October 21, 2014

Sequence Divergence as the main modulator of structure-sequence correlations

Amir Shahmoradi, Claus Wilke, Eleisha Jackson, ...

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 site-specific structural properties of proteins are capable of predicting the site-specific evolutionary sequence variations. The strength and significance of these structure-evolution relations, however, appear to vary widely among different proteins, with correlations ranging from 0.1 to 0.8. In search for potential determinants of the observed variation in sequence-structure correlations, here we present the results of a comprehensive study of interrelations among more than 200 structural and evolutionary characteristics in a dataset of 209 monomeric enzymes. We identify sequence divergence as the main determinant of the strengths of virtually all structure-evolution relationships.

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), substitutions at specific sites may be tolerated more than other sites in the protein sequence, if the substitution does not significantly change the energy landscape and therefore the functional conformation of the protein. Indeed, several independent work have identified site-specific structural properties that can explain the general patterns of sequence variability in proteins (xx). As one of the earliest discovered examples, 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 identified recently and proposed to influence or explain the site-specific evolutionary variations of proteins. Among the simplest properties is 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, named residue *contact number (CN)*(xx). Variants of this quantity that eliminate the free-parameter (i.e., radius) 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 sites in proteins ($\Delta\Delta G$) to the rate at which mutations accumulate at the corresponding sites 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.

Not all proteins exhibit the same correlation strength and association between sequence variation and structural properties. By analyzing a data set of 216 monomeric enzymes, Yeh et al. xx find a wide range of $\rho \sim 0.1 - 0.8$ for the distributions of the correlations of sequence variability with two site-specific properties: contact number and residue solvent accessibility. Similarly, Echave et al. xx find 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. Interestingly however, it appears that sequence-structure correlations also 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 evolution, implies the existence of one or more structural or evolutionary characteristics of protein that modulate sequence-structure correlations in all proteins. Motivated by this observation, 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 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.

2 Materials and Methods

Sequence Data, Alignments and Evolutionary Rates

The results presented in this work are based on data set of 213 monomeric enzymes 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 and including representatives of 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 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 former include site-specific Relative Solvent Accessibility (RSA) (Tien et al. 2012 xx), Weighted Contact Number (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, protein compactness. In addition, the distribution of each residue-level property can be summarized by its statistical moments as pdb-level property of the protein. 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 calculate the relevant Voronoi cell properties of all sites in all proteins using the software VORO++ (xx), 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 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. All

All data including a list of 213 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

4 Discussion and Concluding Remarks

ACKNOWLEDGEMENTS

....

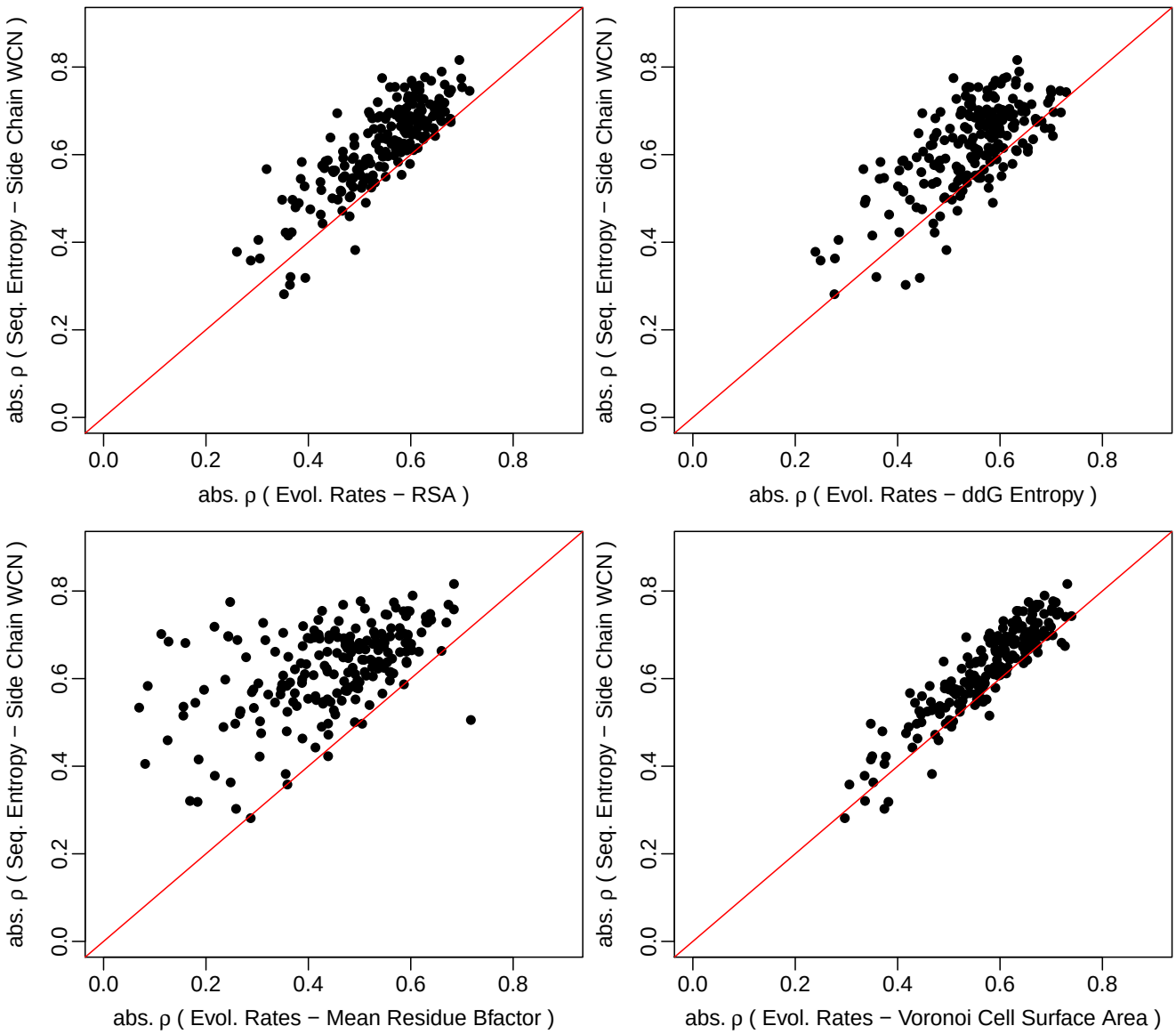

Figure 1: A comparison of the strength of Spearman correlation of sequence evolutionary rates (r4sJC) with *side chain* Weighted Contact Number vs. correlations of other structural properties with evolutionary rates. Detailed description of the structural properties is given in Section . The red lines 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.