Variation in site entropy explains differences in structure-sequence relationships of proteins

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

June 2, 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

[*]These authors contributed equally.

[**]Corresponding author
Email: wilke@austin.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. There have two works that have investigated structural predictors of site variability which both present different results based on the strength of correlations of structural predictors. According to Yeh et al. (2014), local packing density is the strongest predictor of site-wise variability. However, recently, Shahmoradi et al. (2014) compared local packing density and solvent accessibility in viral proteins and found that relative solvent accessibility is a stronger predictor of site-wise variability. Here we present research that suggests that differences in the correlations between these two datasets are due to differences in the site variability among proteins within in each dataset. Specifically proteins with a larger variance in entropy among sites exhibit stronger structure-sequence correlations between both local packing density and solvent accessibility.

1 Introduction

Proteins are subject to a number of biophysical and functional constraints (Scherrer et al., 2012; Wilke and Drummond, 2010). These constraints result in 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 was Relative Site Accessibility (RSA). Franzosa and Xia (2009) identified RSA as the strongest predictor of evolutionary rate and found that residues that are buried within the core of proteins tend to be more conserved than exposed residues close to the surface of the protein. In their analysis, they considered the ability of 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 also found that RSA as a significant predictor of evolutionary rate and supported this linear relationship (Ramsey et al., 2011; Scherrer et al., 2012). However, these papers all have the same flaw. During the course of their analysis they binned the protein sites and averaged over all sites within a bin when determining the trend of RSA. This process may have produced artifacts that account for this strong linear trend between RSA and evolutionary rate.

Recently, Yeh et al. (2014) performed a similar analysis on a series of enzyme monomer proteins and found that packing density, as defined by Contact Number and Weighted Contact Number (Liao et al., 2005; Yeh et al., 2014; Huang et al., 2014), was the strongest determinant of site variability. Soon Afterwards, Shahmoradi et al. (2014) 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 as opposed to local packing density. Additionally the effect seen between CN and WCN was of a much smaller magnitude as compared to Yeh et al. (2014). Therefore the relationship between local pacing density, RSA and site-specific measures of sequence variability is not well understood. Here we attempt to reconcile the work done in this area. We find that site variability is the primary 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.

2 Materials and Methods

Structures, sequences, and measures of sequence properties

The results presented in this work are based on two datasets. The first is a dataset of 209 monomeric enzymes taken from Huang et al. (2014), originally from Yeh et al. (2014).The original dataset was comprised of 213 proteins but we removed four of the proteins (PDB IDS: 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 residues in length. For each structure we had a corresponding alignment of up to 300 homologous sequences. The second dataset was taken from Shahmoradi et al. (2014) and is comprised of nine viral proteins. The viral proteins range from 122 - 557 residues in length and each

structure is accompanied by a sequence alignment of up to 2362 homologous sequences. Sequence alignments for both datasets were constructed by aligning the amino-acid sequences using the alignment software MAFFT (Katoh et al., 2002, 2005), specifying the auto flag to select the optimal algorithm for the given dataset. The alignments were then used to calculate site-specific measures of sequence variability for each individual protein in both datasets. To do so, we relied on two independent methods of measuring sequence variability. First, we calculated the Shannon entropy (H_i) – the sequence entropy 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. Sequence entropy is a measure of variability at each site. We also calculated a measure of site-specific evolutionary rate for each protein using the software Rate4site. First 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 JTT model of sequence evolution (Mayrose et al., 2004).

Calculation of Structural Properties

In our analysis we used two types of measures of local packing density used in previous studies: Contact Number (CN) and Weighted Contact Number (WCN). For the purposes of our comparison between the two datasets of interest we used the CN and WCN for the enzyme proteins calculated in Huang et al. (2014) and the CN and WCN values for the viral proteins from Shahmoradi et al. (2014). In both of these works and in Yeh et al. (2014), WCN and CN are both defined the same. Contact Number is defined as the number of C_α within a pre-redefined radius, r_0. In this case, $r_0 = 13$ as in the previous papers. Weighed Contact Number for a residue, i, is defined as in Liao et al. (2005) and Huang et al. (2014) as:

$$\mathrm{WCN}_i = \sum_{i \neq j}^{N} \frac{1}{r_{ij}^2} \tag{2}$$

where r_j is the length between the C_α of residue i and residue j in a protein of length N (Yeh et al., 2014).

