January 29, 2015

Predicting Sequence Variability from Voronoi Tessellation of Proteins

Amir Shahmoradi[1], Claus 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*

A dataset of 209 monomeric enzymes is considered here to carry out a comprehensive search for the potential structural determinants of the site-specific evolution of protein sequence. Based on Voronoi tessellation of protein structures, we define new site-specific structural properties that are on average capable of describing up to 30% of sequence variability observed in the dataset. We show that the Voronoi cell area and volume outperform other structural proxy measures of site-specific sequence variability previously considered in the literature, such as the Relative Solvent Accessibility, Contact Number (CN), and measures of local flexibility such as the Debye-Waller factor. Using a variety of atomic coordinates in the definition and calculation of structural properties, we show that the best representative set of coordinates for individual sites in proteins is the average of the side-chain atomic coordinates. This choice of coordinates for the definition of structural properties results in the best predictions of site-specific sequence variability. Example structural properties that show significant improvement in this regard include those derived from Voronoi tessellation, contact number and representative B-factors for individual sites in proteins. On average, the choice of geometrical centers of side-chains vs. the commonly chosen coordinates of the backbone atom C_α can result in 0.07 to 0.12 improvements in sequence-structure correlation strengths. We finally argue and show that there is no uniquely-defined best kernel for the calculation of the Weighted Contact Number, which is commonly defined as the sum of the inverse-square of the reciprocal distances between pairs of sites in proteins. By contrast, other definitions perform equally well or better. This finding highlights the diverse energy landscapes of proteins and the fact that no single potential-of-mean-force can uniquely describe all interactions between individual sites in proteins.

1 Introduction

Over the past decade, extensive attempt has been made to predict protein sequence evolution from structural properties. A variety of site-specific structural characteristics have been proposed, including the Relative Solvent Accessibility (xx), Contact Number (xx), measures of thermodynamic stability changes due to mutations at individual sites in proteins (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, it is yet unknown whether these seemingly independent quantities are merely different manifestations of an underlying more fundamental characteristics of individual sites in proteins. It is perceivable that quantities such as B factor, RSA, and CN can all serve as *approximate* measures of local packing density of individual sites in proteins, or the local flexibility of individual amino acids. Recently, Huang et al 2014 (xx) argued, through an extensive mathematical model, for Contact Number as the primary determinant of sequence variability, in contrast to local flexibility. Nevertheless, 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 can be obtained from the accessible volume to each site in protein structure. An estimate of the accessible volume to each site can be obtained through Contact Number. The definition of this CN however, depends on an arbitrary parameter: the radius of neighborhood definition around each site. An alternative more accurate definition of site-specific flexibility can be obtained from partitioning of the protein 3D structure.

Motivated by this gap in the current understanding of the role of flexibility and other structural properties on sequence-structure relation in globular proteins, here we employ tessellation methods from the field of computational geometry to calculate several new site-specific quantities for globular proteins, including new measures of 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. For structural properties that are calculated based on a set of representative site coordinates, particularly Contact Number, 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. We also show that the original kernel proposed for the definition of Weighted Contact Number by (xx) and supported further by (xx) and extensively used in other studies, has no significant advantage whatsoever in predicting B factors or the sequence variability, when compared to other possible types of kernels. A discussion of the methodology used in this work, the results and implications of our findings on the energy landscape of proteins and sequence-structure relations will be presented in the following sections.

2 Methods

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. 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. 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.

In the context of protein studies, the atomic coordinates of C_α backbone atoms have been widely used as the set of centroid points to partition protein 3D structure 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 (e.g., xx), energy landscape (xx) or protein–protein interactions (xx).

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. The 209 proteins (Echave papers, Wilke ddg paper xx) were 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).

Eliminating Degeneracy in Structural Property Definitions

Depending on the choice of coordinates used, there exist degeneracies in the definition of the some site-specific variables. For example, the quantity WCN is generally calculated from the coordinates of α-carbon atoms in the 3-dimensional structure of proteins. There is however no reason to believe this set of atomic coordinates represent individual sites in proteins the best. Indeed, some earlier works have already suggested the use of center-of-mass of side chains coordinates to represent the 3D structure of protein (xx). More recently, (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 entropy. The same definition degeneracy also exists for the set of atomic B factors (xx) that are used to represent site-specific flexibility, although the popular choice of residue flexibility is α-carbon atomic B factor (e.g., Halle 2001 xx). Similar to WCN and Bfactor, there is also ambiguity as to which set of residue atomic coordinates best represent individual sites in proteins for the calculation of Voronoi cells.

