June 12, 2015

Dissecting the Role of Long-Range Amino Acid Interactions and Local Packing Density in Protein Sequence Evolution

Amir Shahmoradi[1,2], Claus O. Wilke[2]

[1] *Department of Physics, The University of Texas at Austin, TX 78712, USA; amir@physics.utexas.edu*
[2] *Department of Integrative Biology, The University of Texas at Austin, TX 78712, USA; wilke@austin.utexas.edu*

What are the best structural determinants of protein sequence evolution? A number of site-specific structural characteristics have been proposed over the past decade to answer this question. Most importantly, the role of local packing density has been highlighted and shown recently to be the dominant factor in shaping the observed patterns of site-specific sequence variability in proteins. The most commonly used measures of local packing density such as the Contact Number and the Weighted Contact Number represent by definition, the combined effects of local packing density and long-range amino acid interactions on sequence variability. Here we propose a methodology to segregate and quantify the extent to which long-range interactions determine the general patterns of sequence variability in proteins, independently of local packing density. We use Voronoi partitioning of protein's 3-dimensional structure to obtain a parameter-free measure of the local packing density, defined as the inverse volume of the Voronoi cell corresponding to each site in protein. Using a dataset of 209 monomeric enzymes, we show that the long-range amino acid interactions can explain $\sim 10\%$ of variability observed in protein sequence, whereas $\sim 30\%$ of sequence variability can be attributed to site-specific packing density.

1 Introduction

A variety of site-specific structural characteristics have been proposed over the past decade to predict protein sequence evolution from structural properties. Among the most important and widely discussed are the Relative Solvent Accessibility (RSA) (e.g., Goldman et al., 1998; Bustamante et al., 2000; Conant and Stadler, 2009; Franzosa and Xia, 2009; Ramsey et al., 2011; Scherrer et al., 2012; Meyer and Wilke, 2013; Meyer et al., 2013; Yeh et al., 2014a,b; Shahmoradi et al., 2014; Sikosek and Chan, 2014; Meyer and Wilke, 2015), Contact Number (e.g., Rodionov and Blundell, 1998; Hamelryck, 2005; Liao et al., 2005; Bloom et al., 2006; Huang et al., 2014; Marcos and Echave, 2015; Yeh et al., 2014b,a; Shahmoradi et al., 2014; Meyer and Wilke, 2015), measures of thermodynamic stability changes due to mutations at individual sites in proteins (e.g., Echave et al., 2014), and measures of local flexibility, such as the Debye-Waller factor (hereafter B factor) (e.g., Liao et al., 2005; Shih et al., 2012; Shahmoradi et al., 2014) or flexibility measures based on elastic network models (e.g., Liu and Bahar, 2012) and Molecular Dynamics (MD) simulations (e.g., Shahmoradi et al., 2014).

Although these structural characteristics have been individually extensively studied and explored with regards to their association with sequence evolution, it is yet unknown whether these seemingly independent quantities are merely different manifestations of a more fundamental underlying characteristics of individual sites in proteins or each has a unique independent influence on the sequence variability patterns in proteins. It is perceivable that quantities such as B factor and RSA, all serve as different proxy measures of local packing density of individual sites in proteins. Franzosa and Xia (2009) used a variety of local packing density measures to show that RSA is the key determinant of sequence evolution with packing density having only peripheral influence. Recently however, Huang et al. (2014) argued through an extensive mathematical formulation within the framework of Elastic Network Models, for the local packing density as the dominant factor in shaping the observed sequence variability patterns among enzymatic proteins in contrast to RSA and local flexibility measures.

A crude measure of site-specific packing density for each site in protein is often obtained through a quantity widely known as Contact Number (e.g., Liao et al., 2005). Alternative packing density measures have also been proposed leading to better predictions of sequence evolution and site-specific flexibility, a prime example of which is the Weighted Contact Number (WCN) proposed by Lin et al. (2008). These proxy measures of packing density however, involve adjustable free parameters in their definitions. Moreover, the packing density as defined by these quantities is more representative of the local environment around the site of interest, and does necessarily correspond to the site-specific packing density.

Motivated by the existing gaps in the current understanding of the role of site-specific packing density and long-range amino acid interactions in protein sequence evolution, here we derive a new set of structural characteristics that, unlike CN and WCN, do not involve free adjustable parameters in their definitions. In particular, we propose a methodology to derive a minimally-biased measure of local packing density based upon which the role of long-range amino-acid interactions on sequence variability can be segregated and isolated from the effects of site-specific packing density. This is done by employing tessellation methods from the field of computational geometry to calculate several new characteristics of sites in proteins, which can serve as proxy measures of site-specific packing density and flexibility. In this regard, the two commonly-used definitions of local packing density (CN & WCN) may not be considered *site-specific*. An important question therefore remains unanswered as to what extent the site-specific packing density and long-range interactions influence sequence variability, independently of each other.

2 Methods

Local Packing Density Definition and Measurement

The local packing density is commonly measured by a quantity known as Contact Number (CN). In its simplest mathematical form, the Contact Number for a given site i in a protein of amino acid sequence length of N is defined as the number of amino acids within a fixed radius r_0 of neighborhood around the site (e.g., Franzosa and Xia, 2009),

$$\mathrm{CN}_i = \sum_{j\neq i}^{N} H\big(r_0 - r_{ij}\big), \tag{1}$$

in which r_{ij} represents the distance between sites i & j and,

$$H\big(r_0 - r_{ij}\big) = \int_{-\infty}^{r_0 - r_{ij}} \delta\big(x\big)\,\mathrm{x}, \tag{2}$$

is the Heaviside step function, with $\delta(x)$ standing for the Dirac delta function. Individual sites are generally represented by the coordinates of C_α backbone atoms for the calculation of CN. A major problem with the traditional definition of contact number however, is the existence of the arbitrary parameter r_0 in the definition of CN. There is no consensus on the optimal value of this cutoff distance, although it is typically chosen in the range $5\mathring{A}$ to $18\mathring{A}$ (e.g., Lin et al., 2008; Franzosa and Xia, 2009; Weng and Wang, 2014).

In an attempt to provide a more general definition of CN, some studies (e.g., Lin et al., 2008) have already suggested an alternative definition known as the Weighted Contact Number (WCN): For a given site i in a protein of length N, WCN_i is defined as the sum of the inverse-squared of distances between the amino acid of interest and all other sites in protein,

$$\mathrm{WCN}_i = \sum_{j\neq i}^{N} r_{ij}^{\alpha=-2}. \tag{3}$$

Although WCN is in general a better predictor of site-specific sequence variability, the proposed definition of WCN still involves an adjustable parameter: the exponent of the power-law kernel (α). The value of the exponent that results in the best predictions appears to be in the range $-3 \lesssim \alpha \lesssim -2$ and is typically fixed to $\alpha = -2$ as shown in Eqn 3 (e.g., Yang et al., 2009). A similar exponent value was also used by Huang et al. (2014) to argue for the *local* packing density of sites as the dominant factor in shaping protein sequence variability. The specific value $\alpha = -2$ however, implies that the long-range amino acid interactions also play a non-negligible role in sequence evolution, independently of the local packing density. Similarly, the best performing cutoff values in the definition of Contact Number are in the range of $10\mathring{A} \lesssim \alpha \lesssim 15\mathring{A}$, also indicative of the significant influence of the long-range amino acid interactions on sequence variability (e.g., Franzosa and Xia, 2009; Shahmoradi et al., 2014), aside from site-specific packing density.

