March 13, 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) (xx), Contact Number (xx), measures of thermodynamic stability changes due to mutations at individual sites in proteins (e.g., the quantity ddG rate in)(xx echave 2014), and measures of local flexibility, such as the Debye-Waller factor (hereafter B factor) (xx) or flexibility measures based elastic network models (xx Bahar et al.) and Molecular Dynamics (MD) simulations (xx shahmoradi 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 (xx) 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 (xx) have argued, through an extensive mathematical model, for the local packing density as the dominant factor in sequence variability patterns, as opposed to RSA and local flexibility measures.

It is notable that 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 originally introduced by (xx). 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 (xx). 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, franzosa, xx), and sometimes up to $18\mathring{A}$ (e.g., xx).

In an attempt to provide a more general definition of CN, some studies (xx) 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 has been traditionally fixed to $\alpha = -2$ as shown in Eqn 1. 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, are not defined with 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 site-specific quantities for 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, 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 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 taken from xx Echave papers, Wilke ddg paper xx randomly picked from the Catalytic Site Atlas 2.2.11 (Porter et al. 2004) with protein sizes in the sample ranging from 95 to 1287, including representatives from all six main EC functional classes (Webb 1992) and domains of all main SCOP structural classes (Murzin et al. 1995). To assess the evolutionary rates at the amino acid level for each protein, first a set of up to 300 homologous sequences were collected by (Yeh et all xx) 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. 2002, 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, hereafter abbreviated as *seqent* – 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 (xx) for the calculation of the Accessible Surface Area (ASA) for each site normalized by the theoretical maximum solvent accessibility values of Tein et al (20112 xx) to obtain the Relative Solvent Accessibility (RSA) for all individual sites in all proteins. The *ddG rate* estimates are taken for all structures in the dataset were all calculated using FoldX program (c.f., Echave 2014 for the details of the methodology employed xx). 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 (Richards 1974, Gerstein 1994 xx). 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 (xx). 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 (Rycroft2009xx) 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 degeneracies in the definition of the 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 are the best representatives for 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 (soyer2000xx). More recently, Marcos2015 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.

In order to identify which set of atomic coordinates best represents individual sites for the calculation of WCN, B factor, and Voronoi cells, here we consider all possible choices of the representative set of atomic coordinates. 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 a Measure of Local Packing Density in Proteins

In order to assess the prediction power site-specific variables derived from Voronoi tessellation, we first calculate the geometric centers of all side-chains for each of the proteins in the 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 should be noted 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 the 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 the above if we use sequence entropy as defined by Eqn. 2 instead of sequence evolutionary rates, although sequence entropy 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 7 different measures for 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 individual side chain results in WCN values that show the best correlation with other structural and sequence properties, such as RSA, Voronoi cell properties, sequence entropy, and evolutionary rates. Specifically, WCN from average side chain coordinates (wcnSC) outperforms WCN based on C_α coordinates (wcnCA) in predicting RSA, $\Delta\Delta G$ entropy, sequence entropy and evolutionary rates (r4sJC) by a median Spearman correlation difference of 0.09, 0.10, 0.07 & 0.08, respectively (Figure 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 (bfSC) results in the best correlations with other structural and sequence properties. Specifically, bfSC outperforms the commonly used C_α B factor (bfCA) in predicting RSA, $\Delta\Delta G$ entropy, sequence entropy and evolutionary rates by a median Spearman correlation difference of 0.11, 0.12, 0.08 & 0.09, respectively (Figure 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 and Concluding Remarks

Throughout the previous sections, a new set of parameter-free site-specific structural quantities were introduced using Voronoi tessellation that are capable of explaining sequence evolutionary rates equally or better than many of the previously considered site-specific structural characteristics, including RSA, ddG rate as defined by Echave et al. (2014xx), site-specific measures of flexibility such as B factor, and the traditional definition of Contact Number and the Weighted Contact Number using C_α atomic coordinates. It is however notable that, once we recalculate WCN using the geometric center of the side chains as the representative coordinates of individual sites, the quantity WCN still outperforms all other structural quantities, including those derived from Voronoi tessellation, in explaining site-specific evolutionary rates.

All observations clearly demonstrate that individual sites in proteins are best represented by the average properties of the side chains of amino acids in the corresponding sites. In particular, the strength of structure and sequence correlations decrease when moving from side chain to backbone atoms. An exception to this general pattern is the correlation of the hydrogen-bond energies of the sites, which correlate more strongly with site characteristics calculated based on the backbone atoms instead of side chain.