In order to identify which set of atomic coordinates result in the highest prediction power for WCN, Bfactor, and Voronoi cells, here we calculate and consider all possible definitions of these variables based on different choices of the representative set of atomic coordinates used. 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 atoms had not been not resolved in the PDB file or the amino acid lacks the heavy atom needed (e.g., C_β for Glycine). The coordinate for that specific site is replaced with the coordinate of the corresponding C_α atom in the amino acid backbone.

We use VORO++ software (xx) to calculate the relevant Voronoi cell properties of all sites in all proteins, and use DSSP (xx) for the calculation of 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.

Contact Number Definition

In its simplest mathematical form, the Contact Number (CN) for a given site in protein is defined as the number of amino acids within a fixed radius of neighborhood around it (xx). Individual sites are generally represented by the coordinates of C_α backbone atoms in the calculation of CN. A major problem with the traditional definition of contact number, is the existence of an arbitrary parameter – the radius of neighborhood – in the definition of CN. There is no consensus on the optimal value of this cutoff distance, typically ranging from $7\mathring{A}$ to $13\mathring{A}$ (e.g., lin, franzosa).

Several The statistical kernel often used in the definition of Weighted Contact Number is generally the square of the reciprocal distance between the contacting pair of sites in the protein.

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`.

Dataset

The results presented in this work are based on a dataset of 209 monomeric enzymes (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 evolutionary rates for each individual protein in dataset. To do so, we relied on two independent methods of measuring sequence variability measure. First, we calculated the Shannon entropy (H_i) – the sequence entropy, hereafter abbreviated as *seqent* – at each alignment column i, based on the assumption that the occurrence of each of the 20 amino acids is equally likely at any given site in the alignments:

$$H_i = -\sum_j P_{ij} \ln P_{ij} \tag{1}$$

where P_{ij} is the relative frequency of amino acid j at position i in the alignment. Alternatively, we also calculated a measure of site-specific evolutionary rate – hereafter abbreviated as *r4s* – for each protein using software rate4site (xx). To do so, first the Maximum Likelihood phylogenetic trees were inferred with RAxML, using the LG substitution matrix and the CAT model of rate heterogeneity (Stamatakis 2014). For each structure, we then used the respective sequence alignment and phylogenetic tree to infer site-specific substitution rates with Rate4Site, using the empirical Bayesian method and the amino-acid Jukes-Cantor mutational model (aaJC) (Mayrose et al. 2004).

Structural Properties

The goal of the presented work is to identify the prominent structural or evolutionary properties of proteins that modulate sequence-structure correlations. These potential modulators represent a unique characteristics of the protein as a whole. In general, the structural and evolutionary properties fall into two major categories. 1. *Residue-level properties*: Site-specific structural or evolutionary properties that are defined and calculated for each specific amino acid site in the protein sequence. Prominent examples of the site-specific structural properties include RSA (Tien et al. 2012 xx), WCN (shih? xx). 2. *PDB-level properties*: structural or evolutionary characteristics that are representative of the protein as a whole. Examples include pdb Contact Order (CO) as defined by xx, protein size and compactness, sequence length, structural resolution of the protein in X-ray crystallography. In addition, the distribution of each residue-level property can be summarized by its statistical moments as pdb-level property of the protein. Prime examples include, the mean and variance of WCN, RSA, sequence entropy, evolutionary rates. A comprehensive list of protein properties and their definitions are given in Table **??**.

3 Results

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

As explained in previous section, there is a high level of redundancy in the initial set of collected protein properties. In particular, depending on the set of atomic coordinates used, there are 7 different measures for some residue characteristics such as the residue Weighted Contact Number, Bfactor and Voronoi cell properties. This in turn results in a large set of secondary variables at pdb-level that basically measure the same protein characteristics, but with different strengths. 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 2).

For the measure of local flexibility in proteins (Bfactor) we similarly find that among all 7 representative measures of site Bfactors, the average of Bfactor 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 Bfactor (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 3).

Similar to WCN and Bfactor, the Voronoi cell properties, most importantly the cell surface area, volume and the cell compactness also correlate best with other structure and sequence properties, only if the average side chain coordinates are used as the seeds of Voronoi cells (Figure 4).

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.

Sequence divergence as the main Determinant of Sequence-Structure Relation

In order to identify the potential contributing factors to the strength of sequence–structure correlations, we first employ one of the simplest nonparametric yet powerful tests of statistical dependence, that is, we construct the Spearman correlation matrix of all pdb-level structure and sequence properties. The choice of Spearman versus the popular Pearson's correlation measure is made in order to minimize the effects of any nonlinear variable relationships on the strengths of the correlations. The resulting correlation matrix reveals a myriad of pdb-level properties each having a small but nonzero contribution to the strength of the structure-sequence correlations.

