Sequence divergence as the main determinant of sequence-structure relationships

Amir Shahmoradi[1*], Eleisha L. Jackson[2], Claus O. Wilke[2]

May 13, 2015

[1] Department of Physics, The University of Texas at Austin, Austin, TX 78712, USA
[2] Institute of Cellular and Molecular Biology, Center for Computational Biology and Bioinformatics, and Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, 78712 USA

[*]Corresponding author
Email: amir@physics.utexas.edu
Phone: +1 512 232 2459

Manuscript type: research article
Keywords: protein evolution, relative solvent accessibility, site variability

Abstract

Recent work has shown that structural properties are capable of predicting site-specific sequence variability for a given protein. However, the strength and significance of these structure-sequence relations 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 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 structural-evolution relationships, explaining $10-30\%$ 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 correlations between structural properties and site variability.

1 Introduction

The result in this paper: Site variabilty is the main determine of the difference in structure-sequence correlations. This is the main. It might be good to mention that we looked at 200 other predictors and found this as the more important one. The issue: There is a difference between the Huang et al paper vs. Sharamoradi et al. In Echave's paper with Huang or Yeh?, they say that the CN is the best predictor. In Amir's paper, we say that RSA is still the best predictor. Why is this different? The difference is in the datasets in terms of variability. This paper should have about 3 figures according to Claus possibly. So far I see 4 plots that need to be made:
1. WCN-Rate4site vs Mean RateSite??? Does this make sense? (Also for RSA)
2. WCN-Entropy vs Mean Entropy
3. RSA-Entropy vs Mean Entropy
Open Questions: Should I make the stress and delta delta G correlations plots for the viral sequences that would complete the story?

Proteins are subject to a number of biophysical and functional constraints (Scherrer et al., 2012; Wilke and Drummond, 2010). These constraints result site-specific patterns of sequence variability within a protein. Recently several site-specific structural properties that can explain patterns of sequence variability in proteins have been identified. One of the earliest examples is Relative Site Accessibility (Ramsey et al., 2011; Franzosa and Xia, 2009). Residues that are buried in the core of proteins tend to be more conserved than exposed residues close to the surface of the protein. Another structural property that has been found to significantly correlate with with site-specific variation is local packing density, often measured by Contact Number (CN) or Weighted Contact Number (WCN) (Liao et al., 2005; Yeh et al., 2014; Huang et al., 2014)

In a recent work, Echave et al. (2015) 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 found 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 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 proteins. By analyzing a data set of 216 monomeric enzymes, Yeh et al. (2014) 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 RSA. Similarly, Echave et al. (2015) 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 implying that several structural properties are representing similar biophysical constraints on site-specific evolutionary rates. SHOULD THIS BE IN RESULTS?? (Figure 1). 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. Where is the stress model, include it!!! Need to mention Huang Paper somewhere!

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. The underlying principles behind the wide variation in correlation strengths among proteins in not well understood. Here we present the results of a comprehensive 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. What?? Can we discuss divergence when talking about entropy? appears to be the primary determinant for the strength of sequence-structure relationships. 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.

2 Materials and Methods

Structures, sequences, and measures of sequence properties

The results presented in this work are based on a dataset of 209 monomeric enzymes (Echave et al., 2015) 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 al., 2014; Huang et al., 2014) for each protein from the *Clean Uniprot* database following the ConSurf protocol (Goldenberg et al., 2008). 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. 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:

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

where P_{ij} is the relative frequency of amino acid j at position i in the alignment. The sequence entropy is a measure of variability at each site. We also calculated a measure of site-specific evolutionary rate – hereafter abbreviated as *r4s* – for each protein using software Rate4site (Citation??). First the Maximum Likelihood phylogenetic trees were inferred with RAxML, using the LG substitution matrix and the CAT model of rate heterogeneity (Stamatakis, 2006, 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).

Calculation of Structural Properties

The goal of the presented work is to identify the prominent structural or sequence 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 (Franzosa and Xia, 2009; Scherrer et al., 2012; Yeh et al., 2014) and WCN (Shih et al., 2012; Yeh et al., 2014) Should I cite Lin paper?. 2. *PDB-level properties*: structural or evolutionary characteristics that are representative of the protein as a whole. Examples include protein Contact Order (CO) as defined by Need Citation!, protein size and compactness, sequence length, and structural resolution of the protein in X-ray crystallography. RE-WRITE THIS In addition, the distribution of each residue-level property can be summarized by its statistical moments as protein-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 **??** Where is this table???!!!.

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. We use VORO++ software (Need citation!) to calculate the relevant Voronoi cell properties of all sites in all proteins, and use DSSP (Kabsch and Sander, 1983) for the calculation of Accessible Surface Area (ASA) for each site. We normalized the ASA for each site by the theoretical maximum solvent accessibility values of Tien et al. (2013) 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.