Based on the observations described in the previous paragraphs, we keep only variables measured from average side chain properties and coordinates throughout the rest of the analysis and omit all other similar measures that show only weaker correlations with other site-specific characteristics. The exclusion of these alternative measures results in a significant reduction in the number of pdb-level variables to be further analysed, without compromising generality and comprehensiveness of the analysis.

One potential caveat with Voronoi tessellation is the edge effects, that is, Voronoi cells that are on the surface of the protein are open. To ensure that

Throughout this work, we have carried out a comprehensive analysis in search for the main determinants of the strength of sequence–structure correlations – some of which are newly reported and discussed in this work. Examples of sequence–structure relations include the correlations of sequence entropy (*seqent*) and measures of evolutionary rates (such as *r4sJC* used in this work) with measures of residue Contact Number (e.g., *wcnSC*), Relative Solvent Accessibility (RSA), $\Delta\Delta G$ entropy (*ddgent*). In addition, we have derived new site-specific properties, based on Voronoi Tessellation of protein 3D structures, that are comparable to or better than several previously known structural properties in explaining site-specific sequence entropy or evolutionary rates (e.g., Figure **??**). Prime examples include Voronoi cell volume (*vvolume*), surface area (*varea*) and

We have also shown that site-specific structural properties – such as Weighted Contact Number, Bfactor and Voronoi Cell properties – that are calculated from the average coordinates of side chain atoms, have the best explanatory powers for the sequence variability measures such as *seqent* and *r4sJC*. Compared to the common choice of back bone *CA* atomic coordinates, site-specific properties averaged over side chain atoms can outperform in predicting sequence evolutionary rates by as much as 0.12 in

terms of Spearman correlation strength.

In search for the determinants of the strength of sequence-structure relations, we compiled a set of more than 200 protein properties for a dataset of 209 monomeric enzymes. By employing several independent parametric and non-parametric statistics, such as Spearman rank test, regularized regression and Principal Regression methods, we identify sequence divergence as the dominant factor in the strength of sequence-structure correlations, capable of explaining $10 - 30\%$ of the observed correlation strengths alone, in both the original and rank-transformed data.

