson et al., 2013). Here we found that designed



entropy performed about as well as the measures of structural flexibility but worse than the measures of buriedness.

## 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. To determine the extent to which the various quantities contain independent information about sequence variability, and to assess whether combining multiple structural quantities yields improved predictive power, we carried out a joint multivariate analysis including most of the structural quantities considered in this work. The approach we used is a principal component (PC) regression, which has previously been used successfully to disentangle genomic predictors of whole-protein evolutionary rates (Drummond et al., 2006; Bloom et al., 2006). In this analysis, we first carry out a principal component analysis of the predictor variables (i.e., the structural quantities such as RSA and RMSF), and we then regress the response (sequence entropy or  $\omega$ ) against the individual components.

Because we wanted to analyze all proteins in our data set individually but in such a way that results were comparable from one protein to the next, we pooled all structural quantities for a single PC analysis and then regressed entropy and  $\omega$  against each PC separately for each protein. 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). We found the second-highest  $r^2$  values, on average, for PC3, 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  $\text{Var}(\chi_1)$  all loaded strongly on PC1, while PC2 and PC3 were 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 think that PC1 can be considered 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.

When we included RMSF derived from crystal structures in our analysis, we found comparable results. 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, results were generally comparable when we considered  $\omega$  instead of sequence entropy (Figures S2 and S3).

## Discussion

We have carried out a comprehensive analysis of how well structural quantities predict evolutionary variation in nine viral proteins. We have found that measures of buriedness generally performed better than 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 contact number showed comparable correlation strengths with evolutionary 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).

We have found here that correlations between sequence entropy and structural quantities were consistently higher than correlations between the evolutionary-rate ratio  $\omega$  and structural quantities. This difference likely reflects the distinct physical processes that entropy and  $\omega$  measure. Entropy is a measure of the amino-acid diversity allowed at a given site. In effect, it reflects how many different amino acids are allowed. By contrast,  $\omega$  measures how rapidly amino-acid changes occur at a given site. While entropy and  $\omega$  are generally correlated, a site can have high entropy and comparatively low  $\omega$  and vice versa. In particular, if mutations rapidly toggle back and forth between two different amino acids at a site, then that site will have high  $\omega$  and low entropy. By contrast, if a site diversified into a number of different alleles deep in the phylogeny but did not experience much further evolution at later times, then that site will have comparatively high entropy but low  $\omega$ . Since structural quantities such as measures of buriedness reflect the biophysical constraints imposed on sites, it makes sense 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 such as immune escape would likely be better predictors of substitution rates than amino-acid diversity.

The correlation strengths we observed were consistently lower than those observed in prior work (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. And the lower the evolutionary divergence, the harder it becomes to resolve differences between more and less conserved sites. 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, and both at buried and at exposed sites (Meyer and Wilke, 2013; Meyer et al., 2013; Suzuki, 2006; Bush et al., 1999).

We found that simple measures of buriedness, such as RSA and contact numbers, were better predictors of evolutionary variation than sequence variability predicted from computational protein design. Thus, 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), it remains limited in its ability to predict evolutionary variation. We had previously found that flexible backbone design with Rosetta resulted in designs whose surface and core were too similar (Jackson et al., 2013), and we had suggested that this discrepancy was caused either by 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 be the cause of at least some of the discrepancies between predicted and observed site variability. In particular, in our PC regression analysis we found that 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. Thus, designed entropy tended to perform better for sites with less structural variability.

Even though RSA and contact number 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 Mittelman, 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).

## References

- Berens P. 2009. CircStat: a MATLAB toolbox for circular statistics. *J. Stat. Software* 31:1–21.
- Berman H M. 2008. The Protein Data Bank: a historical perspective. *Acta crystallographica. Section A, Foundations of crystallography* 64:88–95.

- Bloom J D, Drummond D A, Arnold F H, Wilke C O. 2006. Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.* 23:1751–1761.
- Bordner A J, Mittelman H D. 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:e1000633.
- Bush R M, Bender C A, Subbarao K, Cox N J, Fitch W M. 1999. Predicting the evolution of human influenza A. *Science* 286:1921–1925.
- Drummond D A, Raval A, Wilke C O. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* 23:327–337.
- Fleishman S J, Whitehead T A, Ekiert D C, Dreyfus C, Corn J E, Strauch E M, Wilson I A, Baker D. 2011. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* 332:816–821.
- Franzosa E A, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.* 26:2387–2395.
- Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11:725–736.
- Grant B J, Rodrigues A P C, ElSawy K M, McCammon A J, Caves L S D. 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 T T, del Valle Marcos M L, 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.
- Jackson E L, Ollikainen N, III A W C, Kortemme T, Wilke C O. 2013. Amino-acid site variability among natural and designed proteins. *PeerJ* 1:e211.
- Jones D T, Buchan D W A, 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.
- Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.
- Katoh K, Kuma K I, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucl. Acids Res.* 33:511–518.

