March 26, 2015

Predicting Sequence Variability from Voronoi Tessellation of Proteins

Amir Shahmoradi[1], 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 predictors of protein's sequence evolution? A number of site-specific structural properties have been proposed over the past decade to answer this question. The majority of these quantities however, depend on the set of atomic coordinates used to represent individual sites in proteins, and often involve one or more number of adjustable parameters in their definition. A number of studies have already demonstrated that the choice of C_α atomic coordinates may not be an optimal representation of the protein's 3-dimensional structure, in particular for the calculation of site-specific quantities such as Weighted Contact Number. Expanding on these studies and using a dataset of 209 monomeric enzymes, here we propose a new set of parameter-free structural variables derived from the Voronoi tessellation of protein structure which perform equally well or better than virtually all previously-considered structural quantities in predicting protein sequence evolution. We further show that the ideal representation of the 3-dimensional structure of proteins is the set of geometric average coordinates of atoms in the side chains of individual amino acids versus the common choice of backbone C_α coordinates.

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, 2014; 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., Wilke et al., 2005; 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 elastic network models (e.g., Liu and Bahar, 2012) and Molecular Dynamics (MD) simulations (e.g., Shahmoradi et al., 2014).

Although 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 influence the sequence evolution independently. It is perceivable that quantities such as B factor, RSA, and CN, all serve as a proxy measures of local packing density of individual sites in proteins, or the local flexibility of individual amino acids. Franzosa and Xia (2009) use a variety of structural variables representing the local packing density to show that RSA is the key determinant of sequence evolution with packing density having only peripheral influence. Recently however, Huang et al. (2014) have argued, through an extensive mathematical formulation within the framework of Elastic Network Models, for the local packing density as the dominant factor in sequence variability patterns in contrast to RSA and local flexibility measures.

It is notable that the site-specific flexibility is often represented by C_α atomic B factor, a quantity that is not necessarily an unbiased measure of the amino acid flexibility as a whole in a given site in protein. A more accurate measure of amino acid flexibility requires the calculation of accessible free volume to each site in protein structure. An estimate of the accessible volume for each site in protein can be generally obtained through a quantity widely known as Contact Number introduced and discussed by several authors (e.g., Liao et al., 2005). In its simplest mathematical form, the Contact Number for a given site in protein is defined as the number of amino acids within a fixed radius r of neighborhood around it (e.g., Franzosa and Xia, 2009). 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 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 $7\mathring{A}$ to $13\mathring{A}$ (e.g., Lin et al., 2008; Franzosa and Xia, 2009).

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,

$$WCN_i = \sum_{j \neq i}^{N} \frac{1}{{r_{ij}}^2}, \tag{1}$$

Although WCN is in general a better predictor of C_α atomic B factor and site-specific sequence variability, the proposed definition of WCN still involves an adjustable free parameter, the exponent of the power-law kernel, which is typically fixed to $\alpha = -2$ as shown in Eqn 1 (e.g., Yang et al., 2009). Moreover, no physical model has been so far proposed to support the power-law kernel used in the definition

of WCN and the specific value of exponent often used.

Motivated by the existing gaps in the current understanding of the role of flexibility and other structural properties on sequence-structure relations in proteins, here we propose and derive a new set of site-specific structural properties which, unlike CN and WCN, their definitions does not involve any free parameters, while performing equally well or better than all previously-considered structural quantities in predicting protein sequence evolution. 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 local packing density and site-specific flexibility. Contrary to what is currently perceived about the role of flexibility in sequence variability (e.g., Huang et al., 2014), we show that the newly calculated flexibility measures outperform many of previously studied structural properties, such as RSA and the traditional definitions of Contact Number (CN) and the Weighted Contact Number (WCN), in predicting sequence evolution at residue level.

Furthermore, for structural properties that are calculated based on a set of representative site coordinates, we show that the choice of the geometric average of the side chain atomic coordinates instead of the traditional choice of C_α atomic coordinates, always results in significantly better predictions of site-specific sequence evolution. Similar improvements in correlations with different ste-specific structural properties and sequence variability measures are also observed if the average of side chain B factors, instead of C_α atomic B factor, is used as a proxy measure of site flexibility.

2 Methods

Protein Dataset

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 by 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 site-specific evolutionary rates, we also calculated 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:

$$H_i = -\sum_j P_{ij} \ln P_{ij} \tag{2}$$

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. The *ddG rate* estimates for all structures in the dataset were calculated using FoldX program (c.f., Echave et al., 2014, for details of the methodology employed). In brief, the site-specific quantity, ddG rate, is a proxy measure of the stability

of the entire structure of protein upon substituting an amino acid in a given site with all other 19 amino acids. Therefore, a low ddG rate for a given site would indicate a high chance of structure perturbation upon substitution and therefore high conservation of the specific amino acid in the site on evolutionary timescales.

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`.

Voronoi Tessellation

There is already extensive body of literature on the applications of different methods of structural partitioning in the studies of protein structure and its prediction from sequence (e.g., 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 development of empirical potentials (e.g., 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 space into regions, called *cells*, such that the cell for each centroid point consists of every region in space whose distance is less than or equal to its distance to any other centroid points (Figure 1).

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 protein according to Voronoi tessellation. The properties of individual cells resulting from tessellation are then used to obtain a wide range of information on protein structure, energy landscape or protein–protein interactions. Here in this work, we apply the simplest and most widely used definition of Voronoi tessellation described above on a 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, cell area and volume, number of faces of each cell, the cell eccentricity defined as the distance between the cell's seed and the geometrical center of the cell. In addition, for each cell we also calculate *sphericity*, a measure of the cell *compactness* defined as,

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

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.

