Sequence divergence as the main determinant of sequence-structure relationships

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

May 15, 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

NEED TO EDIT THE 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

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. Franzosa and Xia (2009) identified RSA as the strongest predictor of evolutionary rate. They found that residues that are buried in the core of proteins tend to be more conserved than exposed residues close to the surface of the protein. In their analysis these concerned both RSA and various definitions of residue packing density to predict evolutionary rate. They found that RSA and evolutionary rate shared a significant linear relationship. Afterwards, several other works (Ramsey et al., 2011; Scherrer et al., 2012) also found that RSA as a very significant predictor of evolutionary rate and found this linear relationship as well. However in these papers all have the same flaw. During the course of their analysis they binned the protein sites and average over all sites within a bin when looking at the trend of RSA. This could produce artifacts when looking at the trend.

Later, Yeh et al performed a similar analysis on a series of enzyme monomer proteins and found that packing density, as defined by CN and WCN (Liao et al., 2005; Yeh et al., 2014; Huang et al., 2014), was the strongest determinant of site variability. A year later, Sharamoradi et al also performed a site-wise analysis on a series of viral proteins. In this analysis they found that RSA had the strongest correlation with site variability. Additionally the effect seen between CN and WCN was of a much smaller magnitude as compared to (Yeh et al., 2014). Here we attempt to reconcile the work done in this area. We find that site variability is the primary determinant of structure determinant of the strength of structure-sequence relationships and some differences in previous work can be explained in terms of differing levels of site variability.
COPY EDIT THIS!!!

2 Materials and Methods

Structures, sequences, and measures of sequence properties

The results presented in this work are based on a two datasets. The first is a dataset of 209 monomeric enzymes from (Echave et al., 2015) originally from Huang et al. (2014).The original dataset was comprised of 213 but we removed four of the proteins (1BBS, 1BS0, 1DIN, 2HPL) that had did not have data at insertion sites. Briefly, these proteins are all enzyme monomers 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. For each structure we had a corresponding alignment of up to 300 homologous sequences. The second dataset was from taken from Sharamoradi et al. and is comprised of nine? viral proteins. The viral proteins range from 122 - 557 in length and each structure is accompanied by a sequence alignment of up to 2362 homologous sequences. Sequence alignments for both datasets 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 both datasets. 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. 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

HOW IS CN and WCN Defined and Calculated??? (Cite Shih et al for WCN) Possbily Lin Paper

We used 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.

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

SOME RESULTS.....

4 Discussion

SOME DISCUSSION.....

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

1. WCN-Rate4site vs Variance of RateSite (Also for RSA) and CN-Rate4site vs Variance of RateSite
2. WCN-Entropy vs Mean Entropy and CN-Entropy vs Mean Entropy
3. RSA-Entropy vs Mean Entropy
4. RSA vs Rate4Site
5. Should I make the stress and delta delta G correlations plots for the viral sequences that would complete the story?

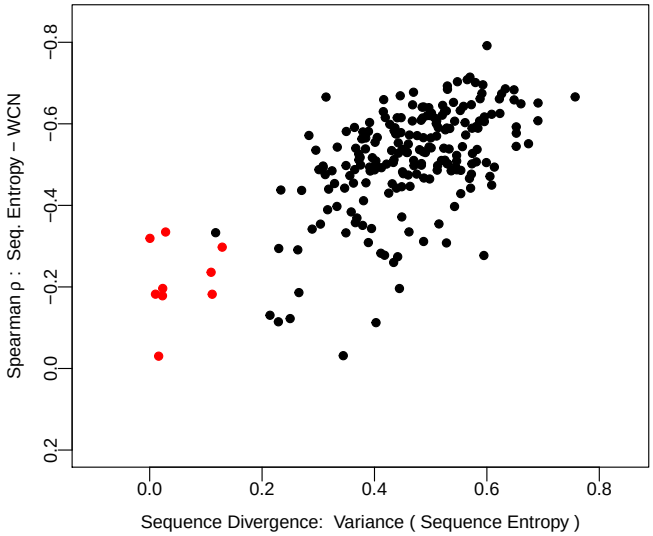

Figure 1: **Sequence–structure correlation strength versus sequence divergence.** The plot illustrates the relationship between the strength of a representative sequence–structure correlation (*seqent–scnSC*) and the sequence divergence as measured by the variance of protein sequence entropy. The black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

.