- Katoh K, Misawa K, Kuma K I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl. Acids Res.* 30:3059–3066.
- Kosakovsky P, Pond S L, Frost S D W, Muse S V. 2005. HyPhy: hypothesis testing using phylogenetics. *Bioinformatics* 21:676–679.
- 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 S M, Lange O F, Thompson J, Jacak R, Kaufman K, Renfrew D P, Smith C A, Sheffler W, Davis I W, Cooper S, Treuille A, Mandell D J, Richter F, Ban Y E A, Fleishman S J, Corn J E, Kim D E, Lyskov S, Berrondo M, Mentzer S, Popović Z, Havranek J J, Karanicolas J, Das R, Meiler J, Kortemme T, Gray J J, 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 R L, 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 D S, Colwell L J, Sheridan R, Hopf T A, Pagnani A, Zecchina R, Sander C. 2011. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6:e28766.
- Marsh J A, Teichmann S A. 2014. Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays* 36:209–218.
- Meyer A G, Dawson E T, Wilke C O. 2013. Cross-species comparison of site-specific evolutionary-rate variation in influenza haemagglutinin. *Phil. Trans. R. Soc. B* 368:20120334.
- Meyer A G, Wilke C O. 2013. Integrating sequence variation and protein structure to identify sites under selection. *Mol. Biol. Evol.* 30:36–44.
- 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:e1003313.
- Ramsey D C, Scherrer M P, Zhou T, Wilke C O. 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 A M, Jiang L, DeChancie J, Betker J, Gallaher J L, Althoff E A, Zanghellini A, Dym O, Albeck S, Houk K N, Tawfik D S, Baker D. 2008. Kemp elimination catalysts by computational enzyme design. *Nature* 453:190–195.
- Ryckaert J P, Ciccotti G, Berendsen H J C. 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 A W, Poole D, Le Grand S, Walker R C. 2013. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* 9:3878–3888.
- Scherrer M P, Meyer A G, Wilke C O. 2012. Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evol. Biol.* Submitted.
- Shih C H, Chang C M, Lin Y S, Lo W, Hwang J K. 2012. Evolutionary information hidden in a single protein structure. *Proteins: Structure, Function, and Bioinformatics* 80:1647–1657.
- Smith C A, 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 S J, Wilke C O. 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 E A, 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 M Z, Meyer A G, Sydykova D K, Spielman S J, Wilke C O. 2013. Maximum allowed solvent accessibilities of residues in proteins. *PLOS ONE* 8:e80635.
- 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 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.

# 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.** Even though most viral proteins have many PDB structures available, the sequence divergence among these structures is low. To calculate RMSF from crystal structures, we here considered all proteins for which we could find at least five homologous structures at 5% pairwise sequence divergence (highlighted in bold).

Viral Protein	BLAST hits <sup>a</sup>	Unique sequences			
		all	$\geq 2\%^b$	$\geq 5\%^b$	$\geq 10\%^b$
Hemagglutinin Precursor	63	17	10	<b>9</b>	7
Dengue Protease Helicase	31	13	7	<b>7</b>	7
West Nile Protease	21	16	10	<b>7</b>	6
Japanese Encephalitis Helicase	31	12	7	<b>7</b>	7
Hepatitis C Protease	302	33	10	<b>5</b>	4
Rift Valley Fever Nucleoprotein	95	9	5	<b>5</b>	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

<sup>a</sup> BLAST hits against all sequences in the PDB, excluding hits with  $< 35\%$  sequence identity and  $< 90\%$  alignment length

<sup>b</sup> 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  $\rho$  between the values obtained from MD time averages and the values obtained from crystal structures. We then calculated the minimum, maximum, mean, and standard deviation of these correlations.

Quantity	min $\rho$	max $\rho$	$\langle \rho \rangle$	SD( $\rho$ )
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