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 (Citation???) that are used to represent site-specific flexibility,

although the popular choice of residue flexibility is α-carbon atomic Bfactor (Halle, 2002). 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. 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.

All data and analysis scripts required to reproduce the work 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

CONDENSE THIS WHOLE SECTION. THIS IS COVERED IN HIS PAPER. 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.m For the WCN 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 properties, 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 3).

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

COLOR Delta Delta G dor viral proteins??? 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 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 backbone atoms of each protein structure, as measured by the standard deviation of hydrogen bond energies (denoted by *sd.hbe*) among all protein 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

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.

5 Acknowledgements

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing High-performance computing resources. ELJ is is funded by a National Science Graduate Research Fellowship, grant number DGE-1110007. COW is funded by Which grants??. AS is funded by Which grants??. ....

References

Echave J, Jackson EL, Wilke CO. **2015**. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *Phys. Biol* **13**:025002.

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

Goldenberg O, Erez E, Nimrod G, Ben-Tal N. **2008**. The consurf-db: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.* **37**:D323–D327.

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

Huang TT, Marcos ML, Hwang JK, Echave J. **2014**. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rate sand their relationship with packing and flexibility. *BMC Evol. Biol.* **14**:78.

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 KI, Toh H, Miyata T. **2005**. Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**:511–518.

Katoh K, Misawa K, Kuma K, Miyata T. **2002**. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.* **30**:3059–3066.

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

Mayrose I, Graur D, Ben-Tal N, Pupko T. **2004**. Comparison of site-specific rate-inference methods for protein sequences; empirical bayesian methods are superior. *Mol. Biol. Evol.* **21**:1781–1791.

Murzin AG, Brenner SE, Hubbard T, Chothia C. **1995**. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol Biol.* **247**:536–540.

Porter CT, Bartlett GJ, Thornton JM. **2004**. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* **32**:D129–D133.

Ramsey DC, Scherrer MP, Zhou T, Wilke CO. **2011**. The relationship between relative solvent accessibility and evolutionary rate in protein evoltion. *Genetics* **188**:479–488.

Scherrer MP, Meyer AG, Wilke CO. **2012**. Modeling coding-sequence evolution within the context residue solvent accessibility. *BMC Evol. Biol.* **12**:179.

Shih CH, Chang CM, Lo WC, Hwang JK. **2012**. Evolutionary information hidden in a single protein structure. *Proteins* **80**:1647–1657.

Stamatakis A. **2006**. Raxml-v1-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**:2688–2690.

Stamatakis A. **2014**. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313.

Tien MZ, Meyer AG, Sydykova DK, Spielman SJ. **2013**. Maximum allowed solvent accessibilities of residues in proteins. *PLOS ONE* **8**:e80635.

Webb EC. **1992**. Enzyme nomenclature 1992: recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes. *San Diego: Academic Press* .

Wilke CO, Drummond DA. **2010**. Signatures of protein biophysics in coding sequence evolution. *Curr. Opin. Struct.* **20**:385–9.

Yeh SW, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J. **2014**. Site-specific structural constraints on protein sequence evolutionary divergence: Local packing density versus solvent exposure. *Mol. Biol. Evol.* **31**:135–139.

Figures

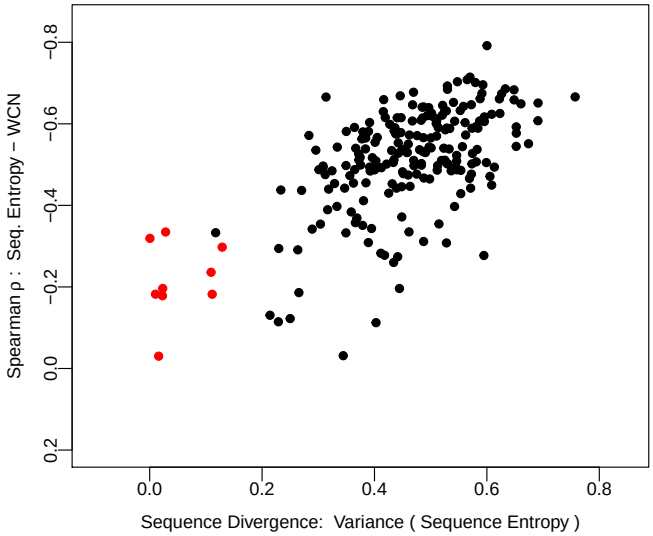

Figure 1: **Sequence–structure correlation strength versus sequence divergence.** The plot illustrates the relationship between the strength of a representative sequence–structure correlation (*seqent–scnSC*) and the sequence divergence as measured by the variance of protein sequence entropy. The black circles represent 209 proteins used in this work. For comparison and validation, the red circles represent data for 9 viral proteins taken from shahmoradi2014xx

.