Eliminating Degeneracy in Structural Property Definitions

Depending on the choice of 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 (e.g., Soyer et al., 2000). More recently, Marcos and Echave (2014) 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 (e.g., 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 we consider all possible choices of the representative set of atomic coordinates in order to identify which set of atomic coordinates best represents individual sites for the calculation of WCN, B factor, and Voronoi cells. 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 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.

3 Results

Voronoi Cell Area and Volume as Proxy Measures of Local Packing Density and Flexibility in Proteins

In order to assess the prediction power of site-specific variables derived from Voronoi tessellation, we first calculate the geometric centers of all side-chains for each of the proteins in dataset and use them as the seeds of Voronoi polyhedra. Plot A of 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 other Voronoi cell properties. In general, we observe that the cell surface area has the best prediction power compared to other cell characteristics, followed by the cell volume, cell eccentricity as defined in previous section, cell's total edge length, and the cell sphericitiy. The cell properties are also highly associated with each other. Although the Voronoi cell volume is the second best correlating variable with ER, it exhibits no significant independent correlation with ER once we control for the cell area, with the median of its distribution centered at ~ 0.0, as illustrated in plot B of Figure 3. On the contrary, 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 almost represent 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. These cell characteristics are however, discrete (integer) quantities and in general have a limited range.

Not shown here for brevity, we also obtain identical results to above if we use sequence entropy as defined by Eqn. 2 instead of sequence evolutionary rates. The use of sequence entropy however, generally results in weaker correlation strengths due to the discreteness and limited range inherent in its definition.

Average Side Chain coordinates as the Best Representation of Protein 3D Structure

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 Weighted Contact Number, B factor and Voronoi cell properties. 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.

For the measure of local packing density in proteins (the Weighted Contact Number) we find that among all possible set of coordinates, the average over coordinates of all heavy atoms of each individ-

ual 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 outperforms WCN based on C_α coordinates in predicting RSA, $\Delta\Delta G$ entropy, sequence entropy and evolutionary rates with median Spearman correlation differences of 0.09, 0.10, 0.07 & 0.08, respectively (Figure 4).

For the measure of local flexibility in proteins (B factor) we similarly find that among all 7 representative measures of site B factors, the average of B factor values over all heavy atoms of each individual side chain results in the best correlations with other structural and sequence properties. Specifically, the average side chain B factor outperforms the commonly used C_α B factor 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 5).

Similar to WCN and Bfactor, 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 (Figure 3).

4 Discussion

Throughout this work, we have carried out a comprehensive analysis and comparison of the main structural determinants of sequence variability, some of which are newly reported and discussed in this work. Examples of sequence–structure relations include the correlations of measures of evolutionary rates such as *r4sJC* used in this work, and sequence entropy with measures of residue Contact Number, Relative Solvent Accessibility (RSA), and $\Delta\Delta G$ Rate as defined by Echave et al. (2014) which is essentially a proxy measure of stability of protein's native conformation upon substitution of amino acids in individual sites in proteins. In addition, we have derived new site-specific characteristics from the Voronoi Tessellation of protein 3D structures, that are capable of explaining sequence variability equally well or better than several previously considered structural quantities, such as B factor, RSA, $\Delta\Delta G$ Rate, and the traditional definitions of Contact Number (CN) and the Weighted Contact Number (WCN) using C_α atomic coordinates (e.g., Figure 6).

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. To ensure that these edge effects do not influence the observed sequence-structure correlations, we identified open cells – i.e., cells that are partially bounded and closed by the cubic box containing the protein – in all proteins by examining the variations in individual cell volumes upon changing the size of the cubic box containing the protein. The open cells in individual proteins were 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.

Our conclusion is that *edge effects* due to Voronoi tessellation have $\lesssim 0.01$ influence on the observed sequence-structure correlation strengths in the dataset of 209 proteins considered in this work. Similar conclusions are reached if the open cells were alternatively ranked by the fractional changes in cell area (vs. cell volume) upon changing the box size. Our conclusion is that Voronoi cell characteristics, in particular cell volume and area can be safely used in predicting sequence variability without recourse to corrections for edge effects. An exception is cell sphericity as defined in Eqn. 3, which turns out to behave differently for open and closed cells. This is well illustrated in the adjacent averaging plots of Figures 7 & 8 where we plot the behavior open and closed cell characteristics averaged over all sites in

all proteins in our dataset, versus sequence evolutionary rate and sequence entropy.

The definitions and values of some site-specific structural quantities, such as WCN and Voronoi cell characteristics, depend on the set of atomic coordinates that are used for representing individual sites in proteins. In order to identify which set of atomic coordinates best represent protein 3-dimensional structure for the calculation of structural quantities, we calculated WCN and Voronoi cell properties using different sets of atomic coordinates representing individual sites in proteins. Moreover, there is also ambiguity as to which set of atomic B factors best represent the flexibility of individual sites in protein.

Our finding is that sequence variability measures correlate best with site-specific structural quantities such as WCN and Voronoi cell characteristics if the geometric center of the side-chains of amino acids are used for the calculation of these structural quantities instead of using the more commonly adopted backbone C_α atomic coordinates. In addition, the average of side-chain atomic B factors appear to correlate significantly better with sequence variability measures and site-specific structural quantities such as Relative Solvent Accessibility, $\Delta\Delta G$ Rate, and measures of amino acid hydrophobicities as illustrated in the plots of Figure 5.