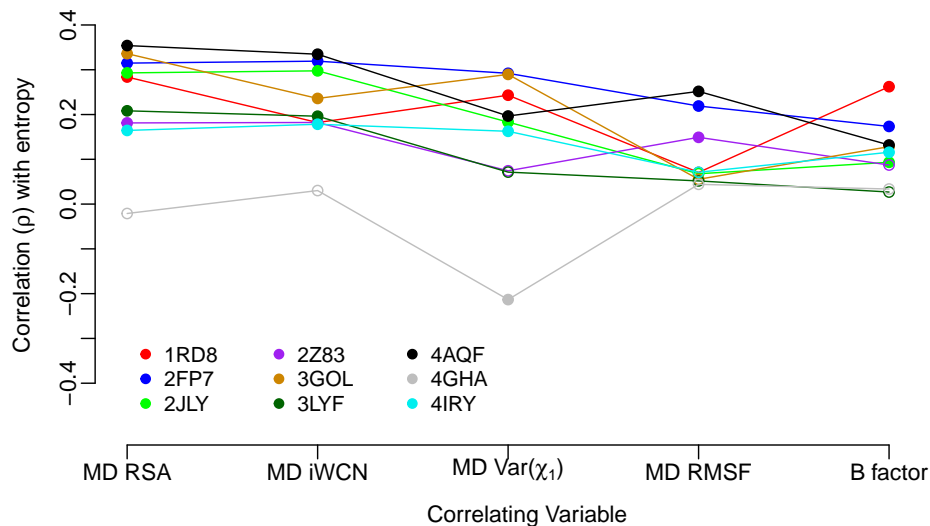


Figure 1: **Spearman correlation of sequence entropy with measures of buriedness and of structural variability.** Each symbol represents one correlation coefficient for one protein structure. Significant correlations ( $P < 0.05$ ) are shown as filled symbols, and insignificant correlations ( $P \leq 0.05$ ) are shown as open symbols. The quantities RSA, iWCN,  $\text{Var}(\chi_1)$ , and 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, the buriedness measures consistently show stronger correlations with sequence entropy.

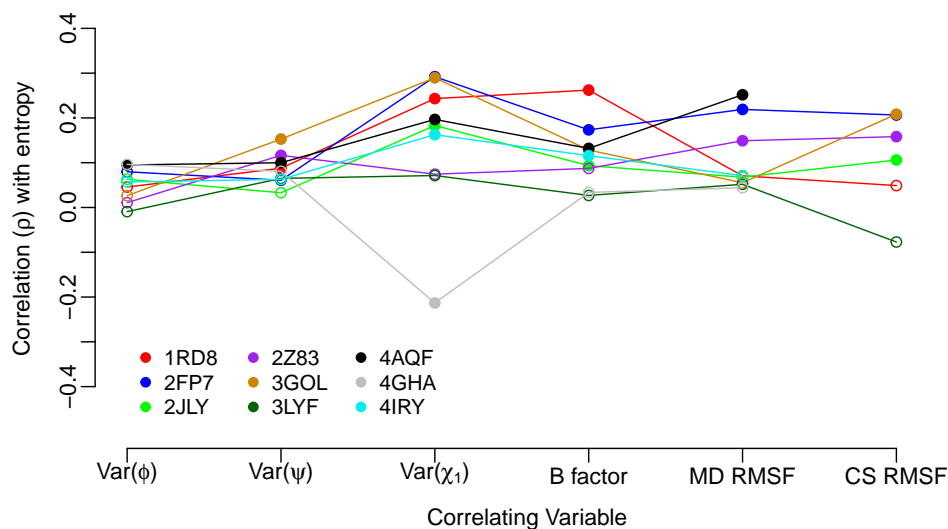
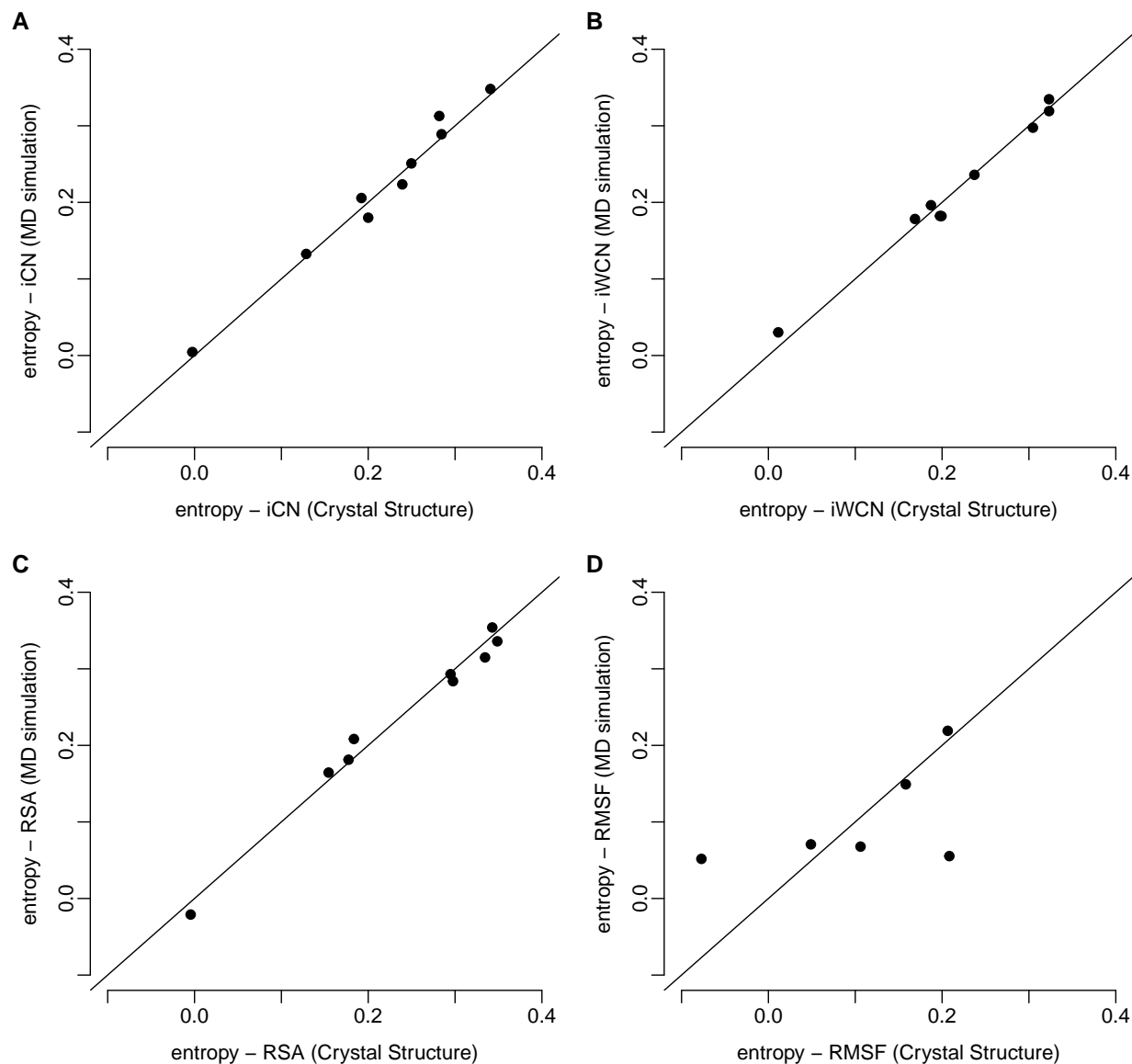