ACKNOWLEDGEMENTS

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

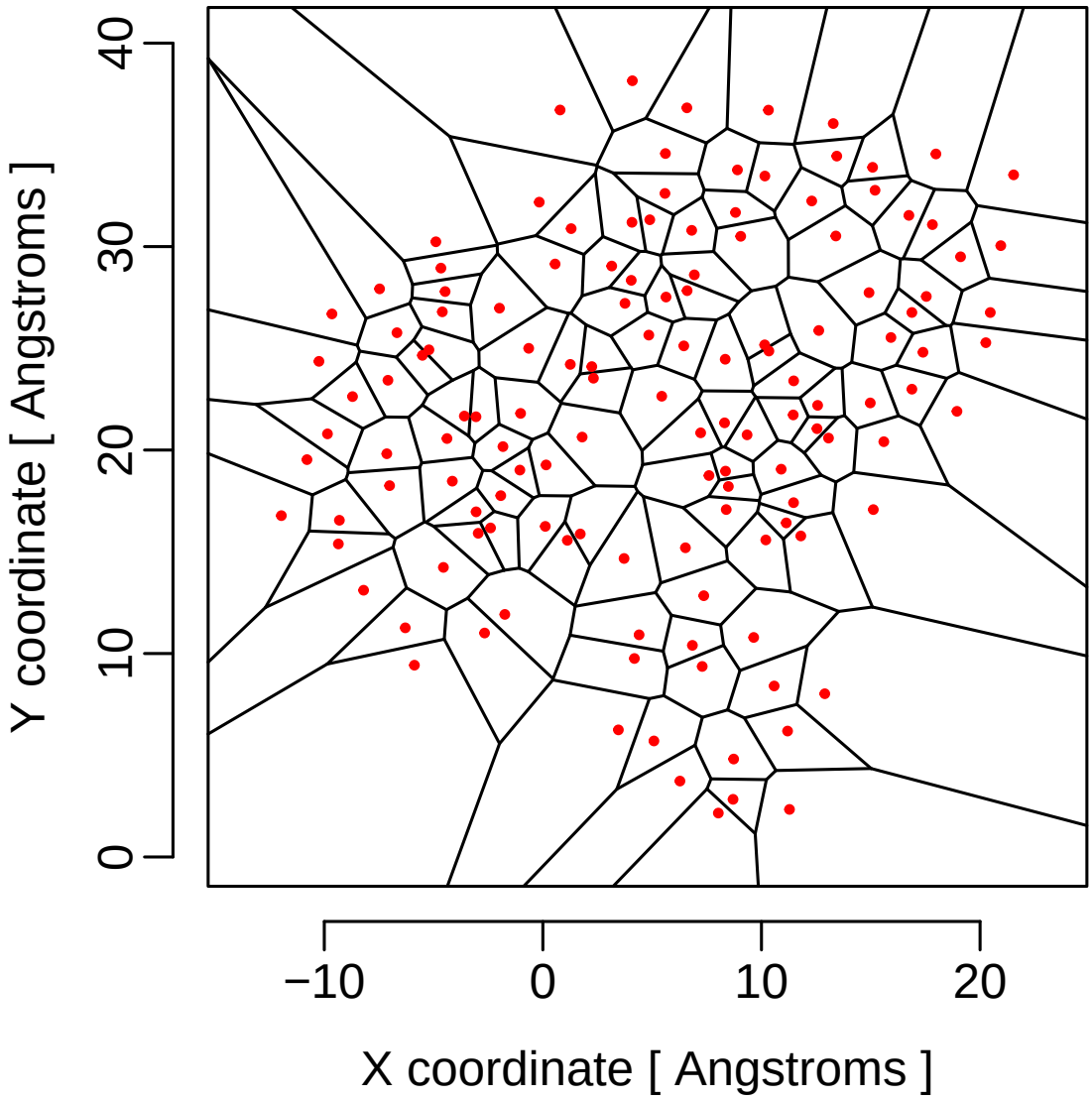

Figure 1: Example 2-dimensional Voronoi diagram for bacteriophage T7 lysozyme (Protein Data Bank ID '1LBA'). The red dots represent the backbone C_α atoms projected on the X–Y plane, used as cell seeds in Voronoi tesellation.