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 rather than by the properties and the spatial coordinates of the backbone C_α atoms. In particular, the strength of structure-structure and sequence-structure correlations decrease monotonically when moving from side chain to backbone atoms. An exception to this general pattern is the correlation of the hydrogen-bond energies of individual sites with other site-specific structural properties. In general, average site-specific H-bond energies correlate more strongly with representative B factors, contact number, and Voronoi cell characteristics, if these quantities are calculated using the backbone Oxygen atoms as representation of individual sites, instead of using the average side chain properties and coordinates. This is well illustrated in the bottom-right plots of Figures 3, 4, & 5. The observed monotonic increase in the correlation strengths with H-bond energies from side chain to backbone O atom can be explained away knowing that the backbone Oxygen atom is responsible for virtually all Hydrogen bonds in proteins. The influence of individual atoms on H-bond energies monotonically decreases from the neighbor backbone atoms C & N to the farthest atoms from Oxygen in the amino acid side chains.

Regardless of the set of atomic coordinates used for representing individual sites in proteins and for calculating site-specific structural quantities it is notable that once WCN is recalculated using the same set of coordinates used for Voronoi tessellation of protein structure, the quantity WCN as defined by Eqn 1 outperforms all other structural quantities – including those derived from Voronoi tessellation – in explaining site-specific sequence variability measured, as depicted in Figure 6. This might not be surprising, knowing that the mathematical definition of WCN includes all sites in proteins which are all likely involved in long-range correlations and electrostatic interactions. In contrast to WCN, measures of site flexibility, in particular, the Voronoi cell volume or its inverse as a more local measure of packing density than WCN, do not include the possible effects of long-range interactions among non-neighbor amino acids in proteins. In this regard, it is reasonable to expect the Voronoi cell area to perform slightly better than cell volume in explaining the general patterns of sequence evolutions, since the Voronoi cell area is primarily determined by the local packing in addition to the number of neighboring amino acids around a given site of interest in protein, whereas the cell volume is more representative of the local packing and site flexibility than the number of neighboring amino acids around the site of interest.

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.

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. 2014. Too packed to change: site-specific substitution rates and side-chain packing in protein evolution. *bioRxiv*. p. 013359.

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. Voronoi 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.

Wilke CO, Bloom JD, Drummond DA, Raval A. 2005. Predicting the Tolerance of Proteins to Random Amino Acid Substitution. *Biophysical Journal.* 89:3714–3720.

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.

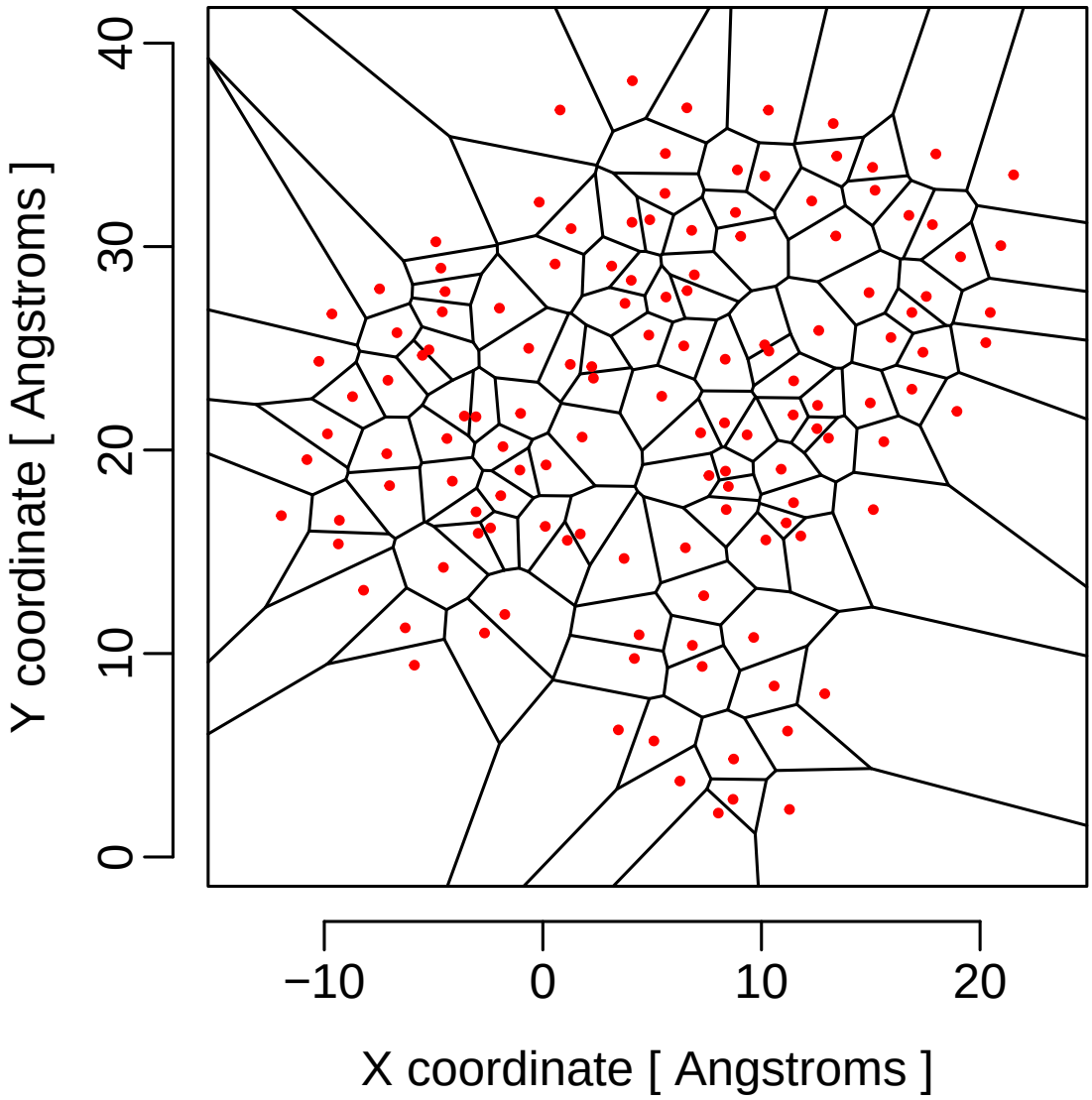

Figure 1: Example 2-dimensional Voronoi diagram for bacteriophage T7 lysozyme (Protein Data Bank ID '1LBA'). The red dots represent the backbone C_α atoms projected on the X–Y plane, used as cell seeds in Voronoi tesellation.

