January 29, 2015

Dissecting the relationship between protein structure and sequence variation

Amir Shahmoradi, Eleisha Jackson, Claus Wilke

Department of Physics
The University of Texas at Austin, TX 78712, USA; amir@physics.utexas.edu

Over the past decade several independent works have shown that some structural properties of proteins are capable of predicting protein evolution. The strength and significance of these structure-sequence relations, however, appear to vary widely among different proteins, with absolute correlation strengths ranging from 0.1 to 0.8. Here we present the results from a comprehensive search for the potential biophysical and structural determinants of protein evolution by studying more than 200 structural and evolutionary properties in a dataset of 209 monomeric enzymes. We discuss the main protein characteristics responsible for the general patterns of protein evolution, and identify sequence divergence as the main determinant of the strengths of virtually all structure-evolution relationships, explaining $\sim 10 - 30\%$ of observed variation in sequence-structure relations. In addition to sequence divergence, we identify several protein structural properties that are moderately but significantly coupled with the strength of sequence-structure relations. In particular, proteins with more homogeneous back-bone hydrogen bond energies, large fractions of helical secondary structures and low fraction of beta sheets tend to have the strongest sequence-structure relations.

1 Introduction

Patterns of amino acid sequence variation are known to be influenced by the function of proteins (xx). The general consensus, based on the flurry of research done over the past several decades, is that the amino-acid sequence determines the 3D structure of proteins, known as the native conformation. This sequence-structure relation, however, does not necessitate a unique one-to-one mapping of sequence and the functionality of the protein. According to stability threshold model of proteins (xx), some substitutions at specific sites may be tolerated, if the new amino acid does not significantly change the energy landscape of the protein and therefore, its functional *native* conformation. Indeed, several independent work have identified site-specific structural properties that can explain the general patterns of sequence variability in proteins (xx). One of the earliest discovered examples of such relations, is the correlation of the site-specific Relative Solvent Accessibility (RSA) with sequence variation, that is, residues that are buried in the core of proteins tend to be more evolutionary conserved than exposed residues close to the surface of the protein.

Other structural properties have also been recently identified and proposed to influence or explain the site-specific evolutionary variations of proteins. Among the simplest properties is the residue *contact number (CN)*, a measure of local density of the protein defined as the number of amino acids within a spherical neighborhood of a specific residue of interest (xx). Variants of this quantity that eliminate the free-parameter (i.e., the radius of the spherical neighborhood) in the definition of CN have also been proposed (xx) and have been shown to correlate better with sequence evolutionary rates (xx). In a recent work Echave et al. xx presented a biophysical model that links the thermodynamic stability changes due to mutations at each site in proteins ($\Delta\Delta G$) to the rate at which mutations accumulate in the corresponding site over evolutionary time. They find that the variations in the free energy of the protein due to amino acid substitutions at individual sites can explain the site-specific evolutionary rates, comparable to the predictive powers of other structural variables such as residue solvent accessibility and contact number.

Although the majority of proteins exhibit some degree of correlation and association between sequence variation and structural properties, the strength of these correlations vary widely among different structures. By analyzing a data set of 216 monomeric enzymes, Yeh et al. xx found a wide range of $\rho \sim 0.1 - 0.8$ for the Spearman correlation strengths of sequence variability with two site-specific properties: the Weighted Contact Number (WCN) and the residue-specific solvent accessibility (i.e., RSA). Similarly, Echave et al. xx recently found a wide range of $\rho \sim 0.2 - 0.8$ for the correlation strength of the site-specific stability contribution – quantified by $\Delta\Delta G$ – with evolutionary rates. It appears that sequence-structure correlations tend to correlate strongly with each other, as illustrated in Figure **??**. This implies that for a given protein, the correlation strength of a specific structural property with evolutionary rates can serve as a proxy for the correlation strength of other structural properties with sequence evolutionary rates.

The fact that all relevant structural properties seem to have more or less the same predictive power for sequence variability, implies the existence of one or more structural or evolutionary characteristics of proteins that modulate sequence-structure correlations in all proteins. Motivated by these observations, here we present the results of comprehensive effort in search for the potential underlying structural or evolutionary properties of proteins that can explain the wide range of variations seen in correlation strengths of sequence evolutionary rates with different structural properties. We show that among all properties considered, sequence divergence appears to be the primary determinant for the strength of sequence-structure relations. In addition, we show that proteins with more homogeneous Hydrogen bond (H-bond) energies, higher fraction of Helical structures and lower number of β-sheets generally tend to exhibit the strongest sequence-structure correlations. In the following sections, we present evidence in support of these findings and discuss their implications.