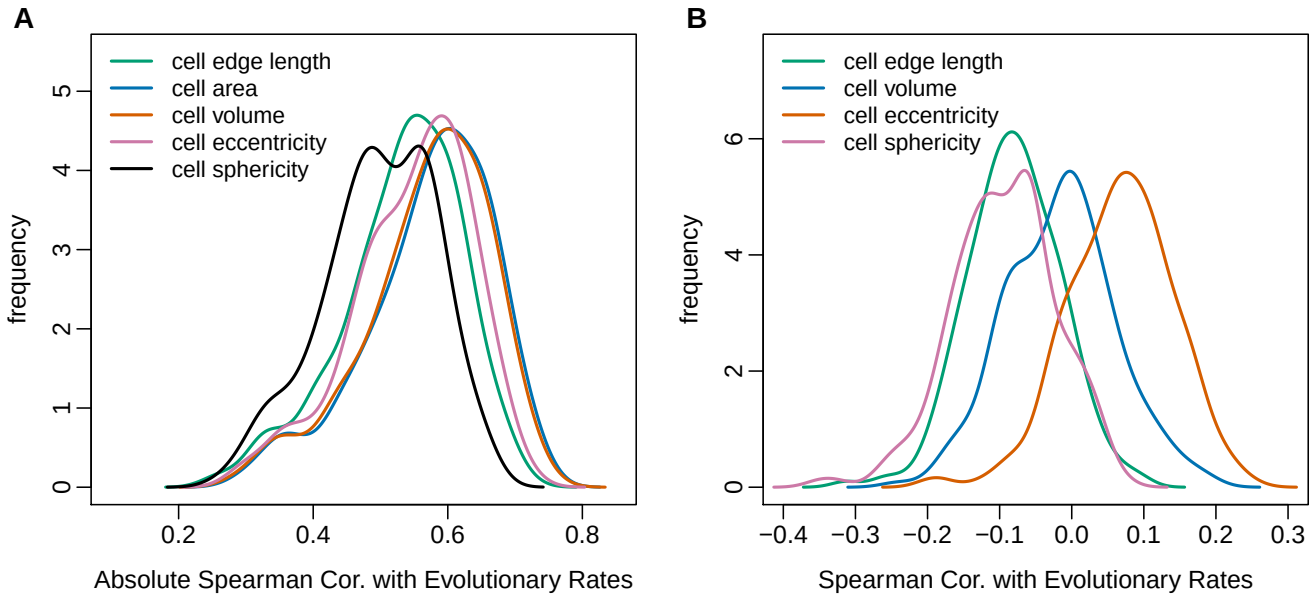

Figure 2: **A:** A comparison of the prediction power of different Voronoi cell characteristics about site-specific evolutionary rates (ER). Note that all cell characteristic correlate positively with ER, except sphericity which strongly negatively correlates with ER. **B:** The partial correlation of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

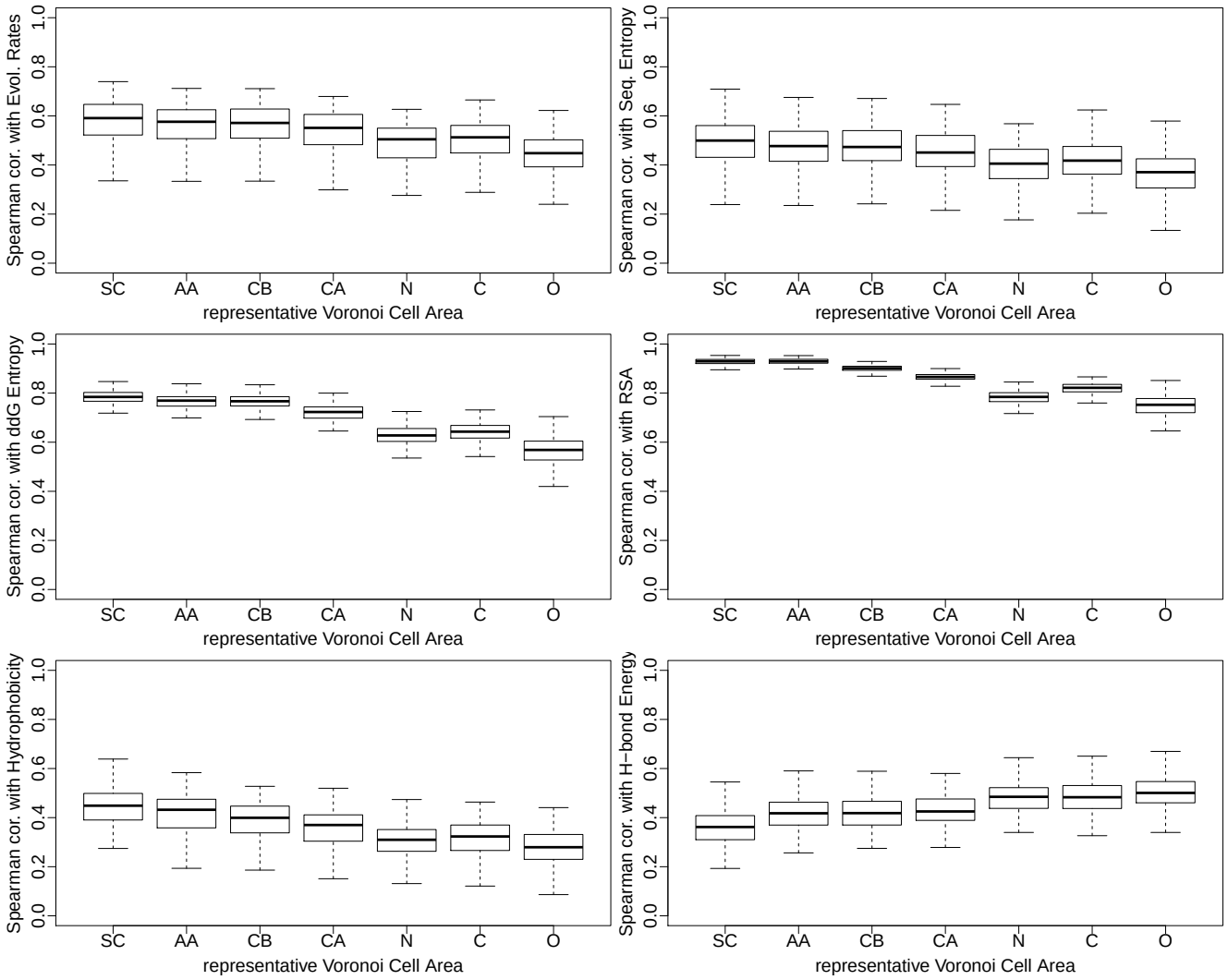

Figure 3: A comparison of the correlation strength of 6 different measures of Voronoi cell areas with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. The Voronoi cells are generated using 6 sets of atomic coordinates: *SC, AA, CB, CA, N, C, O*, used as different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