Protein Dataset and Site-Specific Structure/Sequence Variability Measures

The entire analyses and results presented in this work are based on a dataset of 209 monomeric enzymes (e.g., Yeh et al., 2014b; Echave et al., 2014) 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 amino acids, 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 (Yeh et al., 2014b) 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., 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 sequence variability for each individual protein in dataset. For each structure, the respective sequence alignment and phylogenetic tree were used to infer site-specific substitution rates with Rate4Site, using the empirical Bayesian method and the amino-acid Jukes-Cantor mutational model (Mayrose et al., 2004), hereafter abbreviated as *r4sJC*.

In addition to site-specific evolutionary rates, we also calculate the Shannon entropy (H_i) – the sequence entropy (Shenkin et al., 1991) – 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:

$$H_i = -\sum_j P_{ij} \ln P_{ij} \tag{4}$$

where P_{ij} is the relative frequency of amino acid j at position i in the alignment. We use DSSP software (Kabsch and Sander, 1983) for the calculation of the Accessible Surface Area (ASA) for each site normalized 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 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

Packing Density Definitions and Long-Range Amino Acid Interactions

In order to determine the extent to which long-range amino acid interactions influence the general patterns of sequence variability in proteins, first we investigate the behavior of Contact Number and the Weighted Contact Number in predicting site-specific evolutionary rates (r4sJC) and sequence entropy for a wide range of the free parameters of the two packing density measures (i.e., $r_0 \in [0\AA, 50\AA]$ & $\alpha \in [-30, 30]$ as in Eqns. 1 & 3).

The results for the dataset of 209 monomeric enzymes are plotted in Figure 1 and the values of the free parameters of CN and WCN that yield the strongest correlations are tabulated in Table 2. Evidently, the best-performing free parameters of both models indicate a non-negligible contribution of the long-range amino acid interactions, beyond the immediate neighborhood of the site, to the strengths of the observed correlations. The significance and impact of these non-local interactions on sequence evolution cannot be deciphered solely from the two common definitions of packing density: CN & WCN. Therefore, we present, in the following section, a methodology to segregate the role of local packing density from long-range interactions and quantify their individual contributions to site-specific sequence variability measures.

Voronoi Partitioning of Protein's Structure

There is already an extensive body of literature on the applications of different methods of structural partitioning in the studies of protein structure and its relation to sequence Richards (1974); Gerstein et al. (1994). The Voronoi tessellation and its dual graph, the Delaunay triangulation, have particularly attracted much attention in the studies of protein internal structure and the development of empirical potentials Zomorodian et al. (2006); Zhou and Yan (2014); Xia et al. (2014). For a given a set of centroid points (seeds) in 3-dimensional Euclidean space, the simplest and most familiar case of Voronoi tessellation divides the Euclidean space into regions, called *cells*, such that the cell corresponding to each centroid point consists of every region in space whose distance is less than or equal to its distance to any other centroid points.

In the context of protein studies, the atomic coordinates of C_α backbone atoms have been widely used as the set of Voronoi seeds to partition the 3D structure of proteins. An example of Voronoi tessellation of protein structure in two dimensions (PDB ID: *1LBA*) is shown in Figure 2. The properties of individual cells resulting from tessellation can be then used to obtain a wide range of information on protein structure, energy landscape or protein–protein interactions, also about sequence evolution as will be shown in the following section.

Here in this work, we apply the simplest and most widely used definition of Voronoi tessellation on the dataset of 209 monomeric enzymes. We use VORO++ software Rycroft (2009) to calculate the relevant Voronoi cell properties of all sites in all proteins in the dataset. Among the most important properties are the length of the cell edges, surface area and volume, number of faces of each cell, and the cell eccentricity defined as the distance between the cell's seed and the geometrical center of the cell. In addition, the cell *sphericity* can be calculated as a measure of the cell's *compactness* defined as,

$$\Psi = \frac{\pi^{\frac{1}{3}}(6V)^{\frac{2}{3}}}{A}. \tag{5}$$

in which V & A stand for the volume & area of the cell, respectively. For a perfectly spherical cell, $\Psi = 1$, while it becomes zero for a 2-dimensional object that has no volume but only surface area.

Voronoi Cell Volume as a Proxy Measure of Local Packing Density and Flexibility

In order to assess the prediction power of the site-specific characteristics derived from Voronoi tessellation, first the geometric centers of all side-chains for each of the proteins in dataset were calculated and used as the seeds of Voronoi polyhedra. We show in Appendix A that the choice of the geometric centers of the side chains as Voronoi seeds – in contrast to other sets of atomic coordinates representative of individual sites in protein – results in strongest correlations of Voronoi cell properties with site-specific sequence variability of proteins.

Figure 3 depicts the distributions of the Spearman's correlation coefficients of five most important Voronoi cell characteristics with site-specific evolutionary rates (ER). It is notable that all cell characteristics in the plot correlate positively with ER, except the cell sphericity which is always negatively correlated with ER and with other Voronoi cell properties. In general, it is observed that the cell volume and surface area have the best predictive power compared to other cell characteristics, followed by the cell eccentricity, total edge length, and the cell's sphericitiy.

The Voronoi cell characteristics are also strongly associated with each other. Although the cell volume and area are almost identically the best correlating variables with ER, the cell volume does not exhibit any significant independent correlation with ER once the cell area is controlled for. The median strength of the partial correlation of volume with ER, while controlling for area is centered at ~ 0.0 (Figure 4). Conversely, the cell sphericity and eccentricity both exhibit median partial correlations of ~ -0.1 & ~ 0.07 with ER respectively, when the contribution from the Voronoi cell area is controlled. In conclusion, the cell area, volume, and edge length appear to represent almost the same property of the Voronoi cell. Other Voronoi cell characteristics, such as the number of vertices, faces and edges of the cell also tend to correlate weakly with sequence evolutionary rates. However, these cell characteristics are discrete (integer) quantities and in general have limited ranges.

Not shown here for brevity, almost identical results to the above are obtained were sequence entropy used in place of evolutionary rates, as defined by Eqn. 4. The use of sequence entropy however, generally results in weaker correlation strengths due to the discreteness and limited range that is inherent in the definition of sequence entropy. One potential caveat with the Voronoi tessellation of finite structures, such as proteins, is the presence of *edge effects* in the properties of the cells that remain open, typically on the surface of the structure. We show in appendix B that such effects are generally negligible in our results presented in this section.

Effect of Long-Range Amino Acid Interactions on Sequence Evolution

Figure 5 compares the prediction power of each of the site-specific structural quantities about sequence evolutionary rates (*r4sJC*) for the dataset of 209 monomeric enzymes. Not shown here for brevity, similar results are also obtained for sequence entropy as the measure of sequence variability. For comparison, the results for WCN calculated using the C_α atomic coordinates are also illustrated in the plot, in addition to WCN calculated from the coordinates of the geometric centers of side chains. Notably, the quantity WCN outperforms all other structural quantities in explaining site-specific sequence variability. In particular, a pairwise t-test between the correlation strengths of r4sJC-WCN and the correlation of r4sJC with the inverse of Voronoi cell volume yields a p-value of $< 10^{-16}$. The better performance of WCN compared to local packing density as measured by the inverse of Voronoi cell volume may not be surprising, knowing that WCN by its definition in Eqn. 3 also takes into account the potential long-range interactions among amino acids in different regions of protein.