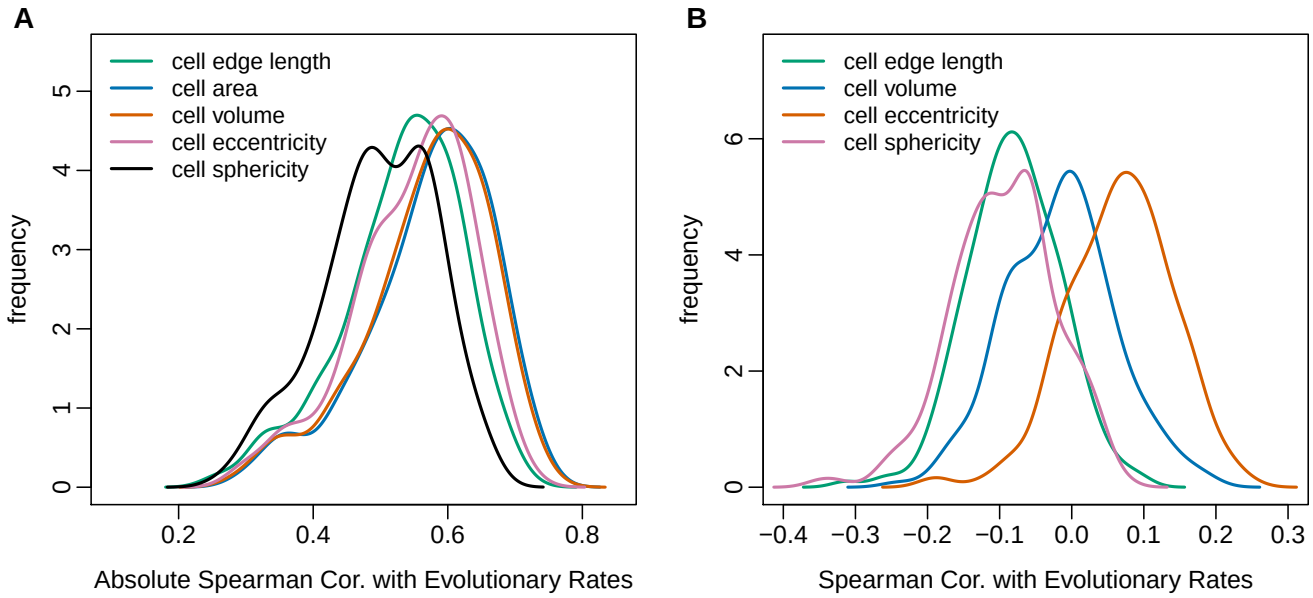

Figure 2: **A:** A comparison of the prediction power of different Voronoi cell characteristics about site-specific evolutionary rates (ER). Note that all cell characteristic correlate positively with ER, except sphericity which strongly negatively correlates with ER. **B:** The partial correlation of the same Voronoi cell characteristics with sequence evolutionary rates while controlling for the cell area.