2 Materials and Methods

Sequence Data, Alignments and Evolutionary Rates

The results presented in this work are based on a dataset of 209 monomeric enzymes (Echave papers, Wilke ddg paper xx) randomly picked from the Catalytic Site Atlas 2.2.11 (Porter et al. 2004) with protein sizes in the sample ranging from 95 to 1287, including representatives from all six main EC functional classes (Webb 1992) and domains of all main SCOP structural classes (Murzin et al. 1995). To assess the evolutionary rates at the amino acid level for each protein, first a set of up to 300 homologous sequences were collected by (Yeh et all xx) for each protein from the *Clean Uniprot* database following the ConSurf protocol (Goldenberg et al. 2009; Ashkenazy et al. 2010). Sequence alignments were then 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. The alignments were then used to calculate the site-specific evolutionary rates for each individual protein in dataset. To do so, we relied on two independent methods of measuring sequence variability measure. First, we calculated the Shannon entropy (H_i) – the sequence entropy, hereafter abbreviated as *seqent* – at each alignment column i, based on the assumption that the occurrence of each of the 20 amino acids is equally likely at any given site in the alignments:

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

where P_{ij} is the relative frequency of amino acid j at position i in the alignment. Alternatively, we also calculated a measure of site-specific evolutionary rate – hereafter abbreviated as *r4s* – for each protein using software rate4site (xx). To do so, first the Maximum Likelihood phylogenetic trees were inferred with RAxML, using the LG substitution matrix and the CAT model of rate heterogeneity (Stamatakis 2014). For each structure, we then used the respective sequence alignment and phylogenetic tree to infer site-specific substitution rates with Rate4Site, using the empirical Bayesian method and the amino-acid Jukes-Cantor mutational model (aaJC) (Mayrose et al. 2004).

Structural Properties

The goal of the presented work is to identify the prominent structural or evolutionary properties of proteins that modulate sequence-structure correlations. These potential modulators represent a unique characteristics of the protein as a whole. In general, the structural and evolutionary properties fall into two major categories. 1. *Residue-level properties*: Site-specific structural or evolutionary properties that are defined and calculated for each specific amino acid site in the protein sequence. Prominent examples of the site-specific structural properties include RSA (Tien et al. 2012 xx), WCN (shih? xx). 2. *PDB-level properties*: structural or evolutionary characteristics that are representative of the protein as a whole. Examples include pdb Contact Order (CO) as defined by xx, protein size and compactness, sequence length, structural resolution of the protein in X-ray crystallography. In addition, the distribution of each residue-level property can be summarized by its statistical moments as pdb-level property of the protein. Prime examples include, the mean and variance of WCN, RSA, sequence entropy, evolutionary rates. A comprehensive list of protein properties and their definitions are given in Table **??**.

Eliminating Degeneracy in Structural Property Definitions

In order to identify the potential determinants of sequence-structure correlations, we first ran a comprehensive search to identify site-specific structural properties that might correlate with measures of sequence variability (i.e., seqent & r4s). There are however degeneracies in the definition of the some site-specific variables. For example, the quantity WCN is generally calculated from the coordinates of α-carbon atoms in the 3-dimensional structure of proteins. There is however no reason to believe this set of atomic coordinates are the best representatives for individual sites in proteins. The same definition

degeneracy also exists for the set of atomic Bfactors (xx) that are used to represent site-specific flexibility, although the popular choice of residue flexibility is α-carbon atomic Bfactor (e.g., Halle 2001 xx).

Similar definition degeneracy also exists for the set of coordinates that can be used for Voronoi tessellation of proteins. A popular tool in condensed matter physics, Voronoi tessellation of a set of points (seeds) is a way of dividing the space into a number of regions such that for each seed there will be a corresponding region consisting of all points closer to that seed than to any other. These regions are called Voronoi cells. The structure of proteins can be considered as a set of 3D coordinates representing individual sites. Similar to WCN and Bfactor, there is also ambiguity as to which set of residue atomic coordinates best represent individual sites in proteins for the calculation of Voronoi cells.