In order to segregate the combined effects of long-range interactions from local packing density, the inverse of the Voronoi cell volume can be used as a maximally local measure of packing density. By controlling for the Voronoi cell volume, one can then quantify the residual influence of WCN – that is, the effects long-range amino acid interactions beyond immediate neighbors – on sequence variability. Figure 6 illustrates the partial correlation strengths of the same structural quantities as in Figure 5, while controlling for the Voronoi cell volume. It should be noted that the strengths of the partial correlations are insensitive to whether the cell volume or its inverse is controlled for, since the Spearman ρ is a non-parametric rank correlation coefficient.

The resulting distribution of the Spearman's partial correlation coefficients of evolutionary rates with wcn (calculated using side-chain coordinates, wcnSC) has an absolute median value of 0.32 with a 50% quartile range of $[0.23, 0.40]$ about the median of the distribution. Therefore, the long-range amino acid interactions appear to explain approximately 10% of the site-specific sequence evolutionary rates. The local packing density as measured by the inverse of Voronoi cell volume is alone capable of explaining a median 35% of sequence evolutionary rates, corresponding to a median Spearman's correlation strength of 0.59 with 50% quartile range of $[0.52, 0.64]$. By contrast, the partial correlation distribution of RSA while controlling for the Voronoi cell volume for the same dataset, indicates a negligible median contribution of ~ 0.006 in explaining the observed variability in evolutionary rates of individual proteins.

4 Discussion

Throughout the previous sections, we carried out a comprehensive analysis and comparison of the most widely studied structural determinants of sequence variability, using a dataset of 209 monomeric enzymes. Examples of important sequence–structure relations include the correlations of measures of sequence variability – such as evolutionary rates (e.g., *r4sJC*) and sequence entropy – with measures of residue Contact Number (CN & WCN) and Relative Solvent Accessibility (RSA). In addition, we derived a new set of site-specific characteristics from the Voronoi partitioning of protein's 3D structure, some of which are capable of explaining sequence variability equally well or better than several structural quantities that were previously considered in the literature – including B factor, RSA, and the traditional definitions of CN and WCN using C_α atomic coordinates (Figure 5).

The commonly used measures of local packing density, most importantly CN and WCN, involve free adjustable parameters in their definitions that can be fine-tuned for each individual protein to obtain the strongest correlations between CN/WCN and sequence variability. Although the best parameter values for CN & WCN vary from one protein structure to another, the median of the values over the entire dataset clearly indicate a non-negligible selection pressure on protein sequence solely due to long-range amino acid interactions, independently of the local packing density. Unlike CN and WCN, the site-specific packing density as defined by the inverse of Voronoi cell volume in this work, represents only the contributions from the nearest neighbors of each site in protein, that is, the first coordination shell.

Although the site-specific packing density as defined by the inverse of the Voronoi cell volume excludes the effects of long-range interactions beyond the first coordinations shell, it remains as best predictor of sequence variability in our dataset among all other structural determinants, most importantly RSA. Contrary to the findings of this work however, some other studies (e.g., Franzosa and Xia, 2009; Scherrer et al., 2012; Shahmoradi et al., 2014) have argued for RSA as the main determinant of sequence evolution with local packing density having a peripheral role. Our conclusion is that RSA and the local packing density – once corrected for long-range interactions – in principle represent the same characteristics of the local environment in protein, that is, both quantities are proxy measures of the number of neighboring amino acids in the first coordination shell. This argument is also further supported by running a pairwise t-test between the correlation strengths of RSA with evolutionary rates and the correlation strengths of the inverse Voronoi cell volume with evolutionary rates for the entire protein dataset, with a resulting p-value ~ 0.29. An important question however remains unanswered in this work and merits further study, as to what extent the findings of the presented analysis – based on a dataset of 209 monomeric enzymes – can be generalized to other types of proteins in nature.

Appendix A Average Side-Chain Coordinates as the Best Representation of Protein 3D Structure

Depending on the choice of the Cartesian coordinates used, there exist degeneracy in the definition of some site-specific structural variables. For example, the quantity WCN is generally calculated from the coordinates of C_α atoms in the 3-dimensional structure of protein. The choice of C_α coordinates is however mainly driven by convenience in WCN calculation and there is no reason to believe this set of atomic coordinates is the best representative of individual sites in proteins. Indeed, some earlier works have already suggested the use of center-of-mass of side chain coordinates to represent the 3D structure of protein Soyer et al. (2000). More recently, Marcos & Echave (2015) Marcos and Echave (2015) have also shown that WCN calculated from side-chain center-of-mass coordinates generally result in significantly better correlations of WCN with sequence variability measures.

Despite the highly popular choice of C_α atomic B factor as a proxy measure of residue flexibility Halle (2002), same definition degeneracy also exists on choice of atomic B factors that are used to represent site-specific flexibility. In addition to WCN and B factor, there is also ambiguity as to which set of residue atomic coordinates best represent individual sites in proteins for the generation of Voronoi polyhedra.

Here in this work, all possible choices of the representative set of atomic coordinates were considered in order to identify which set of atomic coordinates best represents individual sites for the calculation of WCN, B factor, and Voronoi cells. Depending on the set of atomic coordinates that represent the protein structure, there are at least 7 different measures of each individual site-specific structural properties, such as the Weighted Contact Number, B factor and Voronoi cell properties. 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, representative coordinates for each site in protein can be also calculated by averaging over the coordinates of all heavy atoms in the side chains. Also calculated was a representative coordinate for each site by averaging over all heavy atom coordinates in the side chain and the backbone of the amino acid together. In rare cases where the side chain C_β atom had not been resolved in the PDB file or the amino acid lacked C_β (e.g., Glycine), the C_β coordinate for the specific amino acid were replaced with the coordinate of the corresponding C_α atom in the same amino acid. The resulting Spearman's correlation strengths of site-specific evolutionary rates, sequence entropy, $\Delta\Delta G$ rate as defined by Echave et al. (2014), Relative Solvent Accessibility (RSA), amino acid hydrophobicity, and the average backbone Hydrogen bond energy with different measures of WCN and Voronoi cell area are depicted in the plots of Figures 7 and 8 respectively, for different sets of atomic coordinates used in the calculations. The hydrophobicity scales of amino acids residing in individual sites in proteins were taken from Hessa et al. (2005). Other hydropobicity scales were also considered Wimley and White (1996); Kyte and Doolittle (1982), however similar results are obtained for all. The Hydrogen bonds were identified and the corresponding energies were calculated according to the prescription of Kabsch and Sander (1983).

For the measure of local packing density in proteins (the Weighted Contact Number) we find that among all possible sets of coordinates, the average over coordinates of all heavy atoms of each individual side-chain results in WCN values that show the strongest correlation strength with other structural and sequence properties, such as RSA, Voronoi cell properties, sequence entropy, and evolutionary rates. Specifically, WCN from average side chain coordinates outperforms WCN based on C_α coordinates in predicting RSA, $\Delta\Delta G$ rate, sequence entropy and evolutionary rates with median Spearman correlation differences of 0.09, 0.10, 0.07 & 0.08, respectively (Figure 7).

Similar to WCN, the Voronoi cell properties, most importantly the cell surface area, volume, edge length, eccentricity and the cell sphericity also correlate best with other structure and sequence properties, only if the geometric average of side chain coordinates are used as the seeds of Voronoi cells. For brevity, only the results for the cell area are given here. Specifically, the cell area from average side-chain coordinates outperforms cell area calculated based on C_α coordinates in predicting RSA, $\Delta\Delta G$ rate, sequence entropy and evolutionary rates with median Spearman correlation differences of 0.04, 0.06, 0.04 & 0.04, respectively (Figure 8).