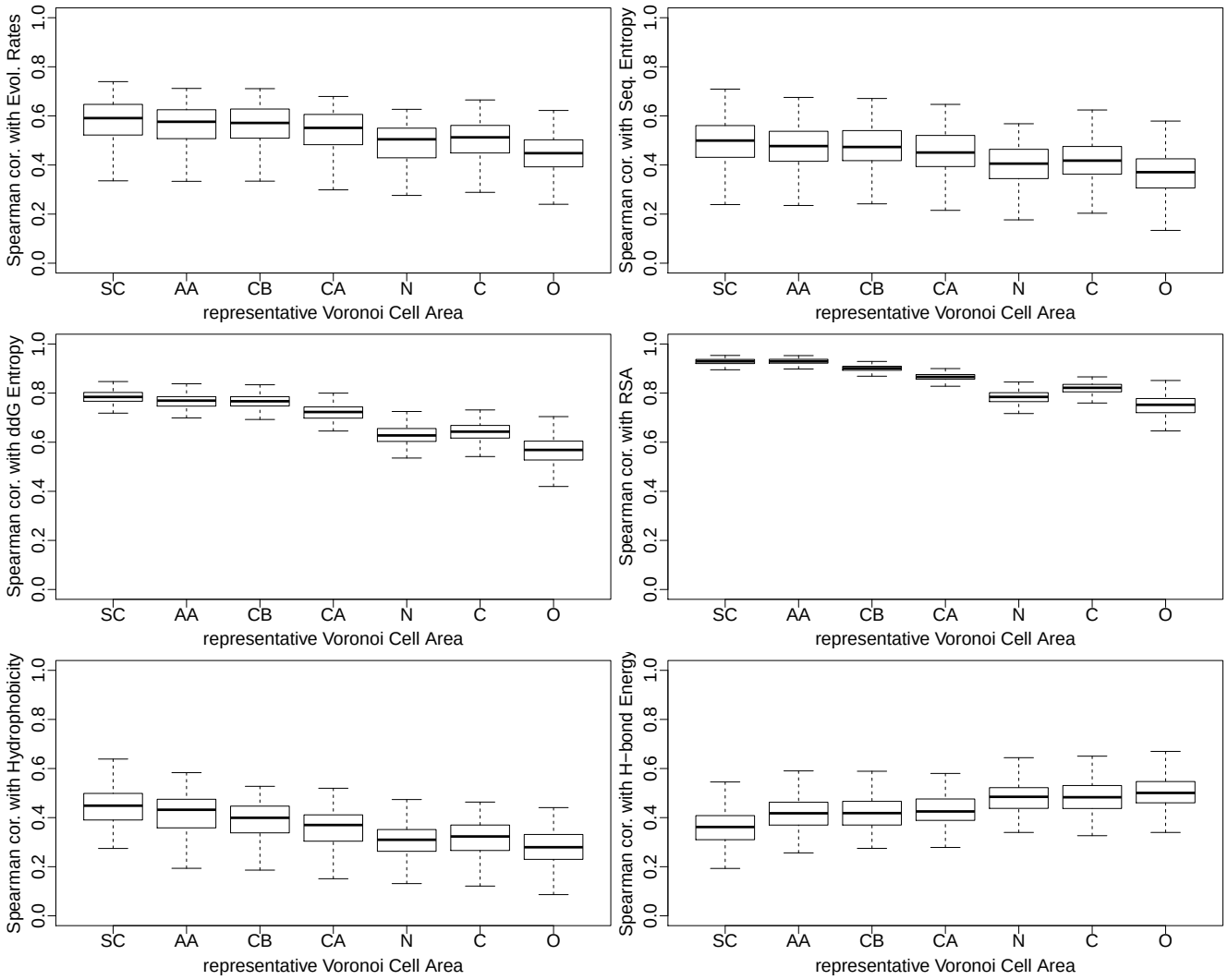

Figure 3: A comparison of the correlation strength of 6 different measures of Voronoi cell areas with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. The Voronoi cells are generated using 6 sets of atomic coordinates: *SC, AA, CB, CA, N, C, O*, used as different 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.

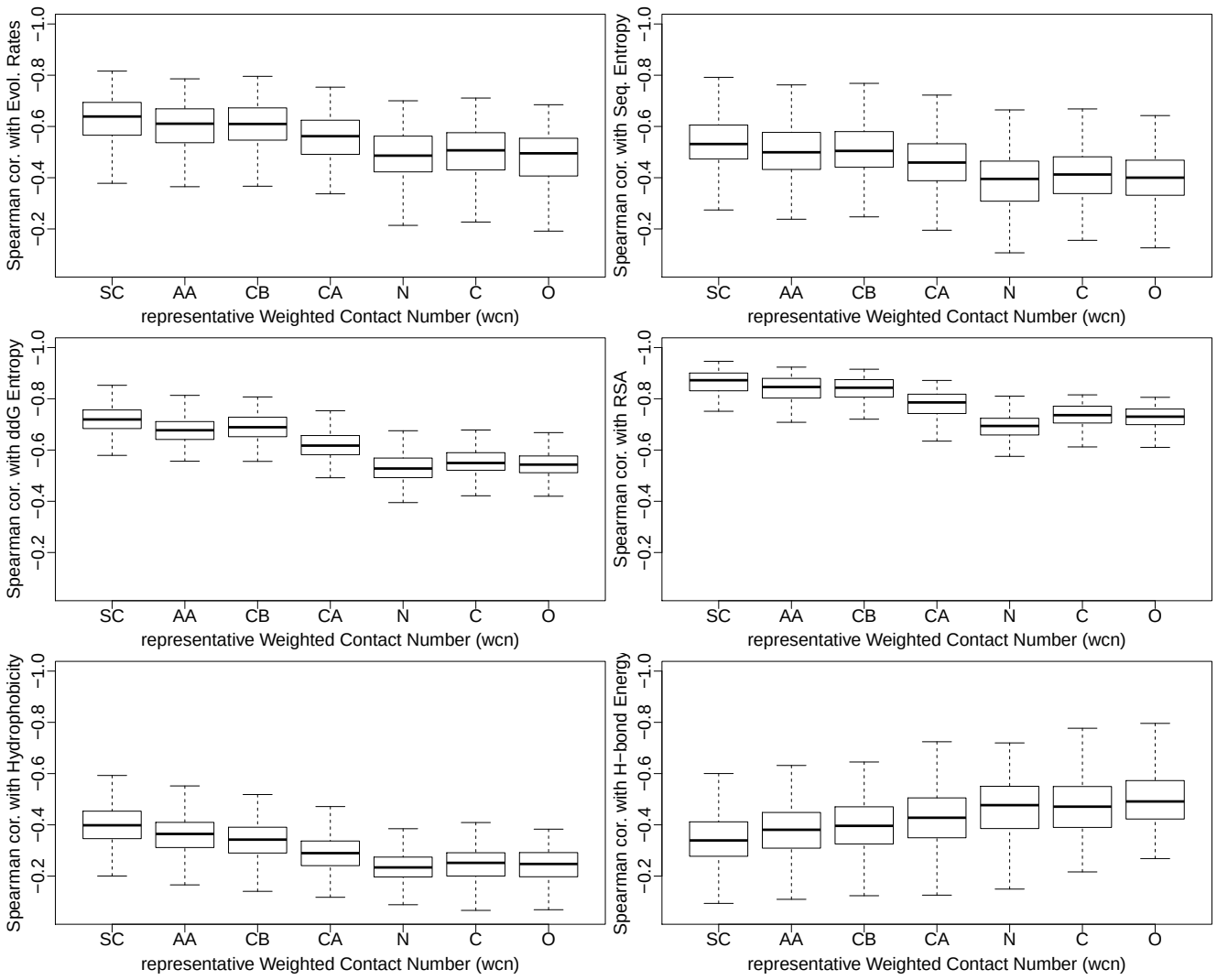

Figure 4: A comparison of the correlation strength of 6 different measures of Weighted Contact Number (WCN) with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. The contact numbers, WCN, are calculated using 6 sets of atomic coordinates: *SC, AA, CB, CA, N, C, 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.