We used DSSP (Kabsch and Sander, 1983) to calculate the Accessible Surface Area (ASA) for each site. We then 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.

All data and analysis scripts required to reproduce the work are publicly available to view and download at `https://github.com/wilkelab/rate_variability_variation`.

3 Results

We calculate the strength between the structural properties and site variability by calculating the Spearman correlation, ρ, between the structural property (solvent exposure and local packing density) and site variability, as measured by either evolutionary rate or site entropy. On average the correlations between CN and entropy are larger in absolute magnitude in the enzyme proteins as compared to viral proteins. The average ρ between each structural property and each measure of site variability (entropy and evolutionary rates) can be seen in Table 1 and Table 2. The viral proteins experience much lower site variability. Figure 1 shows the correlations between CN and entropy for the viral and enzyme proteins. Even though on average the correlations in the enzyme proteins are larger as reported by Yeh et al. (2014), the viral proteins still have correlation strengths that are comparable to some of the enzyme proteins with lower correlations.

For a given protein its mean entropy and the variance might be different. A protein can have a high mean entropy but have a low variance and vice versa. Figure 6A details the relationship between the average entropy across sites within proteins and the variance of entropy across sites. Additionally, the distribution of variance varies greatly between proteins even when they from the same dataset (Figure 6B). Therefore the mean entropy of a protein as well as the variance in entropy at sites may be predictive for structure-sequence relationships.

The average site entropy of a given protein does not seem to be a significant determinant of the strength of structural correlations (Figures 1A, 2A, 3A). However, when examining the variance of entropy there is a clear trend within the enzyme proteins. Proteins with a higher variance in site variability across the protein typically higher correlations in magnitude (Figures 1B, 2B, 3B). If you extrapolate the trend from the enzyme proteins, the viral proteins follow a similar trend. The correlation between the ρ between RSA and entropy and the variance of entropy is positive. Proteins with a larger variance in site entropy have the strongest correlations. The correlation between local packing density (both WCN and CN) and the variance of entropy is negative. Unlike entropy, there is no relationship between the variance of evolutionary rates at sites and any of the measured structural properties (Figures 4, 5). There also is a wide spread in the variance of the evolutionary rates across proteins for both the enzyme and viral proteins.

Although both evolutionary rate, as measured by Rate4Sit,e and entropy are measures of site variability, these quantities are distinctly different. Rate4Site measures the rate at which a site changes over time whereas site entropy measure the absolute variation at a site. A site may have a high evolutionary rate if it changes frequently between the same few amino acids but have low entropy. It appears that the rate at which a site evolves is not important for predicting the strength of structure-sequence relationships. However, the absolute amount of variation at a site can be used to predict the strength of structure-sequence relationships.

In order to further examine the relationship between entropy and the structure-sequence relationships, we used the mean entropy and variance of entropy at sites as predictors of the strength of structure-sequence relationships. Table 4 illustrates the coefficients of various linear models. For WCN-Entropy correlations, mean entropy is not a significant predictor. Although for CN-Entropy correlations, mean entropy is significant the coefficient is 0.079 and therefore predictive power is low. Mean Entropy is also not a significant predictor for

the strength of the relationship between RSA and Entropy. For all linear models where mean entropy and dataset were used to predict structure-sequence correlations, dataset was a significant predictor (Table 4). This means that enzyme and virus proteins have different correlations when using mean entropy as a predictor. This agrees with differences seen in the previous works by Yeh et al. (2014) and Shahmoradi et al. (2014).