It is notable that the standard deviations of the difference distributions for both quantities: WCN, and Voronoi cell area, are an order of magnitude smaller than the observed differences, implying that the correlation coefficients for all proteins in dataset uniformly translate to higher values by moving from C_α atomic coordinates to the geometric centers of the side chains, regardless of the strength of the correlation coefficients.

Appendix B Edge-Effects in Voronoi Partitioning of Protein Structures

One potential caveat with Voronoi tessellation of finite structures in Euclidean space is the *edge effects*. Sites that are close to the surface of protein are often associated with Voronoi cells that are bounded by the cubic box containing the protein (Figure 2). Here to ensure that these edge effects do not influence the observed sequence-structure correlations, the open cells – i.e., cells that are partially bounded and closed by the cubic box containing the protein – are identified in all proteins by examining the variations in individual cell volumes upon changing the size of the cubic box containing the protein to a given extreme value. The open cells in individual proteins are then ranked by the fraction of volume changes observed upon changing the box size and then normalized to the the largest volume observed among closed cells.

It should be noted that the specific extreme value chosen for the box sizes of the proteins or the rank ordering of the open cells does not have any influence on the resulting correlation strengths, since the Spearman's ρ by its definition is a rank correlation coefficient.

Our conclusion is that the *edge effects* due to Voronoi tessellation appear to have $\lesssim 0.01$ influence on the observed sequence-structure correlations in the dataset of 209 proteins considered in this work. Similar conclusions are reached if the open cells were alternatively ranked by other criteria such as the fractional changes in cell area (vs. cell volume) upon changing the box size. The Voronoi cell characteristics, in particular cell volume and cell area can be safely used in predicting sequence variability without recourse to corrections for the edge effects. An exception however is cell sphericity as defined in Eqn. 5, which turns out to behave differently for open and closed cells.

ACKNOWLEDGEMENTS

We thank Austin G. Meyer, Stephanie Spielman and Eleisha Jackson at UT Austin for helpful discussions and comments.

References

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research*. 38:W529–W533.

Bloom JD, Drummond DA, Arnold FH, Wilke CO. 2006. Structural Determinants of the Rate of Protein Evolution in Yeast. *Molecular Biology and Evolution*. 23:1751–1761.

Bustamante CD, Townsend JP, Hartl DL. 2000. Solvent Accessibility and Purifying Selection Within Proteins of Escherichia coli and Salmonella enterica. *Molecular Biology and Evolution*. 17:301–308.

Conant GC, Stadler PF. 2009. Solvent Exposure Imparts Similar Selective Pressures across a Range of Yeast Proteins. *Molecular Biology and Evolution*. 26:1155–1161.

Echave J, Jackson EL, Wilke CO. 2014. Relationship between protein thermodynamic constraints and variation of evolutionary rates among sites. *bioRxiv*. p. 009423.

Franzosa EA, Xia Y. 2009. Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. *Molecular Biology and Evolution*. 26:2387–2395.

Gerstein M, Sonnhammer ELL, Chothia C. 1994. Volume changes in protein evolution. *Journal of Molecular Biology*. 236:1067–1078.

Goldenberg O, Erez E, Nimrod G, Ben-Tal N. 2009. The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Research*. 37:D323–D327.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution. *Genetics*. 149:445–458.

Halle B. 2002. Flexibility and packing in proteins. *Proceedings of the National Academy of Sciences*. 99:1274–1279.

Hamelryck T. 2005. An amino acid has two sides: A new 2d measure provides a different view of solvent exposure. *Proteins: Structure, Function, and Bioinformatics*. 59:38–48.

Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH, von Heijne G. 2005. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature.* 433:377–381.

Huang TT, Marcos MLdV, Hwang JK, Echave J. 2014. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. *BMC Evolutionary Biology.* 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 Research.* 33:511–518.

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology.* 157:105–132.

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

Lin CP, Huang SW, Lai YL, Yen SC, Shih CH, Lu CH, Huang CC, Hwang JK. 2008. Deriving protein dynamical properties from weighted protein contact number. *Proteins: Structure, Function, and Bioinformatics.* 72:929–935.

Liu Y, Bahar I. 2012. Sequence Evolution Correlates with Structural Dynamics. *Molecular Biology and Evolution.* 29:2253–2263.

Marcos ML, Echave J. 2015. Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. *PeerJ.* 3:e911.

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. *Molecular Biology and Evolution.* 21:1781–1791.

Meyer AG, Dawson ET, Wilke CO. 2013. Cross-species comparison of site-specific evolutionary-rate variation in influenza haemagglutinin. *Philosophical Transactions of the Royal Society of London B: Biological Sciences.* 368:20120334.

Meyer AG, Wilke CO. 2013. Integrating Sequence Variation and Protein Structure to Identify Sites under Selection. *Molecular Biology and Evolution.* 30:36–44.

Meyer AG, Wilke CO. 2015. Geometric constraints dominate the antigenic evolution of influenza H3n2 hemagglutinin. *bioRxiv.* p. 014183.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology.* 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 Research.* 32:D129–D133.

Ramsey DC, Scherrer MP, Zhou T, Wilke CO. 2011. The Relationship Between Relative Solvent Accessibility and Evolutionary Rate in Protein Evolution. *Genetics.* 188:479–488.

Richards FM. 1974. The interpretation of protein structures: Total volume, group volume distributions and packing density. *Journal of Molecular Biology.* 82:1–14.

Rodionov MA, Blundell TL. 1998. Sequence and structure conservation in a protein core. *Proteins: Structure, Function, and Bioinformatics.* 33:358–366.

Rycroft CH. 2009. VORO++: A three-dimensional Voronoi cell library in C++. *Chaos: An Interdisciplinary Journal of Nonlinear Science*. 19:041111.

Scherrer MP, Meyer AG, Wilke CO. 2012. Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evolutionary Biology*. 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. *Journal of Molecular Evolution*. 79:130–142.

Shenkin PS, Erman B, Mastrandrea LD. 1991. Information-theoretical entropy as a measure of sequence variability. *Proteins: Structure, Function, and Bioinformatics*. 11:297–313.

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

Sikosek T, Chan HS. 2014. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of The Royal Society Interface*. 11:20140419.

Soyer A, Chomilier J, Mornon JP, Jullien R, Sadoc JF. 2000. Vorono\"\i Tessellation Reveals the Condensed Matter Character of Folded Proteins. *Physical Review Letters*. 85:3532–3535.

Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. 2013. Maximum Allowed Solvent Accessibilites 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*. pp. xiii + 863 pp.

Weng J, Wang W. 2014. Molecular Dynamics Simulation of Membrane Proteins. In: Han Kl, Zhang X, Yang Mj, editors, Protein Conformational Dynamics, Springer International Publishing, number 805 in Advances in Experimental Medicine and Biology, pp. 305–329.

Wimley WC, White SH. 1996. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature Structural Biology*. 3:842–848.

Xia F, Tong D, Yang L, Wang D, Hoi SCH, Koehl P, Lu L. 2014. Identifying essential pairwise interactions in elastic network model using the alpha shape theory. *Journal of Computational Chemistry*. 35:1111–1121.