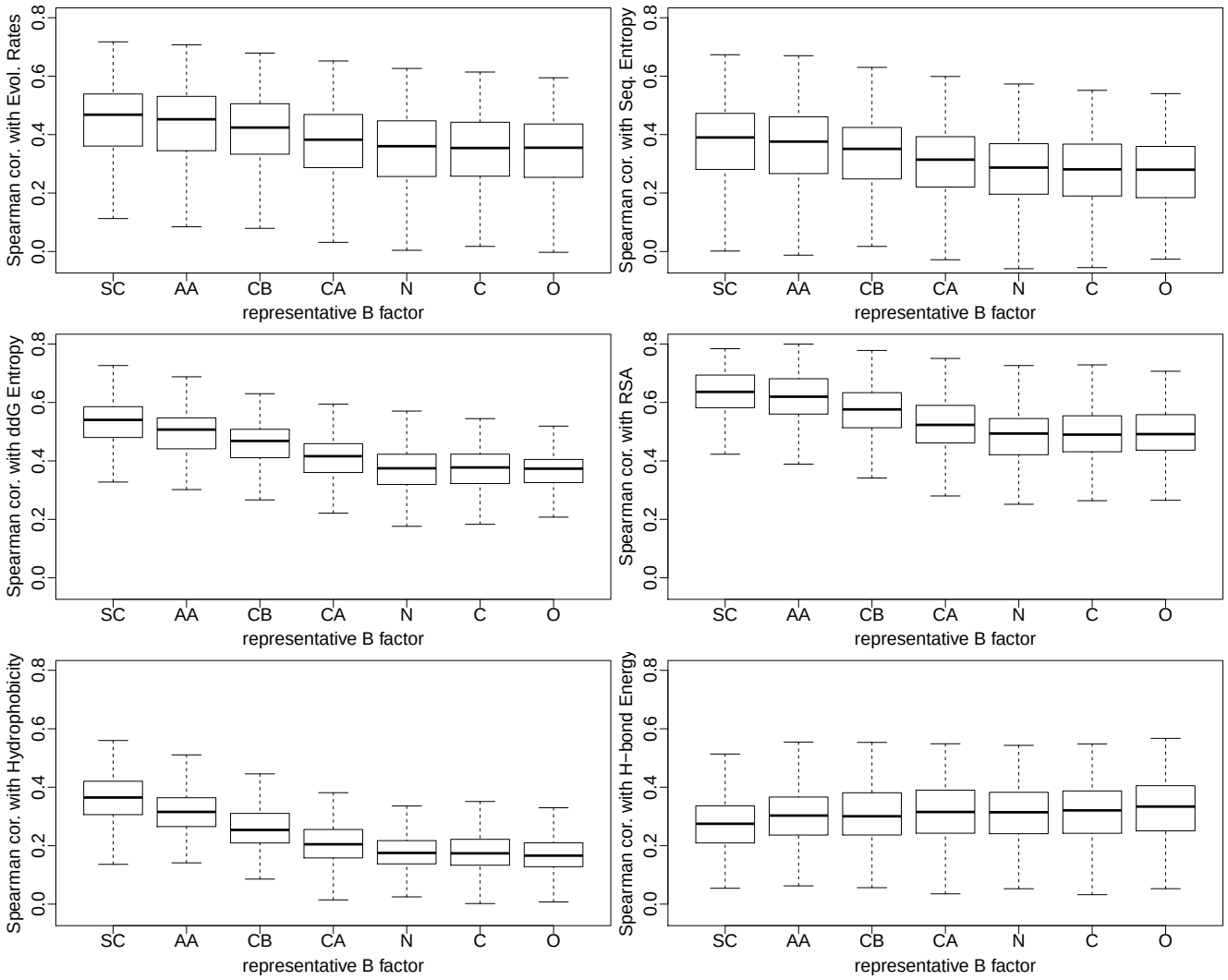

Figure 5: A comparison of the correlation strength of 6 different measures of B factor with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. Shown on the horizontal axes, are the 6 representative atomic B factors: *SC, AA, CB, CA, N, C, O* used as flexibility measures of individual sites in proteins. The two variables *SC* & *AA* stand respectively for the average B factor of all Side Chain (SC) atoms and the entire Amino Acid (AA) atoms, excluding hydrogens.