However, when looking at the variance in entropy there are some stark differences. For all structural predictors (i.e., CN, WCN and RSA), the variance in entropy at sites within a protein is a significant predictor of the strength of structure-sequence relations. For packing density, proteins with a larger variance in site-wise entropy have more negative correlations that are higher in magnitude. Proteins with a higher variance in entropy tend to have stronger RSA-Entropy correlations. When using variance and dataset as predictors of structure-sequence correlations, dataset was not a significant predictor in any model. When looking at the variance of entropy, proteins within both datasets at in a similar fashion and overall trends seen between the variance of site variation and the strength of structure-sequence relationships is preserved across both datasets.

4 Discussion

Structure imposes constraints on the evolution of proteins. Therefore structural predictors can be used to predict the variability at sites. Two main determinants of variability at protein sites are solvent accessibility and local packing density. Sites that are on surface of the protein tend to have higher solvent accessibility and exhibit more site variability. Sites that are densely packed and have more contacts tend to evolve slower and exhibit less sequence variability. However, it has not been well understood which of these two predictors, local packing density and solvent accessibility, is the main determinant of site variability in proteins. Two recent papers (Shahmoradi et al., 2014; Yeh et al., 2014) give conflicting reports on which is the best predictor. In a study done on monomeric enzyme proteins, Yeh et al. (2014) reported that local packing density, as measured by Contact Number and Weighted Contact Number, is the strongest determinant of site variability. Shahmoradi et al. (2014) performed a similar analysis on a series of viral proteins and found that solvent accessibility is more predictive of site variability.

Here we examined the relationship between site variability and the strength of structure-sequence relationships. We found that proteins with a larger variance in site variability as measured by sequence entropy, have stronger structure-sequence relationships on average. However, when looking at the mean entropy values of proteins there is a distinct difference between the virus and enzyme datasets. Overall the enzyme proteins have more site variability than the viral proteins.When predicting structure-sequence relationships the variance in site variability among proteins is a better predictor than mean site variability. Additionally the reason that the enzyme proteins in Yeh et al. (2014) had higher correlation coefficients is that on average, the enzyme proteins have larger variance in entropy as compared to the virus proteins of Shahmoradi et al. (2014). Moreover, since the variance in entropy is a more significant predictor of correlations in local packing density, an increased level of variance in site variability will have a larger effect on the local packing density correlations. This would explain why the local packing density correlations were of a higher magnitude compared to

the RSA correlations within the enzyme proteins of Yeh et al. (2014). THIS NEEDS TO BE EXPLAINED. It appears that site entropy and evolutionary rate are not equal measures of site variability in terms of predicting the strength of structure-sequence relationships. Finally, the variance in evolutionary rate among sites of a protein has a smaller an effect on the correlation magnitude as compared to entropy. This suggests that the absolute variability at sites is a better predictor of whether structural determinants can be used to predict variability at sites as opposed to evolutionary rate. Therefore when examining structural determinants of site variability, entropy is a may be a more meaningful quantity.

Amir's comment: I think it maybe better to not put so much emphasis on variance of entropy, but its implications and meaning. For example entropy variance is an indicator of sequence divergence, to some extent.

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 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

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

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.

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.

Shahmoradi A, Sydykova DK, Spielman SJ, Jackson EL, Dawson ET, Meyer AG, Wilke CO. **2014**. Predicting evolutionary site variability from structure in viral proteins: Buriedness, packing, flexibility, and design. *J. Mol. Evol.* **79**:130–142.

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.