Figure 2: **Spearman correlation of sequence entropy with measures of structural variability.** Each symbol represents one correlation coefficient for one protein structure. Significant correlations ( $P < 0.05$ ) are shown as filled symbols, and insignificant correlations ( $P \geq 0.05$ ) are shown as open symbols. The quantities  $\text{Var}(\psi)$ ,  $\text{Var}(\phi)$ ,  $\text{Var}(\chi_1)$ , and MD RMSF were obtained as time-averages over 15ns of MD simulations. B factors were obtained from crystal structures. CS RMSF values were obtained from alignments of homologous crystal structures. Almost all structural measures of variability correlate weakly but significantly with sequence entropy.



**Figure 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, RMSF from MD simulations leads to very different correlations than RMSF from crystal structures does.

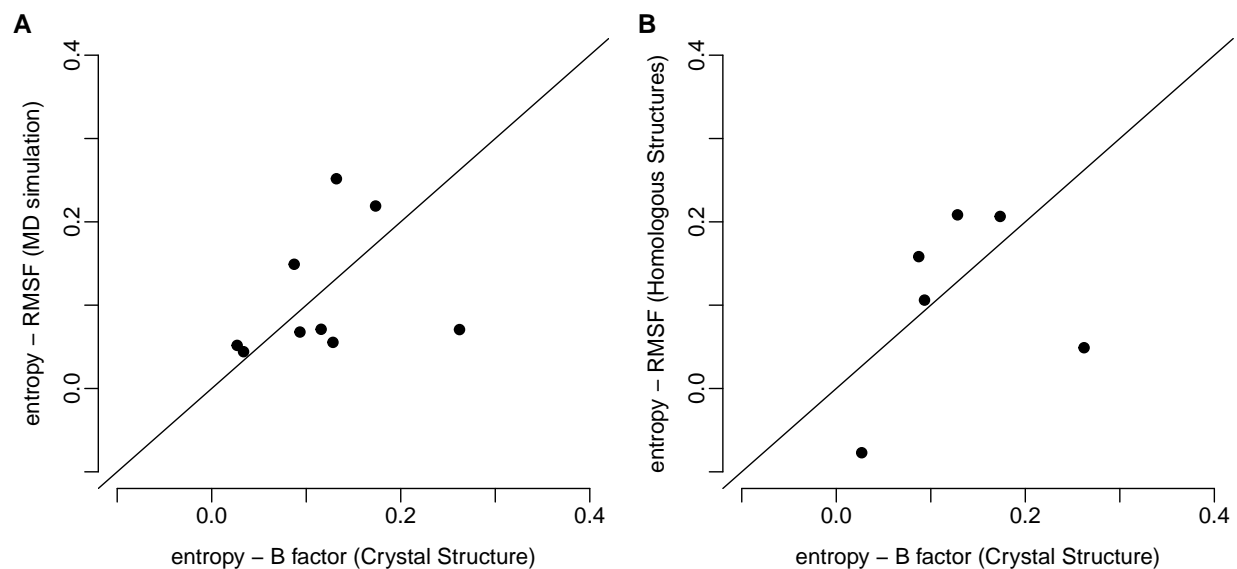


Figure 4: **Spearman correlations of sequence entropy with measures of structural variability.** Vertical and horizontal axes represent Spearman correlations of the indicated quantities. Each dot represents one correlation coefficient for one protein structure. MD RMSF, crystal-structure RMSF, and B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

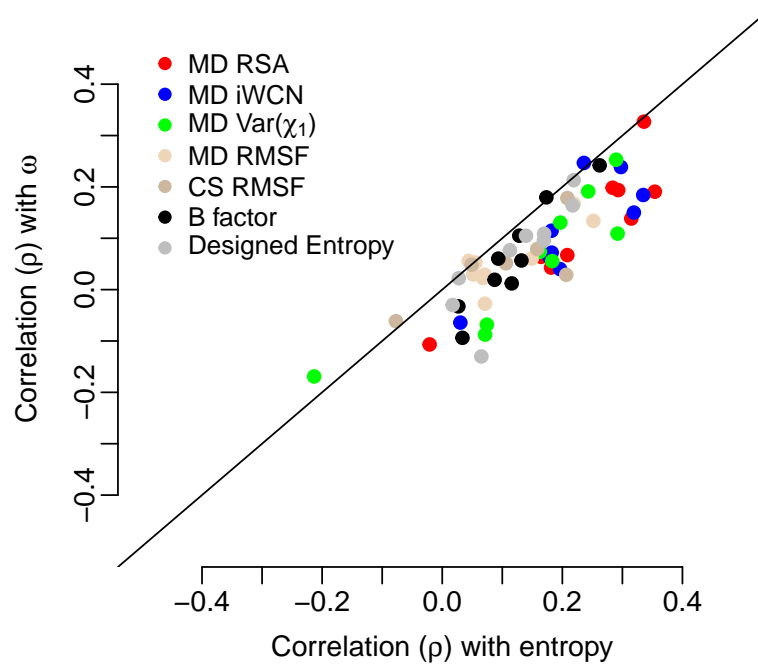
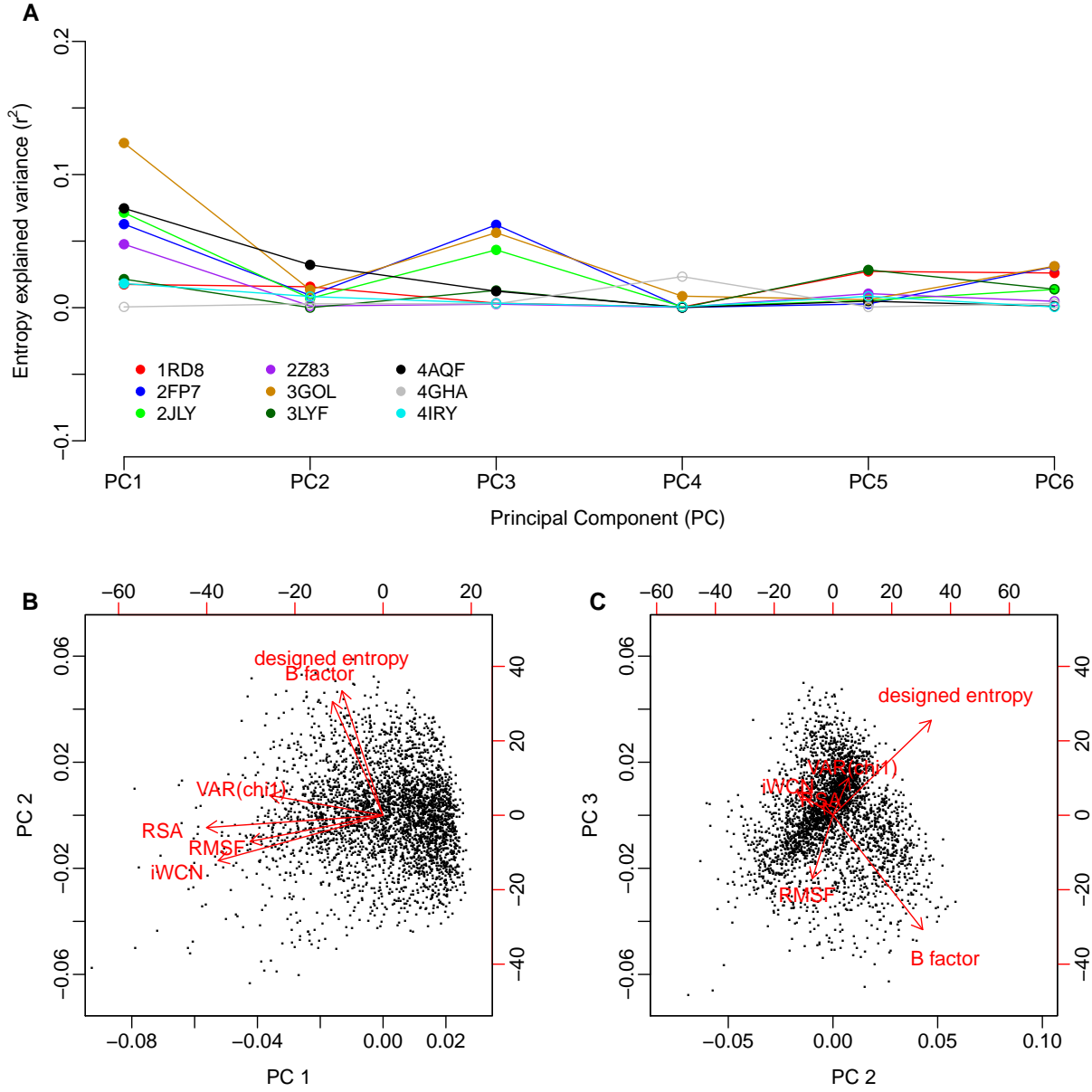