Thus, for the sake of comprehensiveness and in order to identify the best definitions of structural properties such as WCN, Bfactor, and Voronoi cells, here we calculate and consider all possible definitions of properties depending on the choice of the representative set of atomic coordinates used. These include the set of coordinates of all backbone atoms (N, C, C_α, O) and the first heavy atom in the amino acid side chains (C_β). In addition, we calculate representative coordinates for each site in protein by averaging over the coordinates of all heavy atoms in the side chains. We also calculate a representative coordinate for each site that is an average over all heavy atom coordinates in the side chain and backbone of the amino acid. In rare cases where the side chain atoms are not resolved in the PDB file or the amino acid lacks the heavy atom needed (e.g., C_β for Glycine). The coordinate for that specific site is replaced with the coordinate of the corresponding C_α atom in the amino acid backbone.

We use VORO++ software (xx) to calculate the relevant Voronoi cell properties of all sites in all proteins, and use DSSP (xx) for the calculation of Accessible Surface Area (ASA) for each site normalized by the theoretical maximum solvent accessibility values of Tein et al (20112 xx) to obtain the Relative Solvent Accessibility (RSA) for all individual sites in all proteins. In addition to ASA values, we also extract from DSSP output, information about the secondary structure of proteins such as the total number of residues participating in different types of helices, parallel or anti-parallel beta sheets, or loops and turns. To complete the list of pdb-level structural properties, we also calculate the Spearman correlations between all residue-level structure and sequence properties and include them in the analysis to probe their potential effects on the strength of structure-sequence relations.

All data including a list of 209 proteins and their properties together with Python, R and Fortran codes written for data reduction and analysis are publicly available to view and download at `https://github.com/shahmoradi/cordiv`.

3 Results

Average Side Chain coordinates as the Best Representation of Protein 3D Structure

As explained in previous section, there is a high level of redundancy in the initial set of collected protein properties. In particular, depending on the set of atomic coordinates used, there are 7 different measures for some residue characteristics such as the residue Weighted Contact Number, Bfactor and Voronoi cell properties. This in turn results in a large set of secondary variables at pdb-level that basically measure the same protein characteristics, but with different strengths. Therefore, in order to eliminate redundant variables from dataset, we first compare the predictive power of different measures of residue characteristics based on the set of atomic coordinates used.

For the measure of local packing density in proteins (the Weighted Contact Number) we find that among all possible set of coordinates, the average over coordinates of all heavy atoms of each individual side chain results in WCN values that show the best correlation with other structural and sequence prop-

erties, such as RSA, Voronoi cell properties, sequence entropy, and evolutionary rates. Specifically, WCN from average side chain coordinates (wcnSC) outperforms WCN based on C_α coordinates (wcnCA) in predicting RSA, $\Delta\Delta G$ entropy, sequence entropy and evolutionary rates (r4sJC) by a median Spearman correlation difference of 0.09, 0.10, 0.07 & 0.08, respectively (Figure **??**).

For the measure of local flexibility in proteins (Bfactor) we similarly find that among all 7 representative measures of site Bfactors, the average of Bfactor values over all heavy atoms of each individual side chain (bfSC) results in the best correlations with other structural and sequence properties. Specifically, bfSC outperforms the commonly used C_α Bfactor (bfCA) in predicting RSA, $\Delta\Delta G$ entropy, sequence entropy and evolutionary rates by a median Spearman correlation difference of 0.11, 0.12, 0.08 & 0.09, respectively (Figure **??**).

Similar to WCN and Bfactor, the Voronoi cell properties, most importantly the cell surface area, volume and the cell compactness also correlate best with other structure and sequence properties, only if the average side chain coordinates are used as the seeds of Voronoi cells (Figure **??**).

All observations clearly demonstrate that individual sites in proteins are best represented by the average properties of the side chains of amino acids in the corresponding sites. In particular, the strength of structure and sequence correlations decrease when moving from side chain to backbone atoms. An exception to this general pattern is the correlation of the hydrogen-bond energies of the sites, which correlate more strongly with site characteristics calculated based on the backbone atoms instead of side chain.

Based on the observations described in the previous paragraphs, we keep only variables measured from average side chain properties and coordinates throughout the rest of the analysis and omit all other similar measures that show only weaker correlations with other site-specific characteristics. The exclusion of these alternative measures results in a significant reduction in the number of pdb-level variables to be further analysed, without compromising generality and comprehensiveness of the analysis.