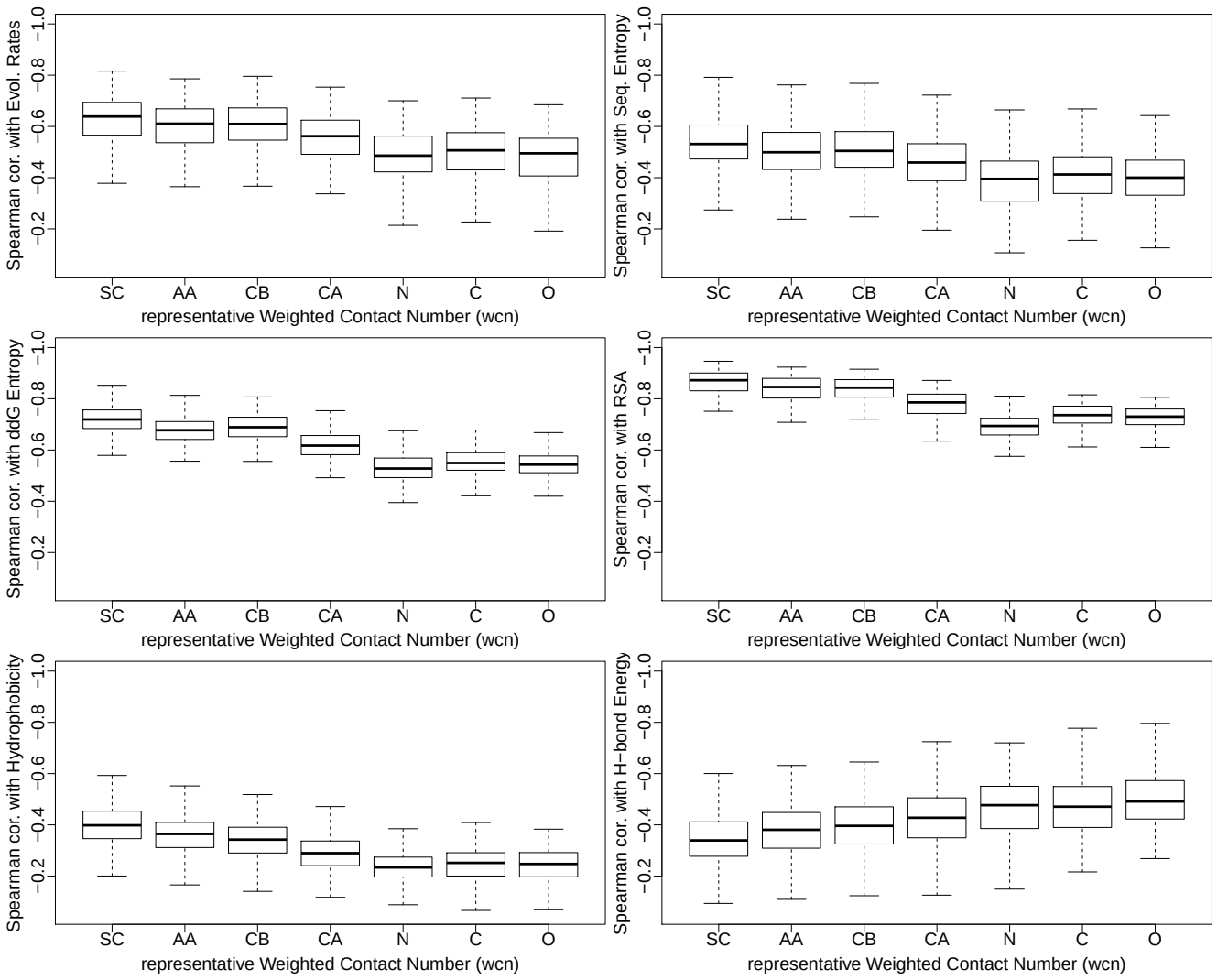

Figure 4: A comparison of the correlation strength of 6 different measures of Weighted Contact Number (WCN) with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. The contact numbers, WCN, are calculated using 6 sets of atomic coordinates: *SC, AA, CB, CA, N, C,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