Figure 5: **Spearman correlations of structural quantities with sequence entropy and with the evolutionary rate ratio  $\omega$ .** All structural quantities generally predict as much as or more variation in sequence entropy than in  $\omega$ .



**Figure 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( $\chi_1$ ) 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

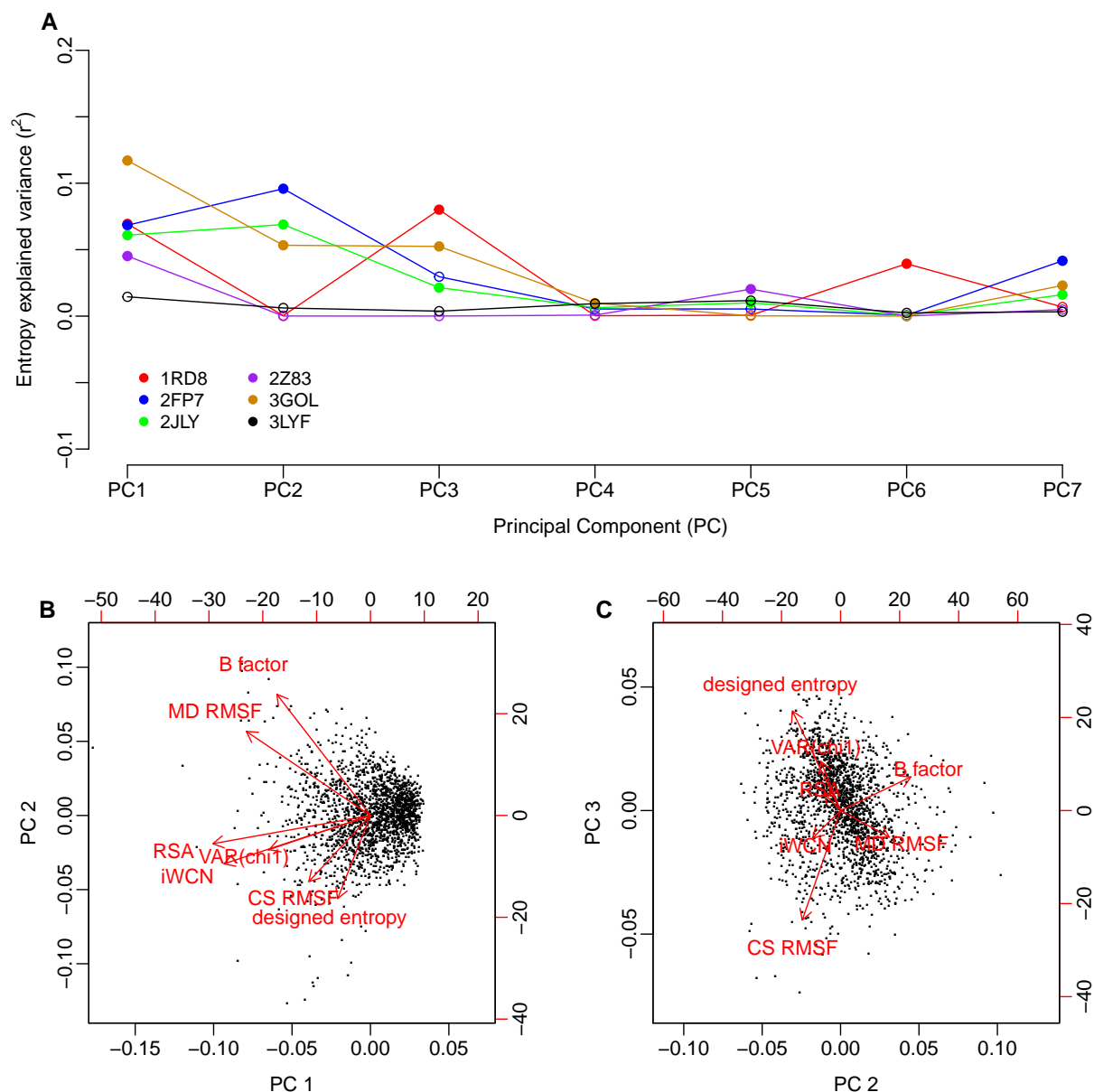


Figure S1: **Principal Component (PC) Regression of sequence entropy against the structural variables, including CS RMSF.** (A) Variance in entropy explained by each principal component. For most proteins, PC1 and either PC2 or PC3 show the strongest correlations with sequence entropy. Significant correlations ( $P < 0.05$ ) 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.