Sequence divergence as the main Determinant of Sequence-Structure Relation

In order to identify the potential contributing factors to the strength of sequence–structure correlations, we first employ one of the simplest nonparametric yet powerful tests of statistical dependence, that is, we construct the Spearman correlation matrix of all pdb-level structure and sequence properties. The choice of Spearman versus the popular Pearson's correlation measure is made in order to minimize the effects of any nonlinear variable relationships on the strengths of the correlations. The resulting correlation matrix reveals a myriad of pdb-level properties each having a small but nonzero contribution to the strength of the structure-sequence correlations.

A hierarchical clustering of the correlation matrix however, reveals two main independent factors that have the strongest influence on the strengths of sequence-structure correlations: 1. The sequence divergence as measured by the standard deviation of sequence entropy and evolutionary rates (denoted by *sd.seqent* & *sd.r4sJC*) among all sites in each protein structure. 2. The homogeneity of the hydrogen bond strengths among the back bone atoms of each protein structure, as measured by the standard deviation of hydrogen bond energies (denoted by *sd.hbe*) among all pdb sites. A reduced-size of the Spearman correlation matrix for the most influential factors on the two strongest sequence–structure (seqent/r4s – wcnSC/varea) relations is illustrated in Figure 2.

For the other weaker sequence–structure relations, i.e. the correlations of seqent/r4s with RSA, $\Delta\Delta G$ entropy (ddgent), and Bfactor (bfSC) we find other pdb-level properties that also contribute to the correlation strengths, comparable to or even stronger than in sequence divergence and hydrogen bond

homogeneity. In general, we observe that for the weaker the sequence-structure correlations, factors that determine the accuracy of the measured residue properties become more influential on the strength of the correlations. In particular, the X-ray crystallographic resolution of the structure and the definition of the $\Delta\Delta G$ entropy play dominant roles, with Spearman correlation coefficients of ρ 0.3, on the strengths of the corresponding sequence-structure relations.

To ensure the accuracy of the results obtained from the Spearman correlation matrix of the pdb-level properties, we also use multivariate linear regression models, with individual sequence-structure correlations as the sole regressand of the regression models, and the set of pdb-level properties as the explanatory variables. Since the number of explanatory variables is comparable to the number of observations (i.e., the number pdb structures in the dataset), we use regularized regression (reference to R package xx) on the entire dataset, and also on the rank transformation of the dataset in order to minimize the effects of potential nonlinearities in data. Depending on value of the free parameter α, this generalized regression model is a compromise between *ridge regression* – which attempts to shrink the coefficients of correlated predictors towards each other – and *lasso regression* – which tends to pick one of the correlated predictors and discard the rest. In addition to regularized regression, we have also employed Principal Component Regression (PCR) on the original dataset and its rank transformation. Both regression methods, PCR & regularized, point to similar set of pdb-level properties as the strongest determinants of sequence-structure correlations.

4 Discussion and Concluding Remarks

Throughout this work, we have carried out a comprehensive analysis in search for the main determinants of the strength of sequence–structure correlations – some of which are newly reported and discussed in this work. Examples of sequence–structure relations include the correlations of sequence entropy (*seqent*) and measures of evolutionary rates (such as *r4sJC* used in this work) with measures of residue Contact Number (e.g., *wcnSC*), Relative Solvent Accessibility (RSA), $\Delta\Delta G$ entropy (*ddgent*). In addition, we have derived new site-specific properties, based on Voronoi Tessellation of protein 3D structures, that are comparable to or better than several previously known structural properties in explaining site-specific sequence entropy or evolutionary rates (e.g., Figure 1). Prime examples include Voronoi cell volume (*vvolume*), surface area (*varea*) and Voronoi cell sphericity defined as,

$$\Psi = \frac{\pi^{\frac{1}{3}}(6V)^{\frac{2}{3}}}{A}. \tag{2}$$

in which V & A represent *vvolume* & *varea* respectively. We have also shown that site-specific structural properties – such as Weighted Contact Number, Bfactor and Voronoi Cell properties – that are calculated from the average coordinates of side chain atoms, have the best explanatory powers for the sequence variability measures such as *seqent* and *r4sJC*. Compared to the common choice of back bone *CA* atomic coordinates, site-specific properties averaged over side chain atoms can outperform in predicting sequence evolutionary rates by as much as 0.12 in terms of Spearman correlation strength.