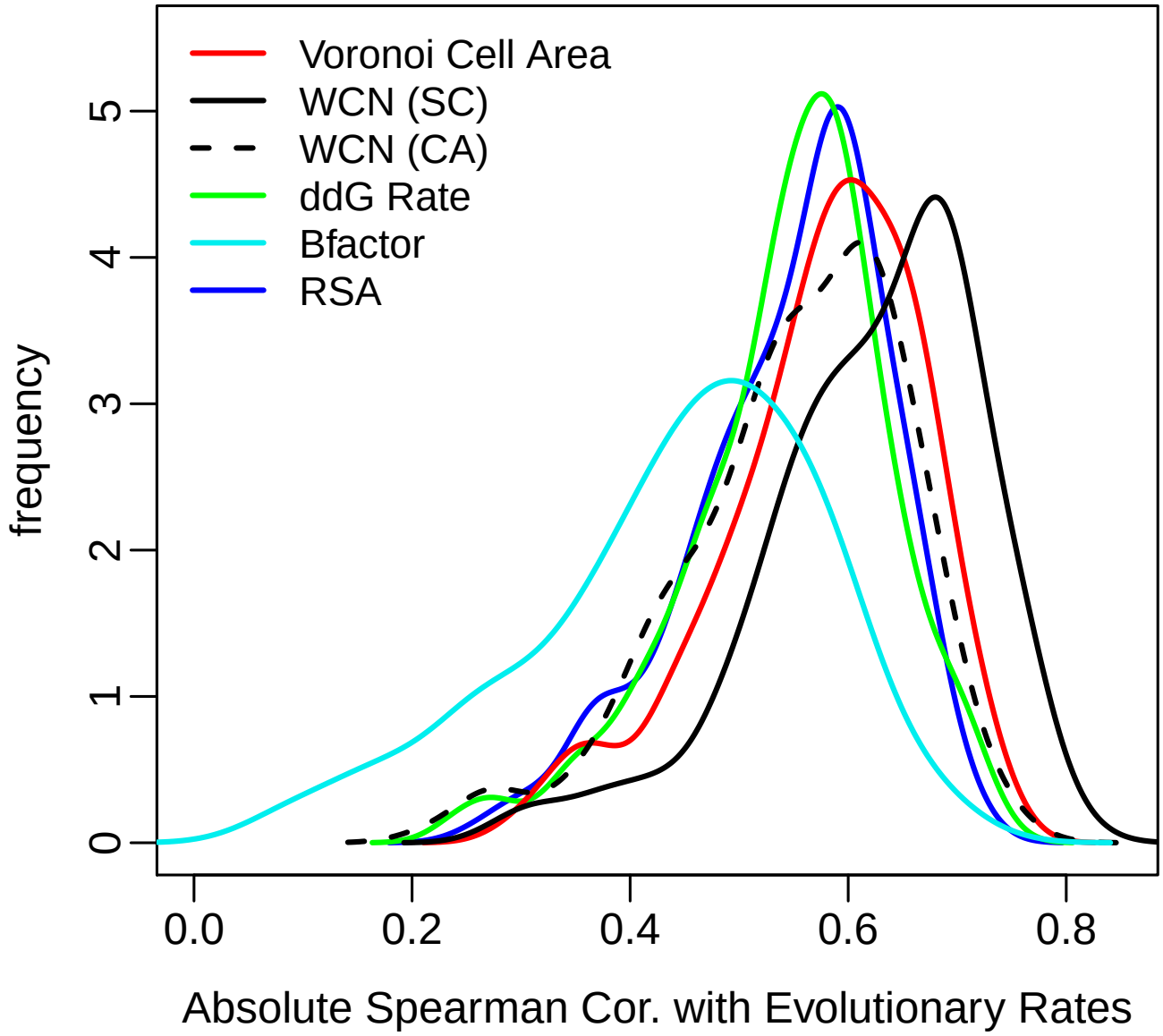

Figure 6: A comparison of the prediction power of four structural variables about site-specific evolutionary rates (ER). All structural quantities correlate positively with ER, with the exception of Weighted Contact Number (WCN) which correlates negatively. For better illustration however, the absolute Spearman's correlation of WCN with ER are shown in the Figure.