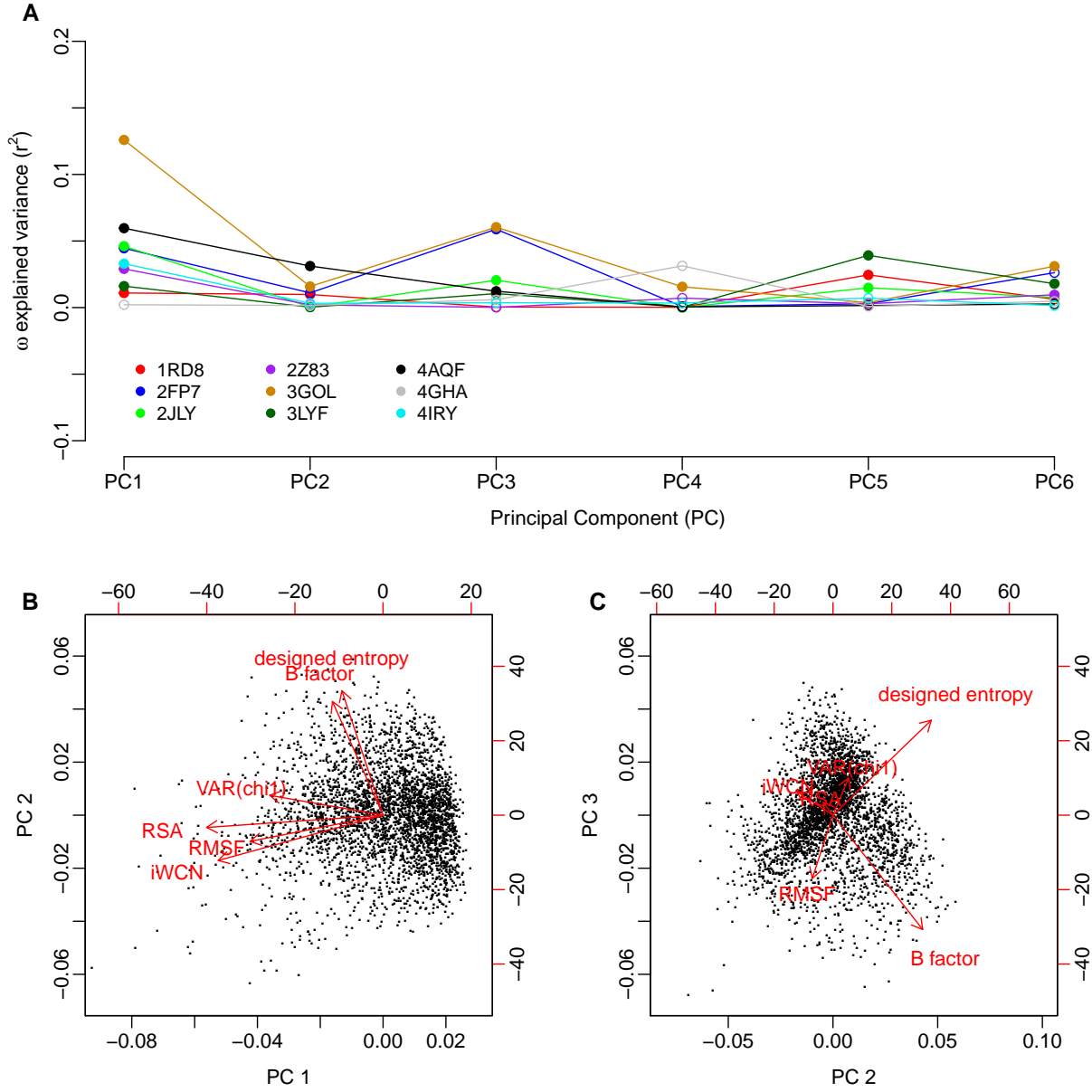
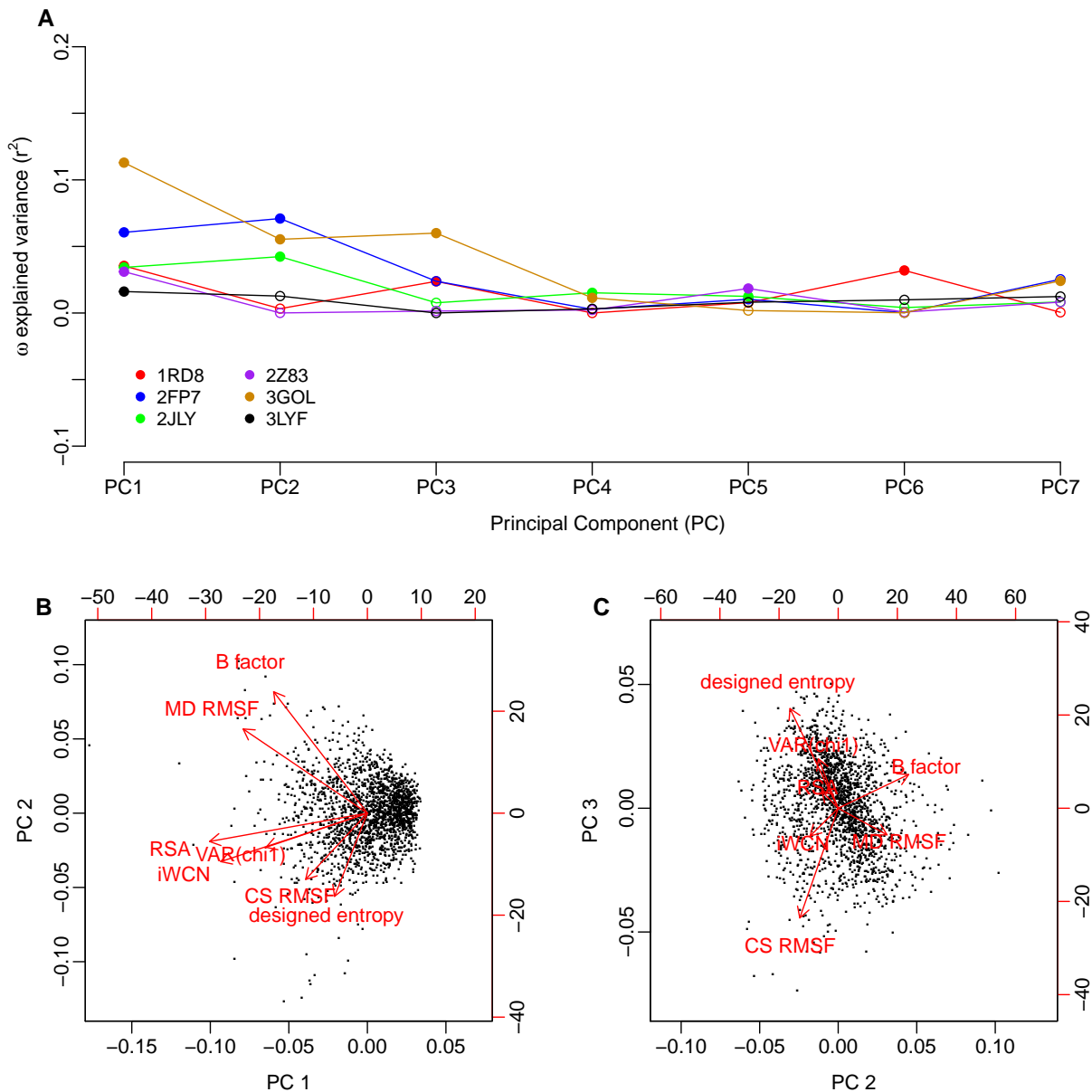


Figure S2: **Principal Component (PC) Regression of  $\omega$  against the structural variables.** (A) Variance in  $\omega$  explained by each principal component. For most proteins, PC1 and PC3 show the strongest correlations with  $\omega$ . Significant correlations ( $P < 0.05$ ) are shown as filled symbols, and insignificant correlations ( $P \leq 0.05$ ) 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.





**Figure S3: Principal Component (PC) Regression of  $\omega$  against the structural variables, including CS RMSF.** (A) Variance in  $\omega$  explained by each principal component. For most proteins, PC1 and either PC2 or PC3 show the strongest correlations with  $\omega$ . 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.