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

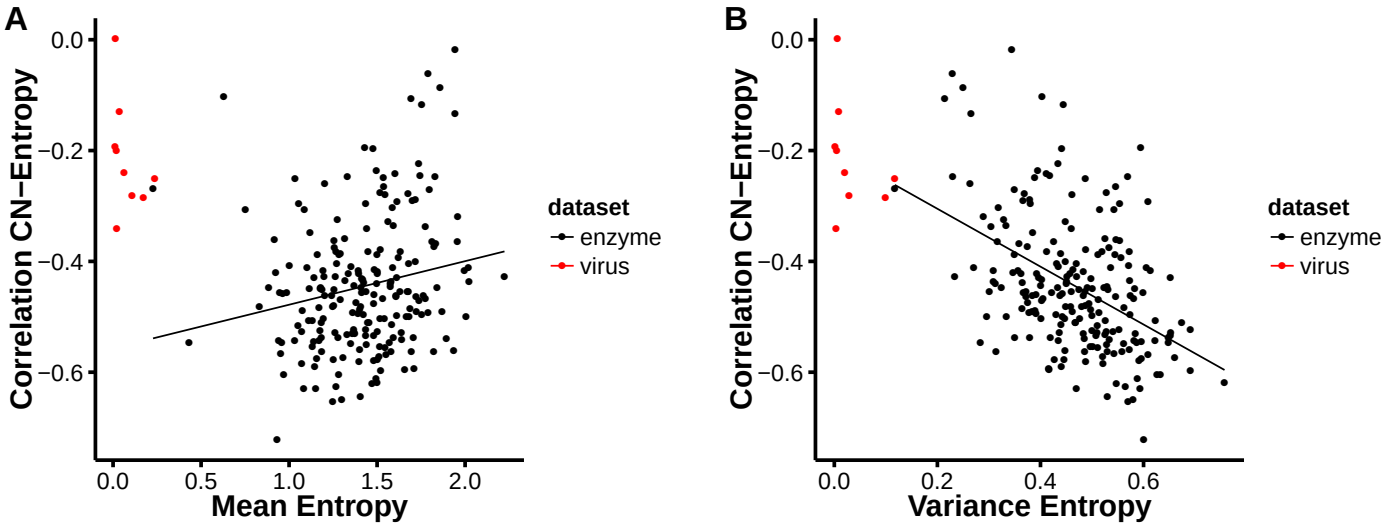

Figure 1: Correlations between Contact Number (CN) and Entropy. For each protein we calculate the Spearman Correlation coefficient between CN and entropy at each site within the protein. A) Comparison between the correlation coefficients and mean entropy of site in each protein. The line indicates a linear regression through the enzyme proteins with mean entropy as a single predictor of the spearman correlations. B) Comparison between correlation coefficients and the variance of entropy of each protein. The enzyme proteins are colored in black and the virus proteins are visualized in red. For both WCN and CN the viral proteins have lower mean entropy and variance of entropy. The line indicates a linear regression through the enzyme proteins with variance entropy as a single predictor of the spearman correlations. Proteins with a larger variance in site entropy among sites tend to smaller CN-Entropy correlations in magnitude.