Yang L, Song G, Jernigan RL. 2009. Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences*. 106:12347–12352.

Yeh SW, Huang TT, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J. 2014a. Local Packing Density Is the Main Structural Determinant of the Rate of Protein Sequence Evolution at Site Level. *BioMed Research International*. 2014:e572409.

Yeh SW, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J. 2014b. Site-Specific Structural Constraints on Protein Sequence Evolutionary Divergence: Local Packing Density versus Solvent Exposure. *Molecular Biology and Evolution*. 31:135–139.

Zhou W, Yan H. 2014. Alpha shape and Delaunay triangulation in studies of protein-related interactions. *Briefings in Bioinformatics*. 15:54–64.

Zomorodian A, Guibas L, Koehl P. 2006. Geometric filtering of pairwise atomic interactions applied to the design of efficient statistical potentials. *Computer Aided Geometric Design*. 23:531–544.

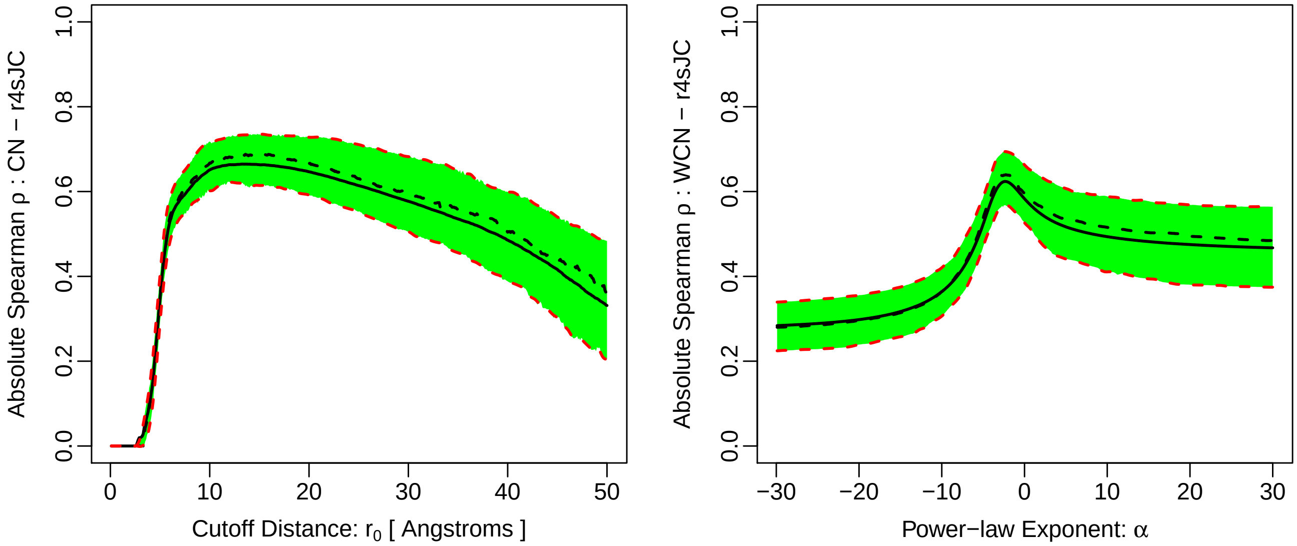

Figure 1: Average absolute Spearman's correlation strengths of Contact Number (CN, as defined by Eqn. 1 using Heaviside kernel in Eqn. 2) and the Weighted Contact Number (WCN, as defined by Eqn. 3 using a power-law kernel) with site-specific evolutionary rates, for different values of the free parameters of the two kernels (r_0 & α respectively). On each plot, the solid black lines represent the mean correlation strength in the entire dataset of 209 proteins at each value of the free parameter, and the dashed black lines indicate the median of the distribution. The green-shaded region together with the red-dashed lines represent the 25% & 75% quartiles of the correlation strength distribution. Note that for the case of WCN with $\alpha > 0$ the sign of the correlation strength ρ is the opposite of the sign of ρ with $\alpha < 0$. In addition ρ is undefined at $\alpha = 0$ and not shown in this plot. The parameter values at which the Spearman's correlation coefficient reaches the maximum over the entire dataset are given in Table 2.

Table 1: Median best free parameter values of the Contact Number (CN) and the Weighted Contact Number with power-law kernel (WCN) that result in the strongest median Spearman's correlation (ρ) of CN & WCN with site-specific sequence variability measures (evolutionary rates (r4sJC) and sequence entropy) in the entire dataset of 209 proteins. Given in parentheses are the corresponding median Spearman correlation coefficients at the best parameter values. The subscripts and superscripts to each value represent the 25% percentile range below and above the median value of the distribution.

Correlation with	$r_0[\mathring{A}]$ (CN)	α (WCN)
r4sJC	$14.3^{+5.3}_{-4.0}$ ($\rho \sim 0.64^{+0.06}_{-0.06}$)	$-2.3^{+0.8}_{-0.4}$ ($\rho \sim 0.65^{+0.05}_{-0.07}$)
Seq. Entropy	$12.4^{+5.5}_{-2.6}$ ($\rho \sim 0.55^{+0.06}_{-0.06}$)	$-2.2^{+0.8}_{-0.4}$ ($\rho \sim 0.55^{+0.07}_{-0.06}$)