In search for the determinants of the strength of sequence-structure relations, we compiled a set of more than 200 protein properties for a dataset of 209 monomeric enzymes. By employing several independent parametric and non-parametric statistics, such as Spearman rank test, regularized regression and Principal Regression methods, we identify sequence divergence as the dominant factor in the strength of sequence-structure correlations, capable of explaining $10 - 30\%$ of the observed correlation strengths alone, in both the original and rank-transformed data.

ACKNOWLEDGEMENTS

....

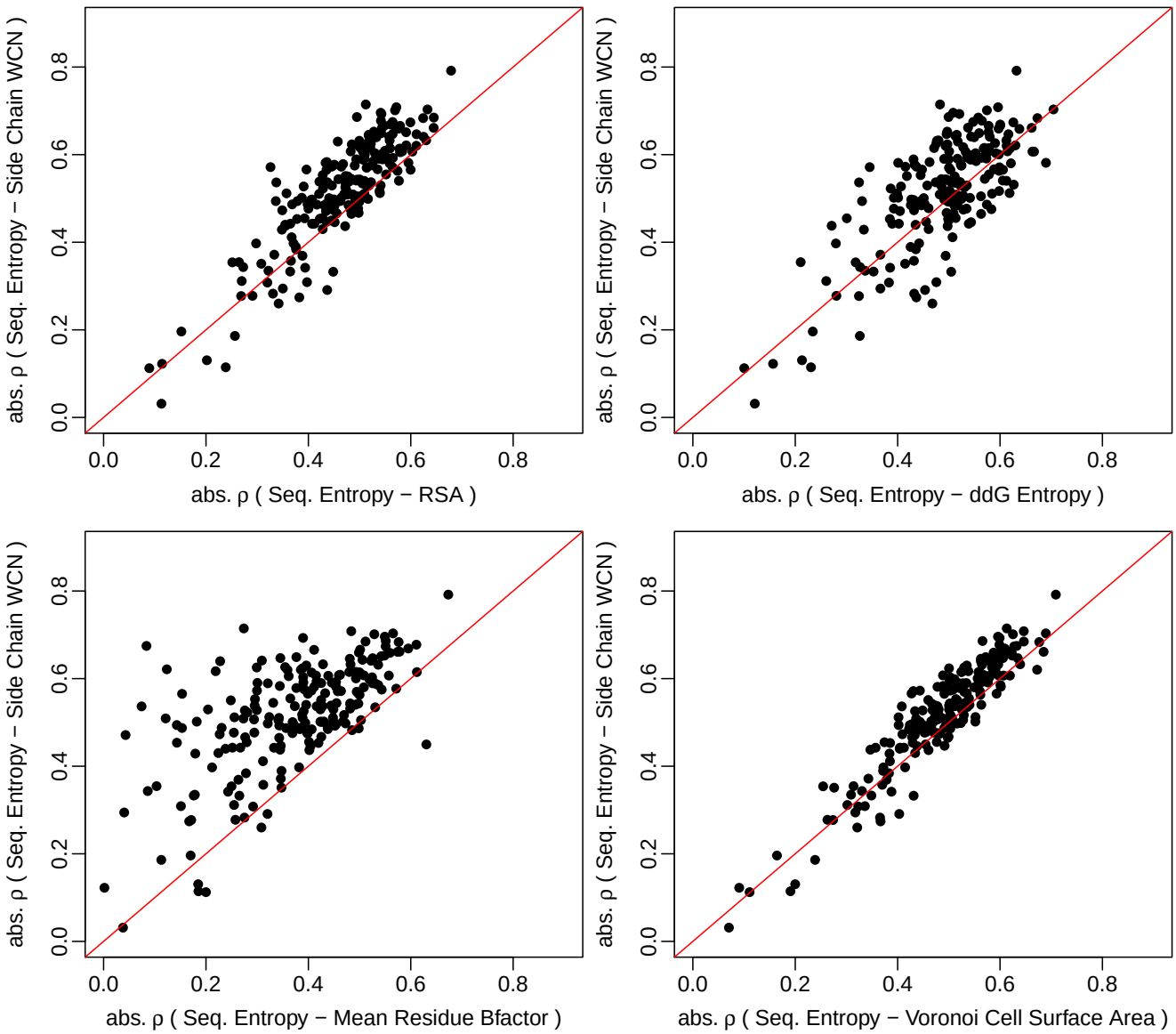

Figure 1: A comparison of the strength of Spearman correlation of sequence evolutionary rates (r4sJC) with *side chain* Weighted Contact Number vs. correlations of other structural properties with evolutionary rates. Detailed description of the structural properties is given in Section . The red lines in each plot represent equality line. It is evident from all plots that for any given protein in dataset, the correlation strength of one structural property is a good proxy measure of the correlation strength of any other structural property with sequence variability measures. For brevity, correlations of structure-rate4site are not shown here but are available online, also in supplementary material.