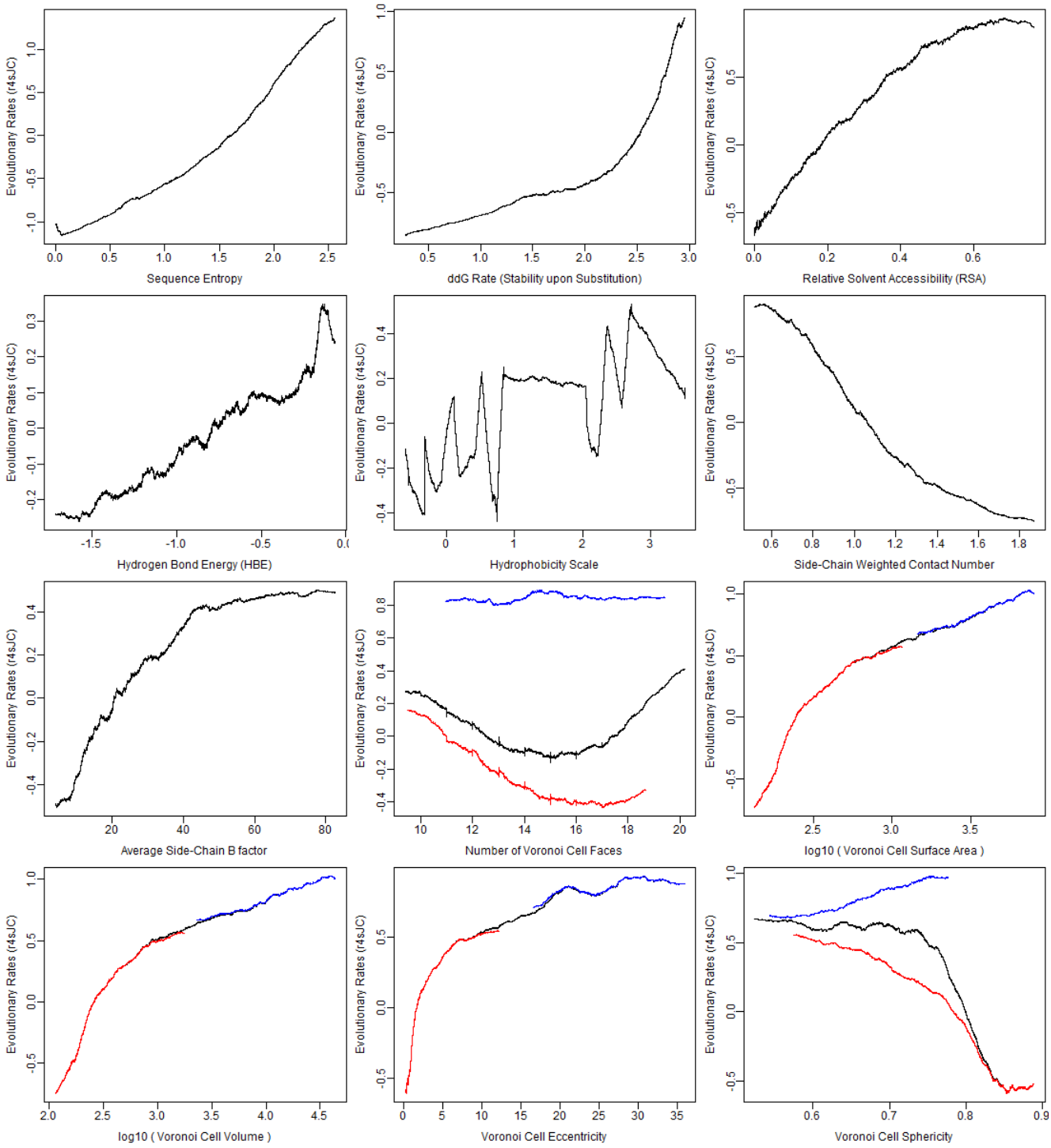

Figure 7: General behavior of site-specific structural characteristics versus site-specific evolutionary rates among all sites in all 209 proteins in dataset. For plots depicting evolutionary rates versus versus Voronoi cell characteristics, the red & blue curves represent respectively the general behaviors of closed & blue open Voronoi cells, in addition to the black curves which represent the behavior of all open and closed cells together. The amino acid hydrophobicity scales were taken from Hessa et al. (2005).

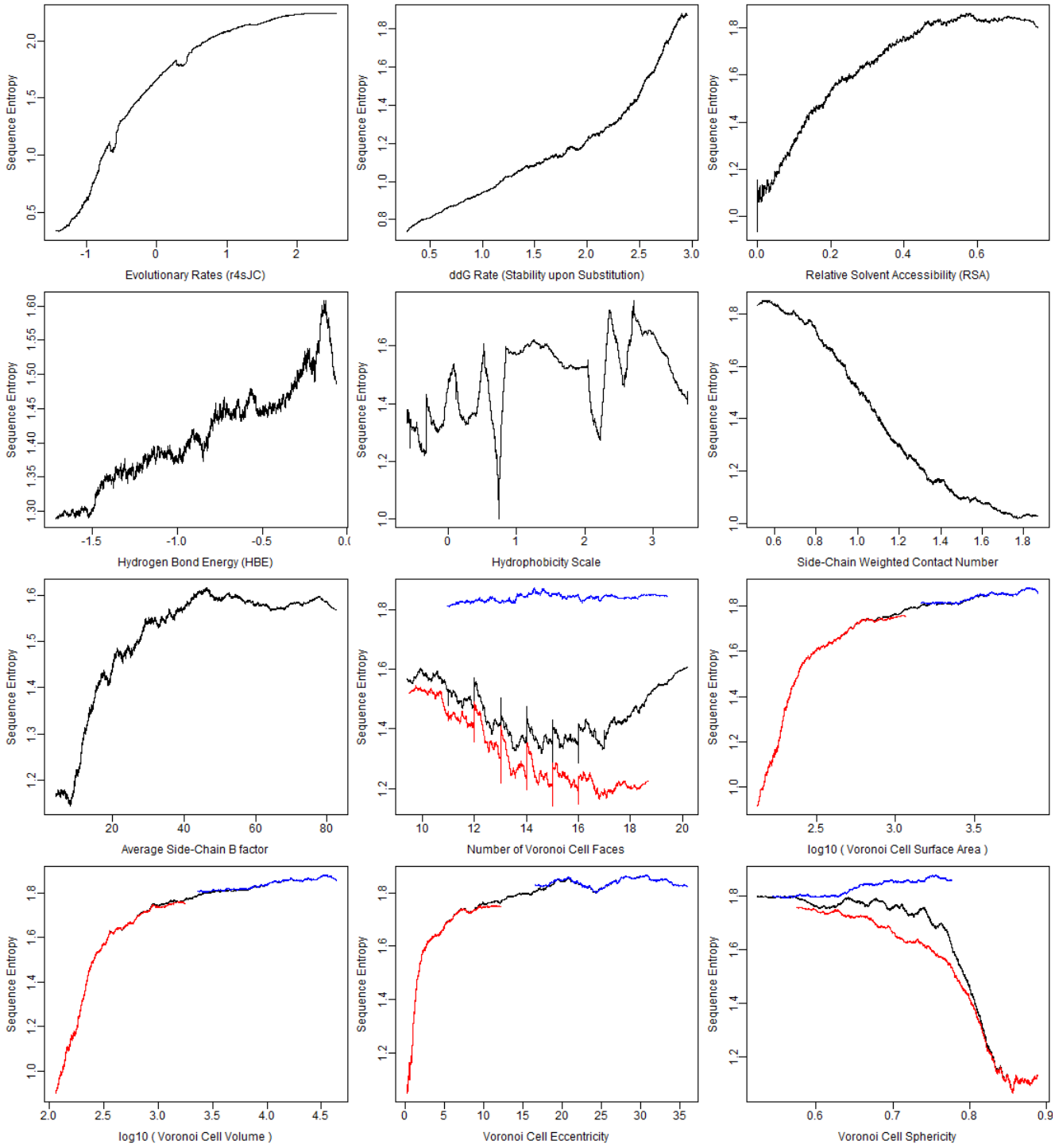

Figure 8: General behavior of site-specific structural characteristics versus site-specific sequence variability, as measured by sequence entropy (Eqn. 2), among all sites in all 209 proteins in dataset. For plots depicting sequence entropy versus Voronoi cell characteristics, the red & blue curves represent respectively the general behaviors of closed & blue open Voronoi cells, in addition to the black curves which represent the behavior of all open and closed cells together. The amino acid hydrophobicity scales were taken from Hessa et al. (2005).