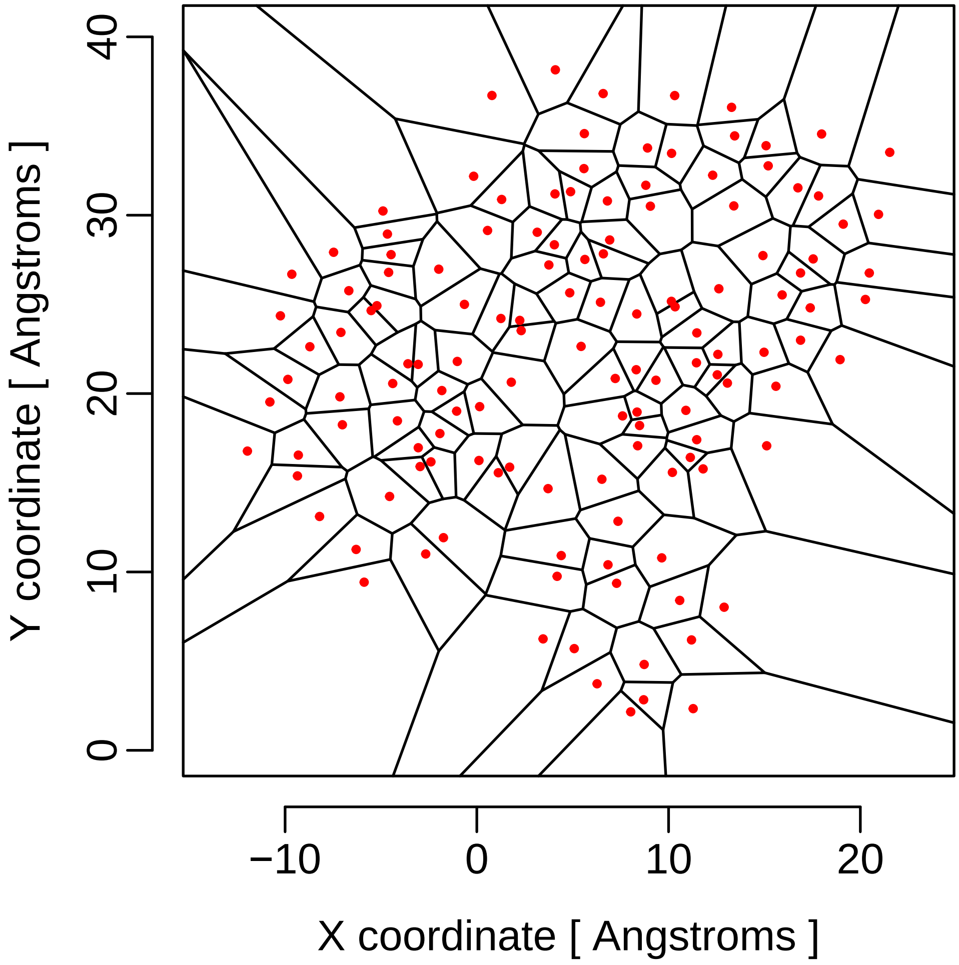

Figure 2: An Example 2-dimensional Voronoi diagram for bacteriophage T7 lysozyme (Protein Data Bank ID '1LBA'). The red dots represent the backbone C_α atoms projected on the X–Y plane, used as cell seeds in Voronoi tessellation.

Table 2: The percentage of variance of site-specific evolutionary rates (r4sJC) explained by site-specific packing density (represented by the inverse volume of Voronoi cells) and long-range amino acid interactions (as described in Sec. 3). For each structural quantity, the first, second (median) and the third quartiles of the percentage distribution of the explained variance are reported on each row, for the entire dataset of 209 proteins.

ctural quantity	variance explained (25% quartile)	variance explained (median)	variance explained (75% qua
specific packing density	27%	34%	41%
-range amino-acid interactions	5%	10%	16%

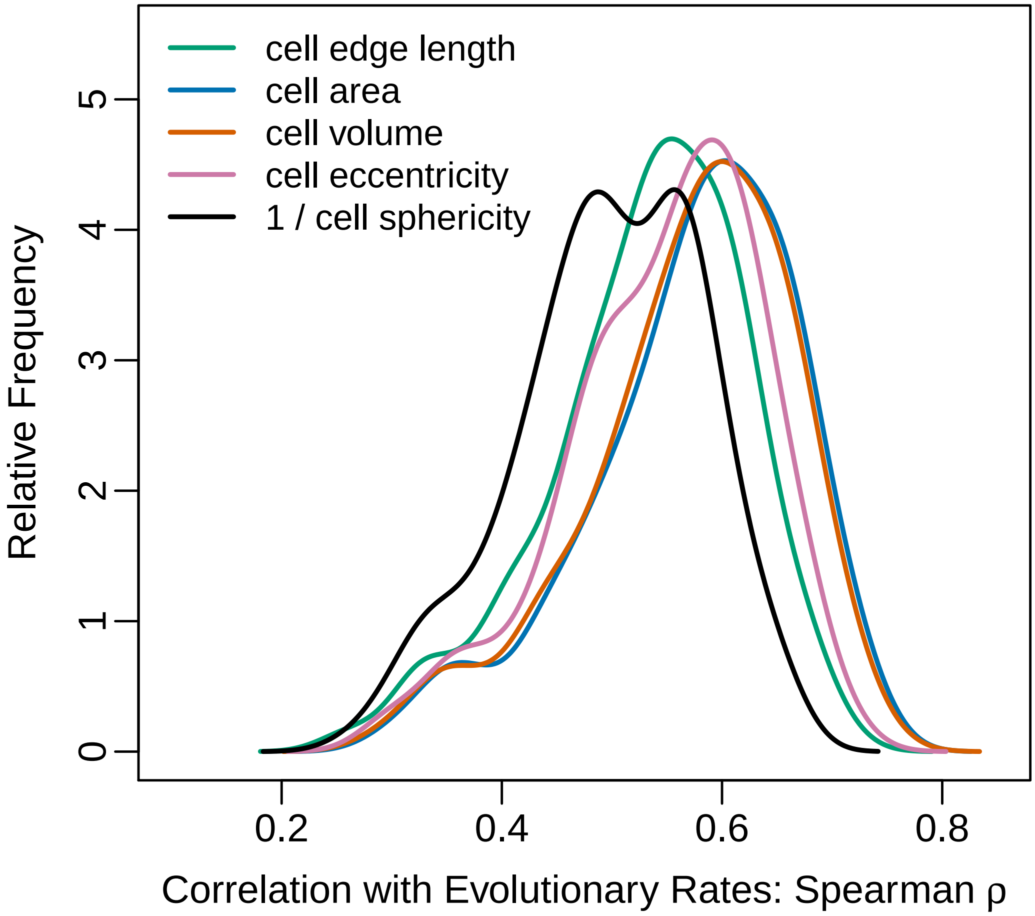

Figure 3: A comparison of the prediction power of different Voronoi cell characteristics about site-specific evolutionary rates (ER). Note that all cell characteristic correlate positively with ER, except sphericity which strongly negatively correlates with ER.

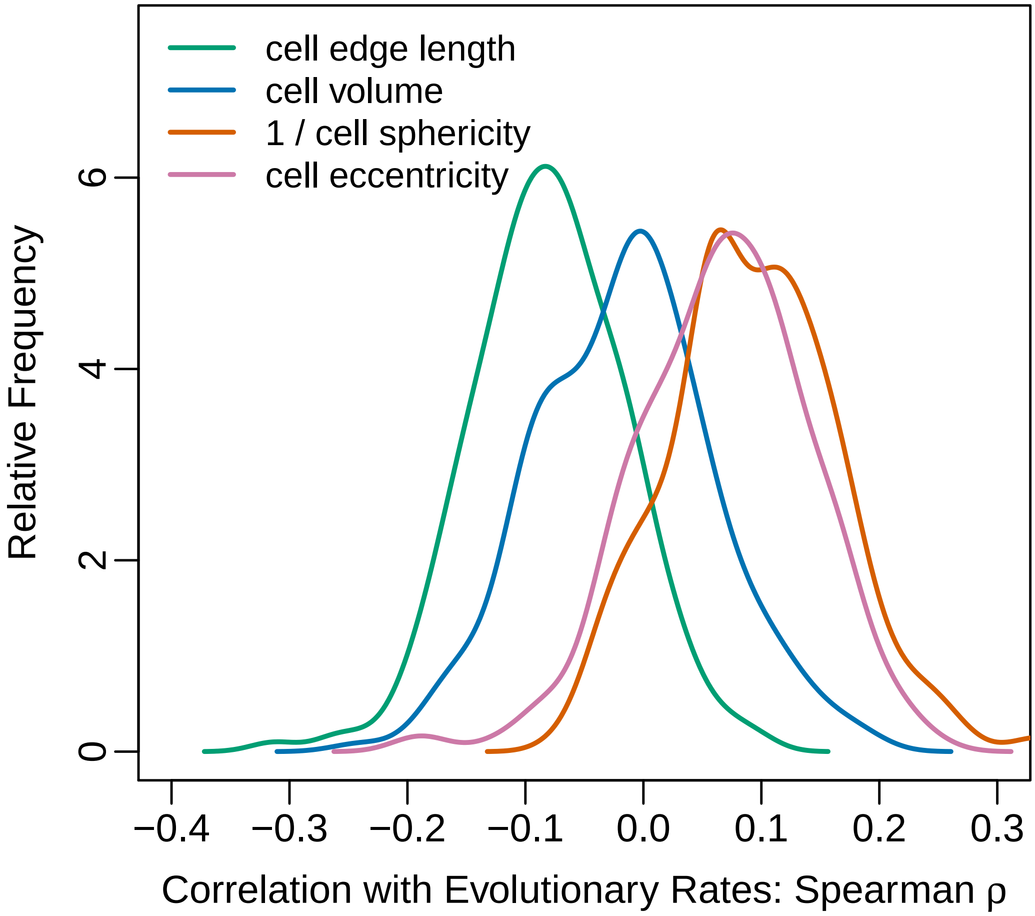

Figure 4: The partial correlation strengths of the same Voronoi cell characteristics with sequence evolutionary rates while controlling for the cell area.