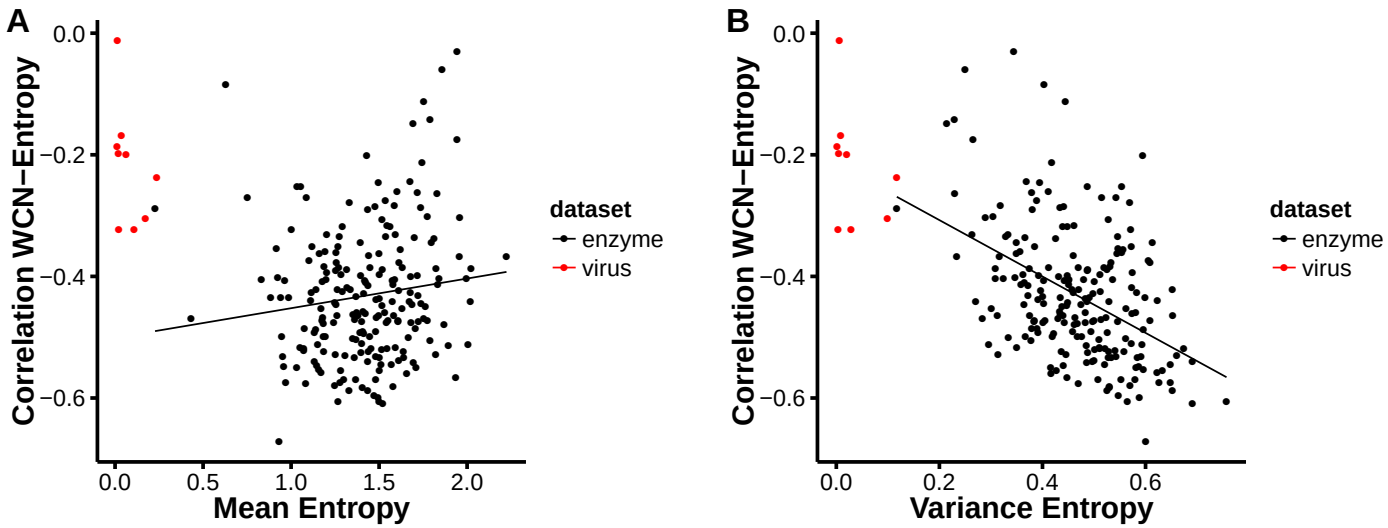

Figure 2: Correlations between Weighted Contact Number (WCN) and Entropy. A) Comparison between the correlation coefficients and mean entropy of site in each protein. B) Comparison between correlation coefficients and the variance of entropy of each protein. The enzyme proteins are colored in black and the virus proteins are visualized in red. Proteins with a larger variance in site entropy have smaller WCN-Entropy correlations in magnitude.

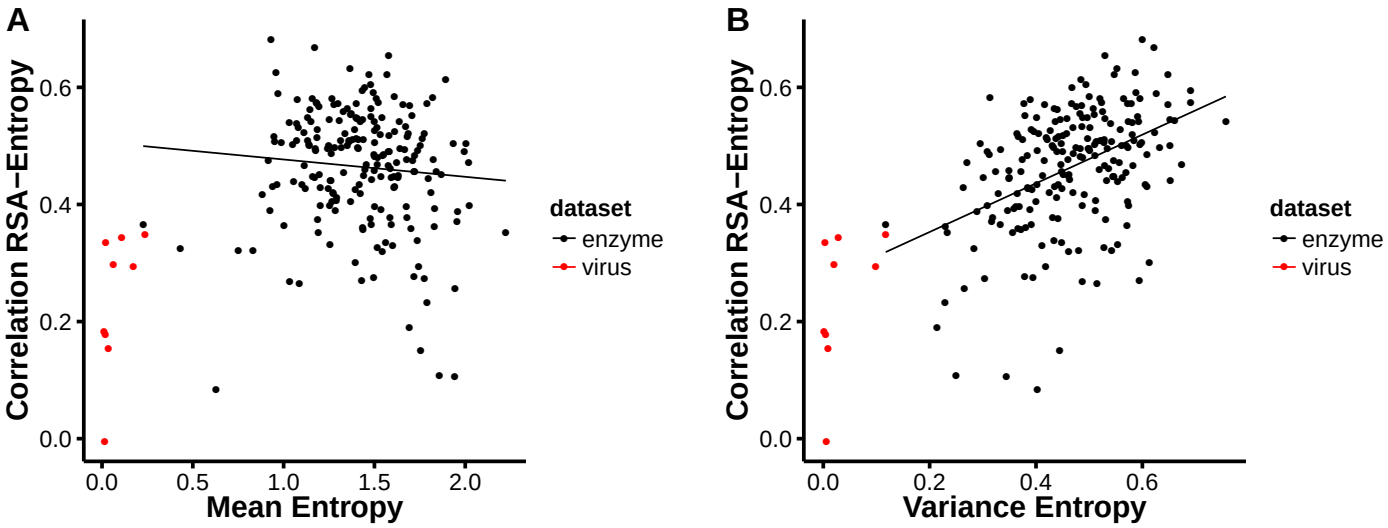

Figure 3: Correlations between Relative Solvent Accessibility (RSA) and Entropy. A) Comparison between the correlation coefficients and mean entropy of site in each protein. B) Comparison between correlation coefficients and the variance of entropy of each protein. The enzyme proteins are colored in black and the virus proteins are visualized in red. Proteins have a larger variance in site entropy have larger RSA-Entropy correlations.