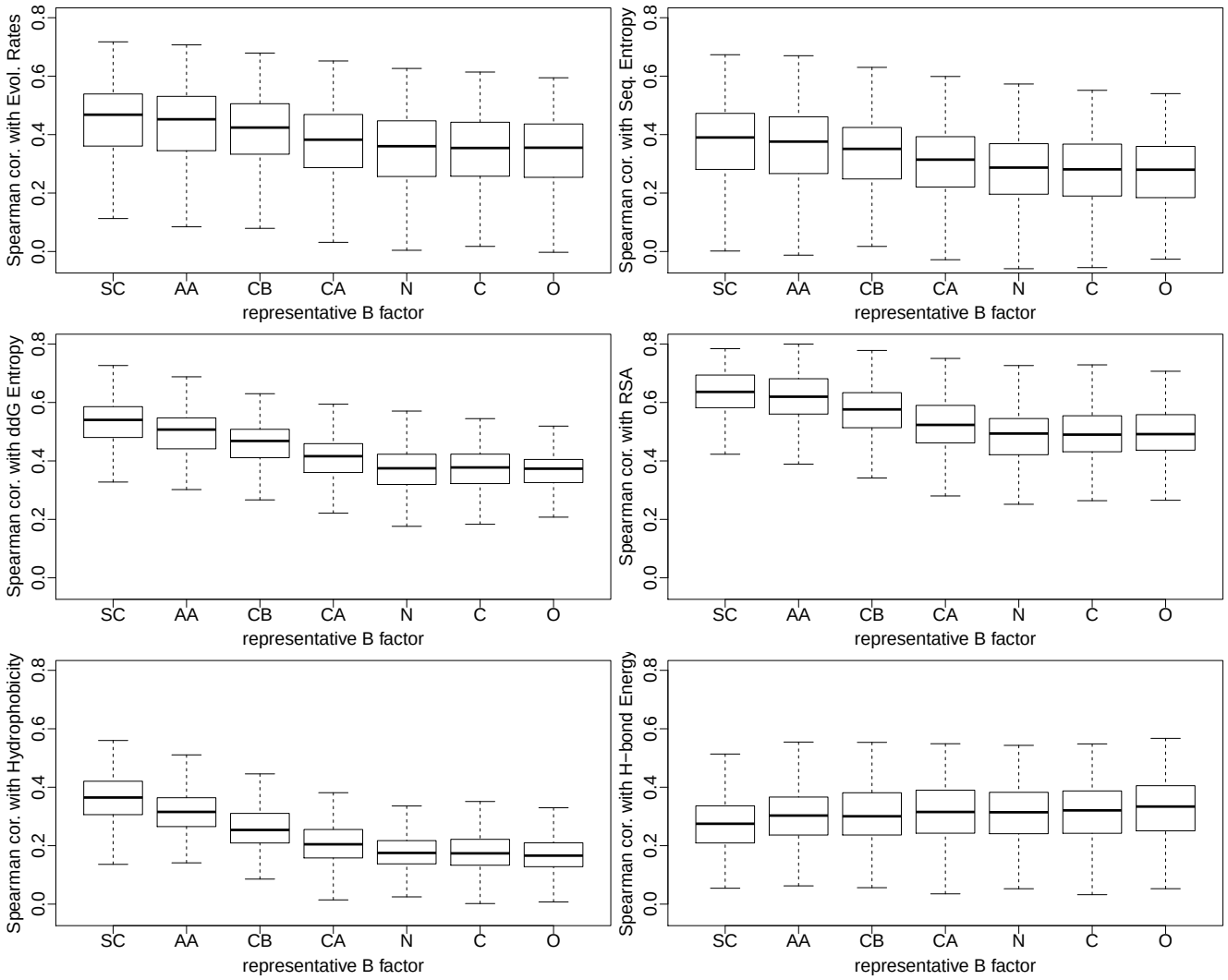

Figure 5: A comparison of the correlation strength of 6 different measures of B factor with 6 coordinate-independent structural or sequence properties for 209 proteins in dataset. Shown on the horizontal axes, are the 6 representative atomic B factors: *SC, AA, CB, CA, N, C, O* used as flexibility measures 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.