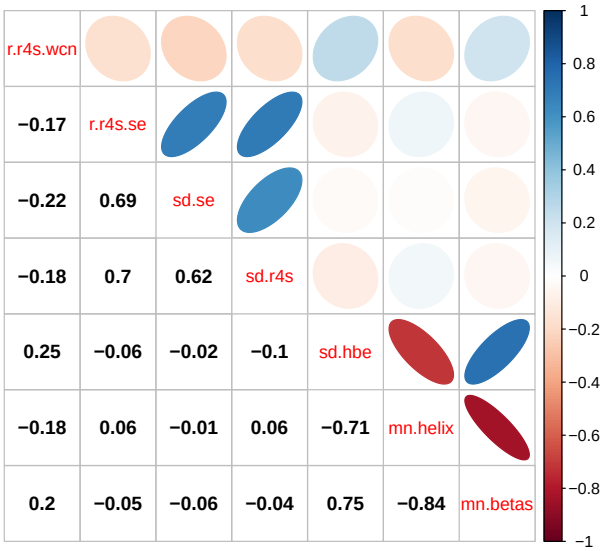

Figure 2: **The Spearman correlation matrix of the strongest sequence–structure correlations and the prominent determinants of the strengths of the corresponding relations**. The variables on the diagonal elements of the matrix from top to bottom represent respectively, the four strongest sequence–structure relations: r4sJC–wcnSC, r4sJC–varea, seqent–wcnSC, seqent–varea, followed by protein properties that modulate the strength of these relations: the correlation between r4sJC–seqent (*r.r4sJC.seqent*), variance of r4sJC (*var.r4sJC*), variance of sequence entropy (*var.seqent*), variance of back-bone hydrogen bond energies (*var.hbe*), and the fraction of amino acids in helical & $\beta-$sheet secondary structures in the proteins (*mn.helix* & *mn.betas* respectively).

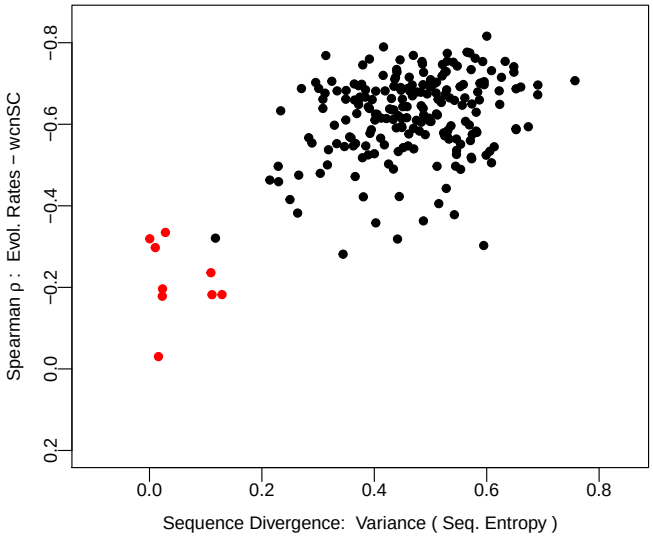

Figure 3: **Sequence–structure correlation strength versus sequence divergence.** The plot illustrates the relationship between the strength of a representative sequence–structure correlation (*seqent–scnSC*) and the sequence divergence as measured by the variance of protein sequence entropy. The black circles represent 209 proteins used in this work. For comparison and validation, the red circles represent data for 9 viral proteins taken from shahmoradi2014xx

.