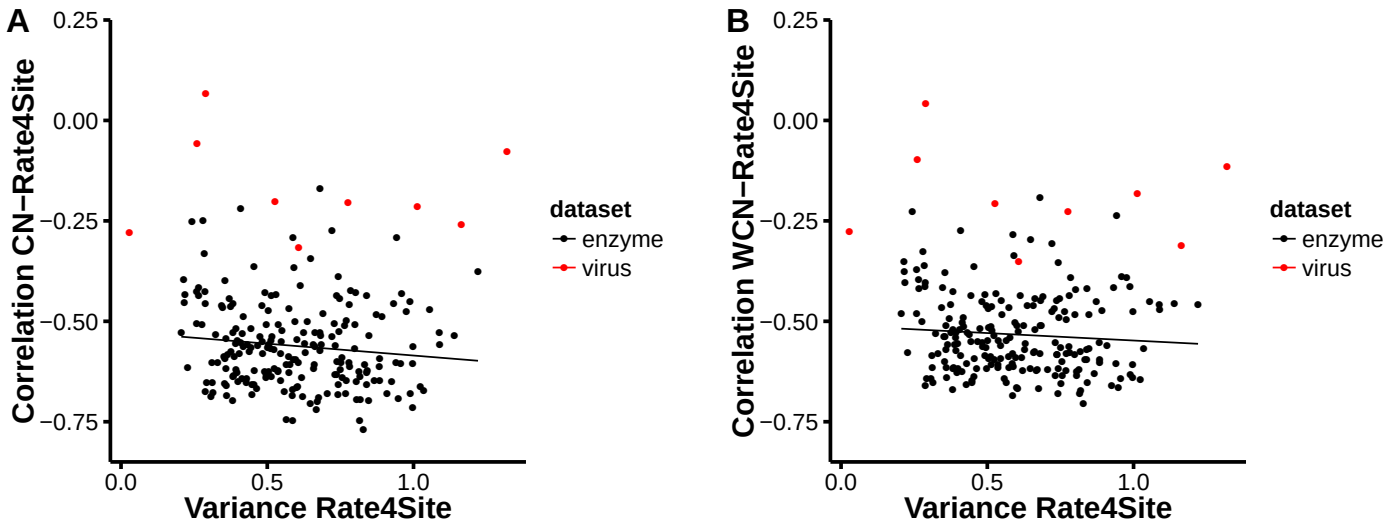

Figure 4: Comparison of Rate4Site correlations and the variance of Rate4Site at sites. A) Comparison between the correlation coefficients between CN and Rate4Site and variance of Rate4Site of site in each protein. B) Comparison between correlation coefficients between WCN and Rate4Site and the variance of entropy of each protein. The enzyme proteins are colored in black and the virus proteins are visualized in red. There is no observable trend between the variance of evolutionary rate and the spearman correlations.

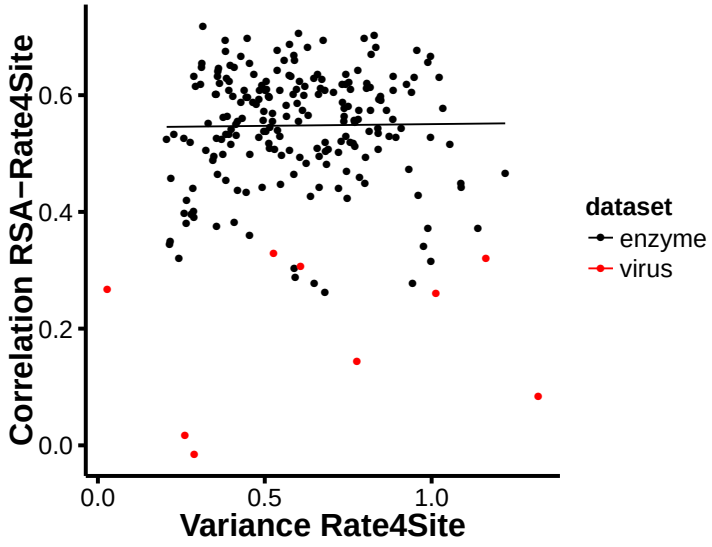

Figure 5: Comparison of the correlations between RSA and Rate4Site and the variance of Rate4Site at sites. The enzyme proteins are colored in black and the virus proteins are visualized in red. The variance of evolutionary rate of a protein, as measured by Rate4Site, does not have a correlation with the spearman correlation between RSA and evolutionary rate for a given protein.