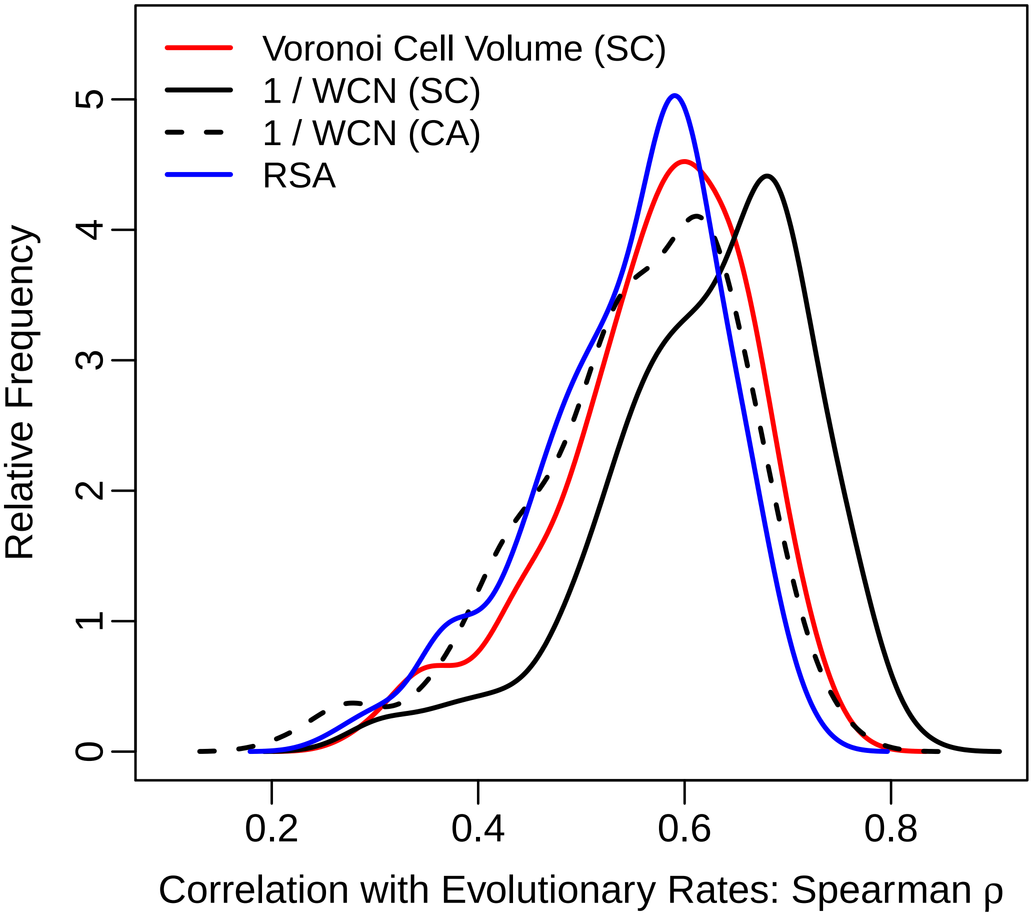

Figure 5: A comparison of the prediction power of five structural variables about site-specific evolutionary rates (ER). All structural quantities correlate positively with ER, with the exception of Weighted Contact Number (WCN) which correlates negatively. For better illustration however, the Spearman's correlation coefficient (ρ) of the inverse of WCN with ER are shown in the Figure. Note that the Spearman's ρ is a rank correlation coefficient, meaning that the use of inverse WCN only changes the sign and not the magnitude of ρ. The abbreviation *SC* refers to the use of average Side-Chain coordinates wherever used, and *CA* refers to the use of backbone C_α atomic coordinates for representation of individual sites in proteins.

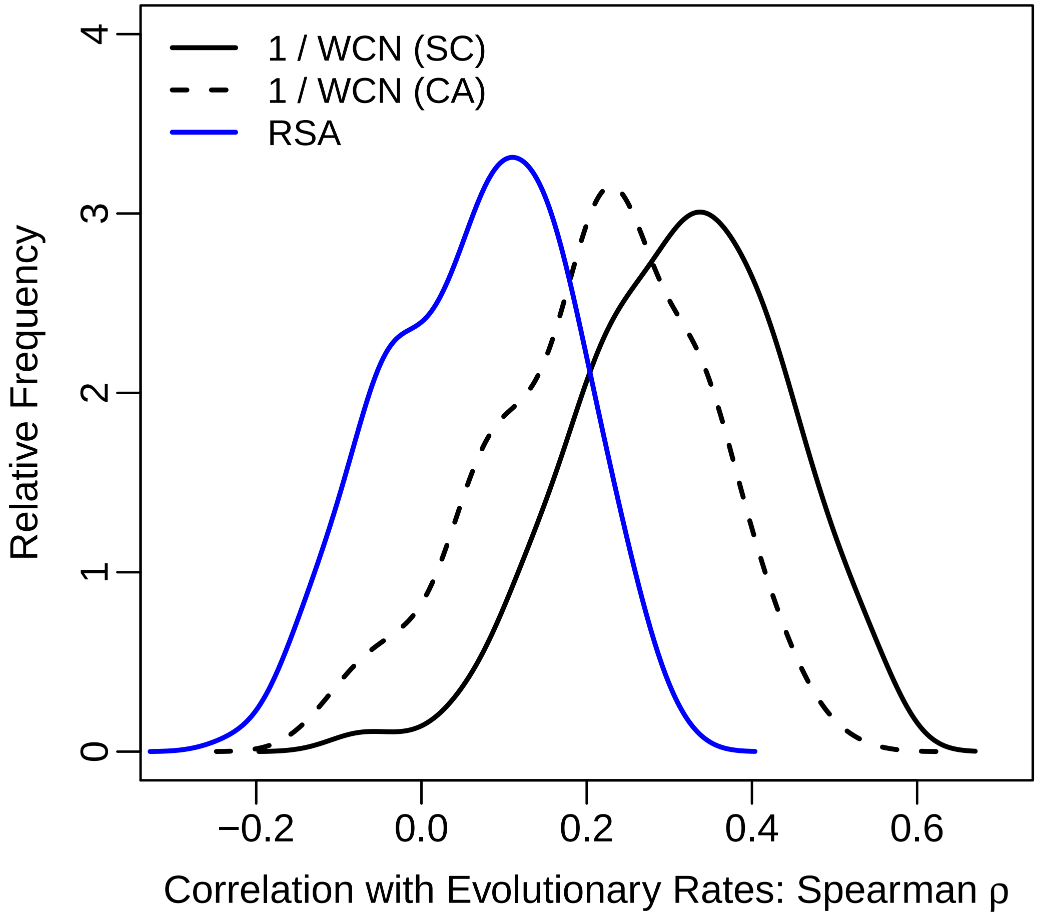

Figure 6: A comparison of the prediction power of four structural variables (as in Figure 5) about site-specific evolutionary rates (ER), while controlling for the voronoi cell volume. All structural quantities correlate positively with ER on average, with the exception of Weighted Contact Number (WCN) which correlates negatively. For better illustration however, the Spearman's correlation coefficient (ρ) of the inverse of WCN with ER are shown in the Figure. Note that the Spearman's ρ is insensitive to the use of inverse WCN in place of WCN. The abbreviation *SC* refers to the use of average Side-Chain coordinates wherever used, and *CA* refers to the use of backbone C_α atomic coordinates for representation of individual sites in proteins.