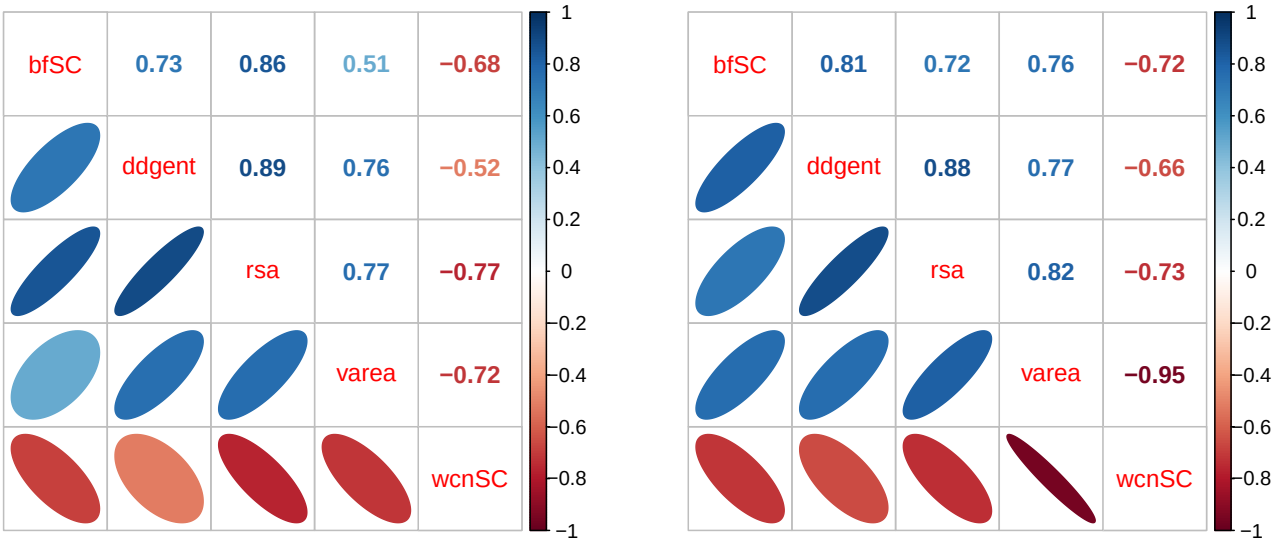

Figure 4: **Spearman correlation matrices illustrating the similarity of the set of structure and sequence properties that contribute to the strength of sequence–structure relations**. **Left:** The correlation matrix for the relation of evolutionary rates (*r4sJC*) with structural quantities on the diagonal of the matrix. **Right:** The correlation matrix for the relation of sequence entropy (*seqent*) with structural quantities on the diagonal of the matrix.

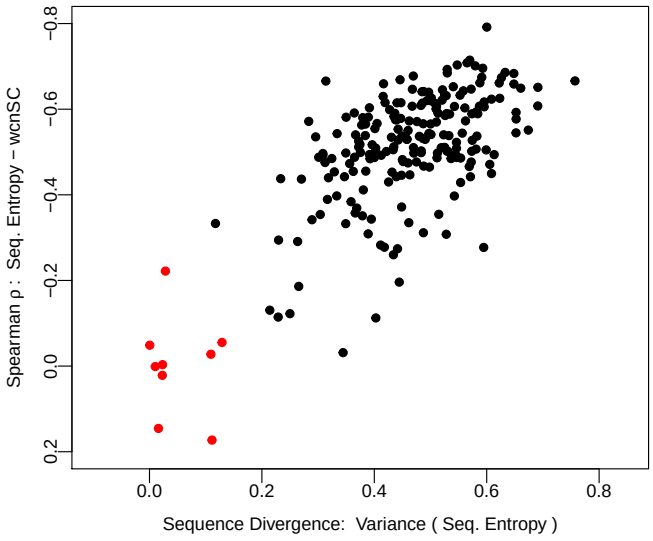

Figure 5: **Sequence–structure correlation strength versus sequence divergence.** The plot illustrates the relationship between the strength of a representative sequence–structure correlation (*seqent–scnSC*) and the sequence divergence as measured by the variance of protein sequence entropy. The black circles represent 209 proteins used in this work. For comparison and validation, the red circles represent data for 9 viral proteins taken from shahmoradi2014xx

.