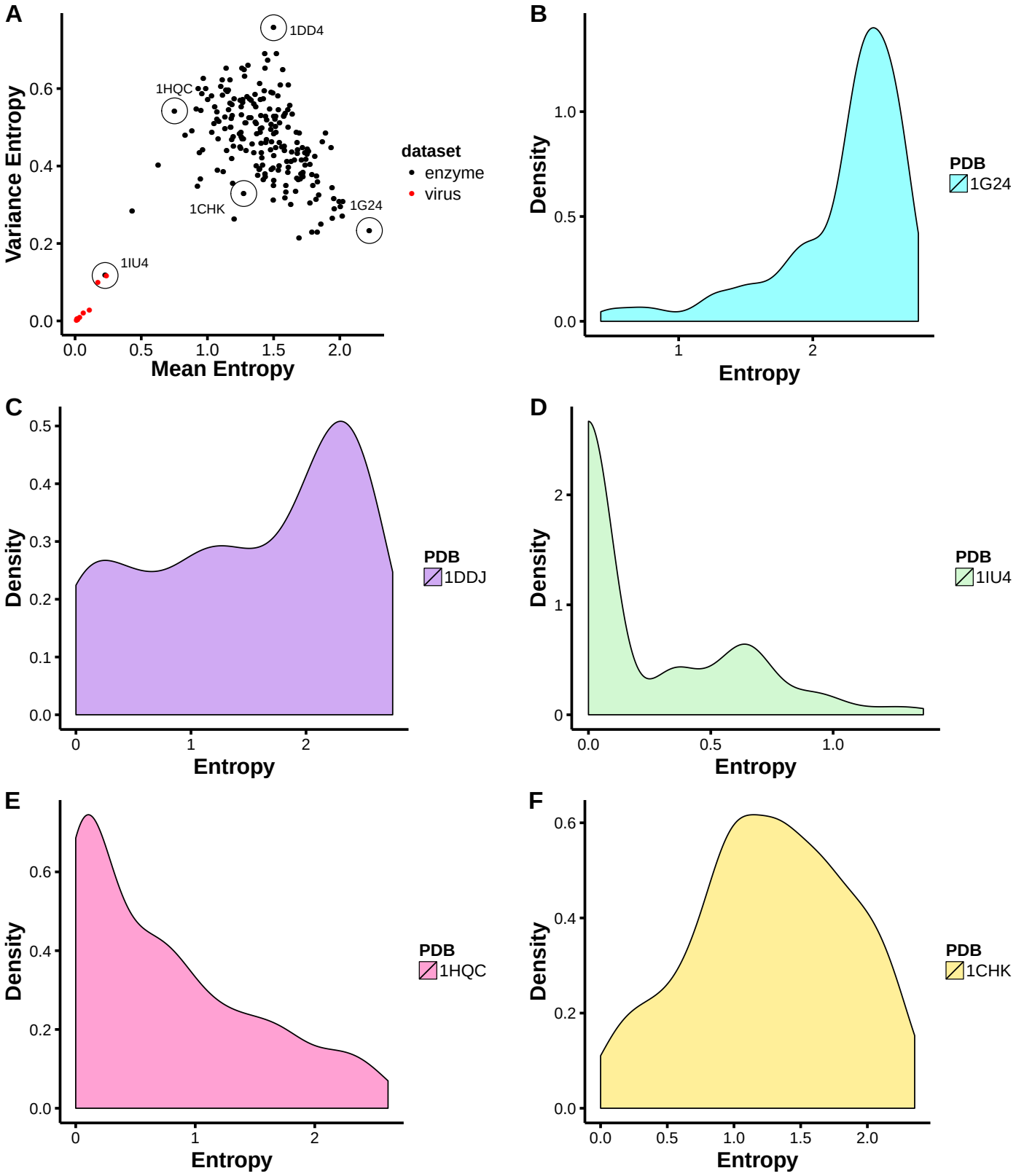

Figure 6: Comparison of the mean entropy and the variance of entropy for individual proteins. A) Variance in entropy at sites compared against overall mean entropy for proteins. Four proteins of interest are highlighted by open circles. The enzyme proteins are colored in black and the virus proteins are colored red B-F) Distributions of site-wise entropy values for the five enzyme proteins highlighted in A. There are a variety of distributions in site entropy for different proteins.

Dataset	Mean ρ Entropy-CN	Mean ρ Entropy-WCN	Mean ρ Entropy-RSA
Enzyme	-0.445	-0.432	0.464
Virus	-0.213	-0.217	0.237

Table 1: Averages of the Spearman Correlation Coefficients between Structural Properties and Entropy. The structural properties analyzed are solvent accessibility as calculated by RSA and packing density as calculated by CN and WCN.

Dataset	Mean ρ Rate4Site-CN	Mean ρ Rate4Site-WCN	Mean ρ Rate4Site-RSA
Enzyme	-0.445	-0.432	0.464
Virus	-0.213	-0.217	0.237

Table 2: Averages of the Spearman Correlation Coefficients between Structural Properties and Evolutionary Rate. Evolutionary rate is calculated with Rate4Site. The structural properties analyzed are solvent accessibility as calculated by RSA and packing density as calculated by CN and WCN.

Model	$\langle H \rangle$	Dataset	Dataset*$\langle H \rangle$
ρ_CN–H = $\langle H \rangle$ + Dataset + Dataset*$\langle H \rangle$	0.079^{**}	0.382^{***}	-0.583
ρ_WCN–H = $\langle H \rangle$ + Dataset + Dataset*$\langle H \rangle$	0.049	0.323^{***}	-0.568
ρ_RSA–H = $\langle H \rangle$ + Dataset + Dataset*$\langle H \rangle$	-0.030	-0.336^{***}	0.903^{*}