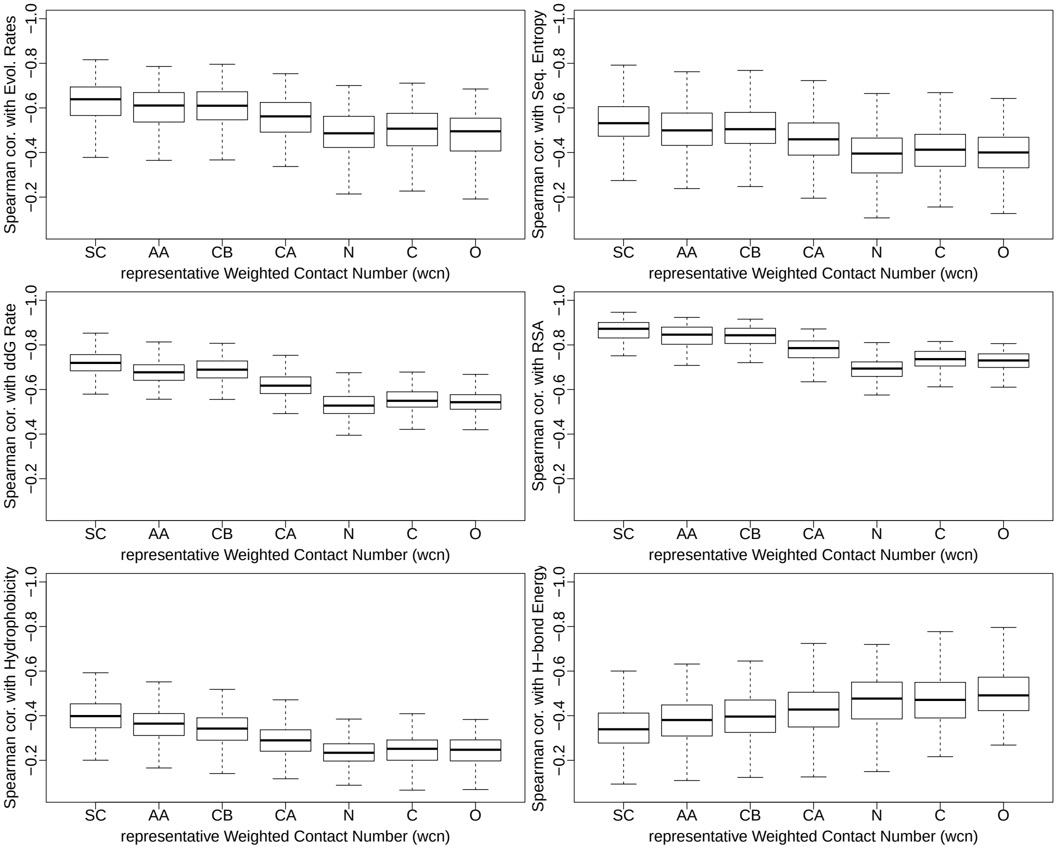

Figure 7: A comparison of the correlation strength of 6 different measures of Weighted Contact Number (WCN) with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. The contact numbers, WCN, are calculated using 6 sets of atomic coordinates: *SC, AA, CB, CA, N, C, O*, used as different representations of individual sites in proteins. The two labels *SC* & *AA* stand respectively for the geometric average coordinates of the Side Chain (SC) atoms and the entire Amino Acid (AA) atoms, excluding hydrogens.

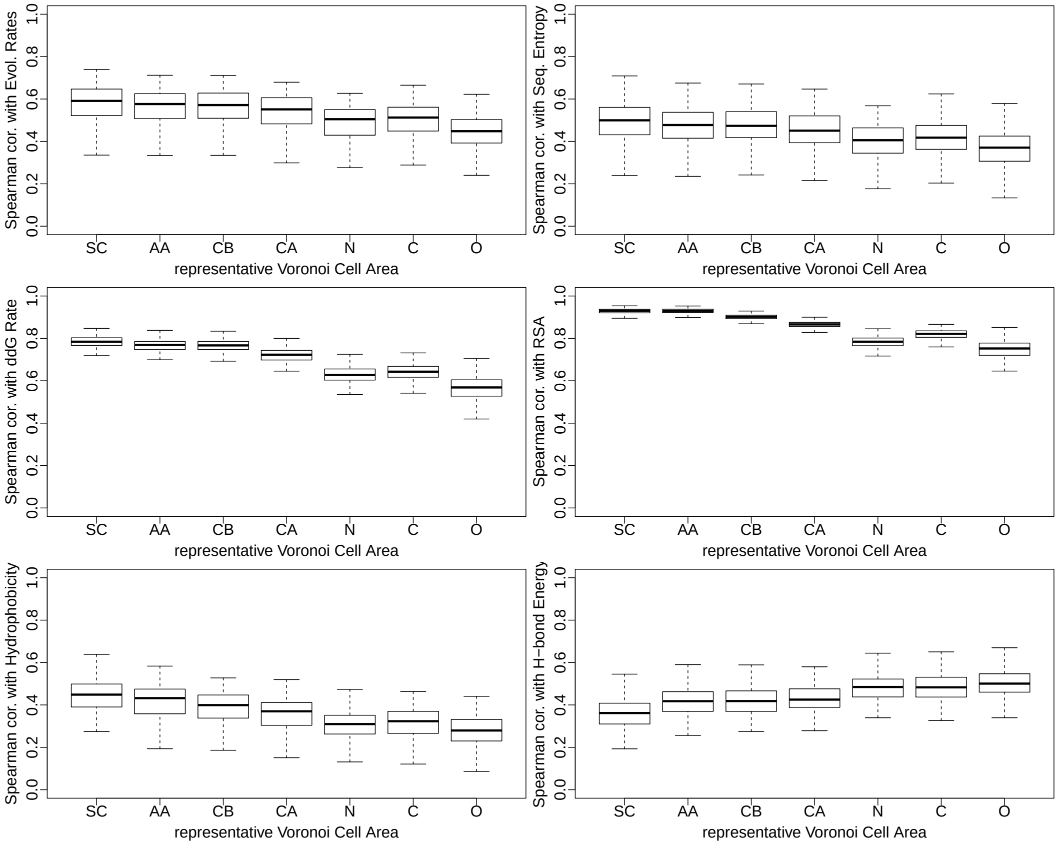

Figure 8: A comparison of the correlation strength of 6 different measures of Voronoi cell areas with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. The Voronoi cells are generated using 6 sets of atomic coordinates: *SC, AA, CB, CA, N, C, O*, used as different representations of individual sites in proteins. The two labels *SC* & *AA* stand respectively for the geometric average coordinates of the Side Chain (SC) atoms and the entire Amino Acid (AA) atoms, excluding hydrogens.