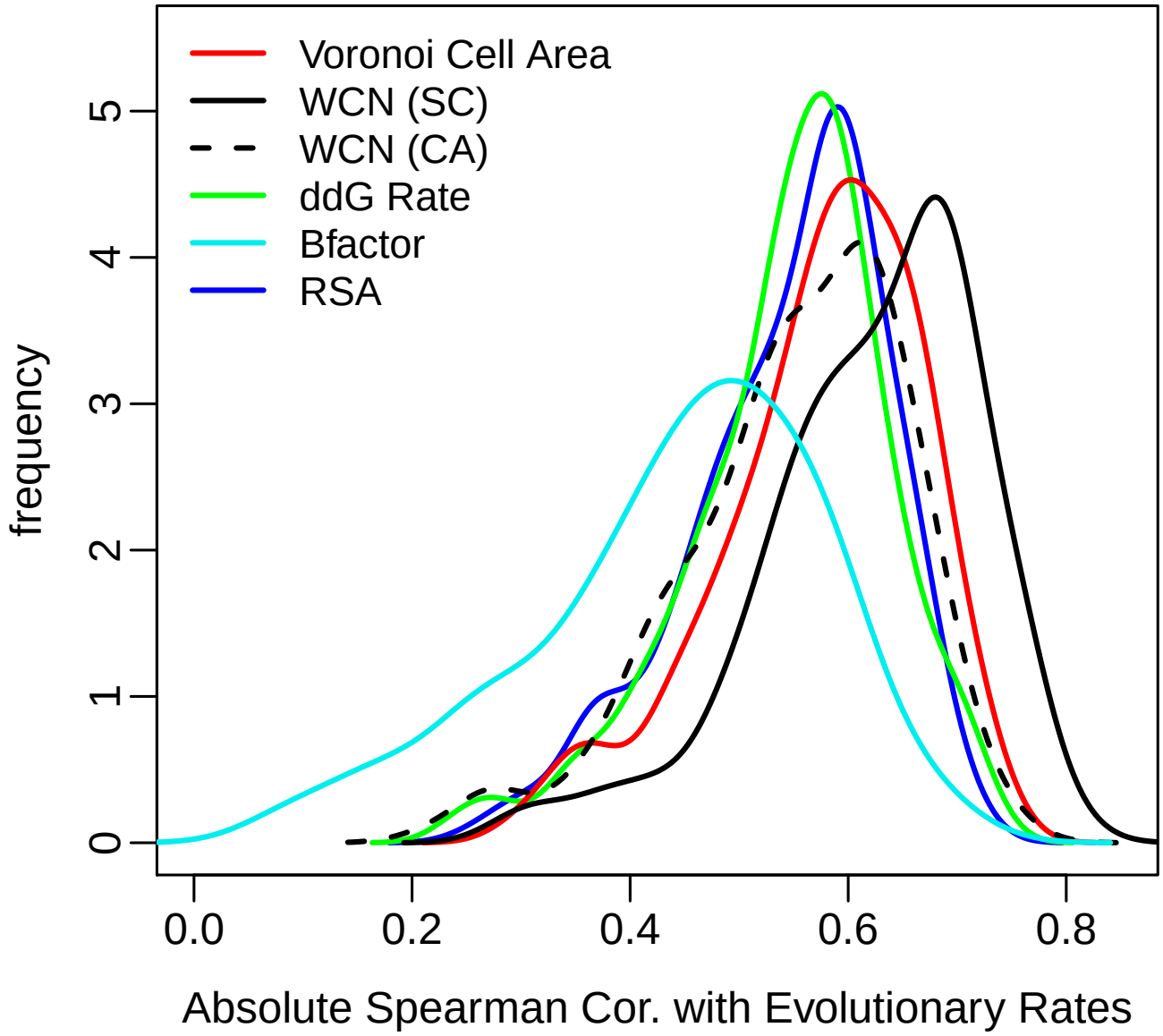

Figure 6: A comparison of the prediction power of four structural variables about site-specific evolutionary rates (ER). All structural quantities correlate positively with ER, with the exception of Weighted Contact Number (WCN) which correlates negatively. For better illustration however, the absolute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