ρ_CN–H = $\langle H \rangle$ + Dataset + Dataset*$\langle H \rangle$	0.077^{**}	0.336^{***}	NA
ρ_WCN–H = $\langle H \rangle$ + Dataset + Dataset*$\langle H \rangle$	0.047	0.278^{***}	NA
ρ_RSA–H = $\langle H \rangle$ + Dataset + Dataset*$\langle H \rangle$	-0.027	-0.264^{***}	NA

Table 3: Linear models predicting structural correlations with various quantities. Coefficient x represents the coefficient of the first predictor, coefficient y is the coefficient of the second predictor and coefficient z is the coefficient of the third predictor in each model. *** means the p-value is less than 0.001. ** means the p-value is less than 0.01. * means the p-value is less than 0.05.

Model	Variance H	Dataset	Dataset*Variance H
ρ_CN–H = Variance H + Dataset + Dataset*Variance H	-0.522^{***}	0.013	-0.265
ρ_WCN–H = Variance H + Dataset + Dataset*Variance H	-0.464^{***}	0.023	-0.325
ρ_RSA–H = Variance H + Dataset + Dataset*Variance H	0.415^{***}	-0.076	0.927
ρ_CN–H = Variance H + Dataset + Dataset*Variance H	-0.524^{***}	0.004	NA
ρ_WCN–H = Variance H + Dataset + Dataset*Variance H	-0.466^{***}	0.012	NA
ρ_RSA–H = Variance H + Dataset + Dataset*Variance H	0.422^{***}	-0.044	NA

Table 4: Linear models predicting structural correlations with various quantities. Coefficient x represents the coefficient of the first predictor, coefficient y is the coefficient of the second predictor and coefficient z is the coefficient of the third predictor in each model. *** means the p-value is less than 0.001. ** means the p-value is less than 0.01. * means the p-value is less than 0.05.