A hierarchical clustering of the correlation matrix however, reveals two main independent factors that have the strongest influence on the strengths of sequence-structure correlations: 1. The sequence divergence as measured by the standard deviation of sequence entropy and evolutionary rates (denoted by *sd.seqent* & *sd.r4sJC*) among all sites in each protein structure. 2. The homogeneity of the hydrogen bond strengths among the back bone atoms of each protein structure, as measured by the standard deviation of hydrogen bond energies (denoted by *sd.hbe*) among all pdb sites. A reduced-size of the Spearman correlation matrix for the most influential factors on the two strongest sequence–structure (seqent/r4s – wcnSC/varea) relations is illustrated in Figure **??**.

For the other weaker sequence–structure relations, i.e. the correlations of seqent/r4s with RSA, $\Delta\Delta G$ entropy (ddgent), and Bfactor (bfSC) we find other pdb-level properties that also contribute to the correlation strengths, comparable to or even stronger than in sequence divergence and hydrogen bond homogeneity. In general, we observe that for the weaker the sequence-structure correlations, factors that determine the accuracy of the measured residue properties become more influential on the strength of the correlations. In particular, the X-ray crystallographic resolution of the structure and the definition of the $\Delta\Delta G$ entropy play dominant roles, with Spearman correlation coefficients of ρ 0.3, on the strengths of the corresponding sequence-structure relations.

To ensure the accuracy of the results obtained from the Spearman correlation matrix of the pdb-level properties, we also use multivariate linear regression models, with individual sequence-structure correlations as the sole regressand of the regression models, and the set of pdb-level properties as the explanatory variables. Since the number of explanatory variables is comparable to the number of observations (i.e., the number pdb structures in the dataset), we use regularized regression (reference to R package xx) on the entire dataset, and also on the rank transformation of the dataset in order to minimize the effects of potential nonlinearities in data. Depending on value of the free parameter α, this generalized regression model is a compromise between *ridge regression* – which attempts to shrink the coefficients of correlated predictors towards each other – and *lasso regression* – which tends to pick one of the correlated predictors and discard the rest. In addition to regularized regression, we have also employed Principal Component Regression (PCR) on the original dataset and its rank transformation. Both regression methods, PCR & regularized, point to similar set of pdb-level properties as the strongest determinants of sequence-structure correlations.

4 Discussion and Concluding Remarks

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 Voronoi cell sphericity defined as,

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

in which V & A represent *vvolume* & *varea* respectively. 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

....

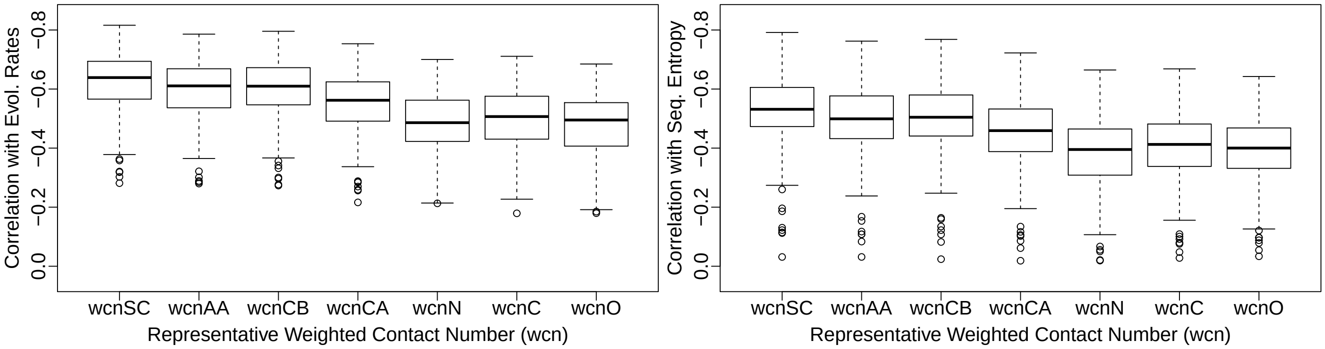

Figure 1: A comparison of the predictive power of different measures of Weighted Contact Number (WCN) about different structure or sequence properties for 209 proteins in dataset. In each plot, the Spearman correlation strengths of a given structure or sequence property on the vertical axis with different measures of WCN on the horizontal axis are compared against each other. The capital letters in the variable names on the horizontal axis denote the set of atomic coordinates used to calculate WCN. The variable *wcnSC* denotes WCN measure based on the average coordinates over all heavy atoms in the side chain and the variable *wcnAA* denotes WCN measure based on the average coordinate over all heavy atoms in the side chain and backbone of the amino acid.

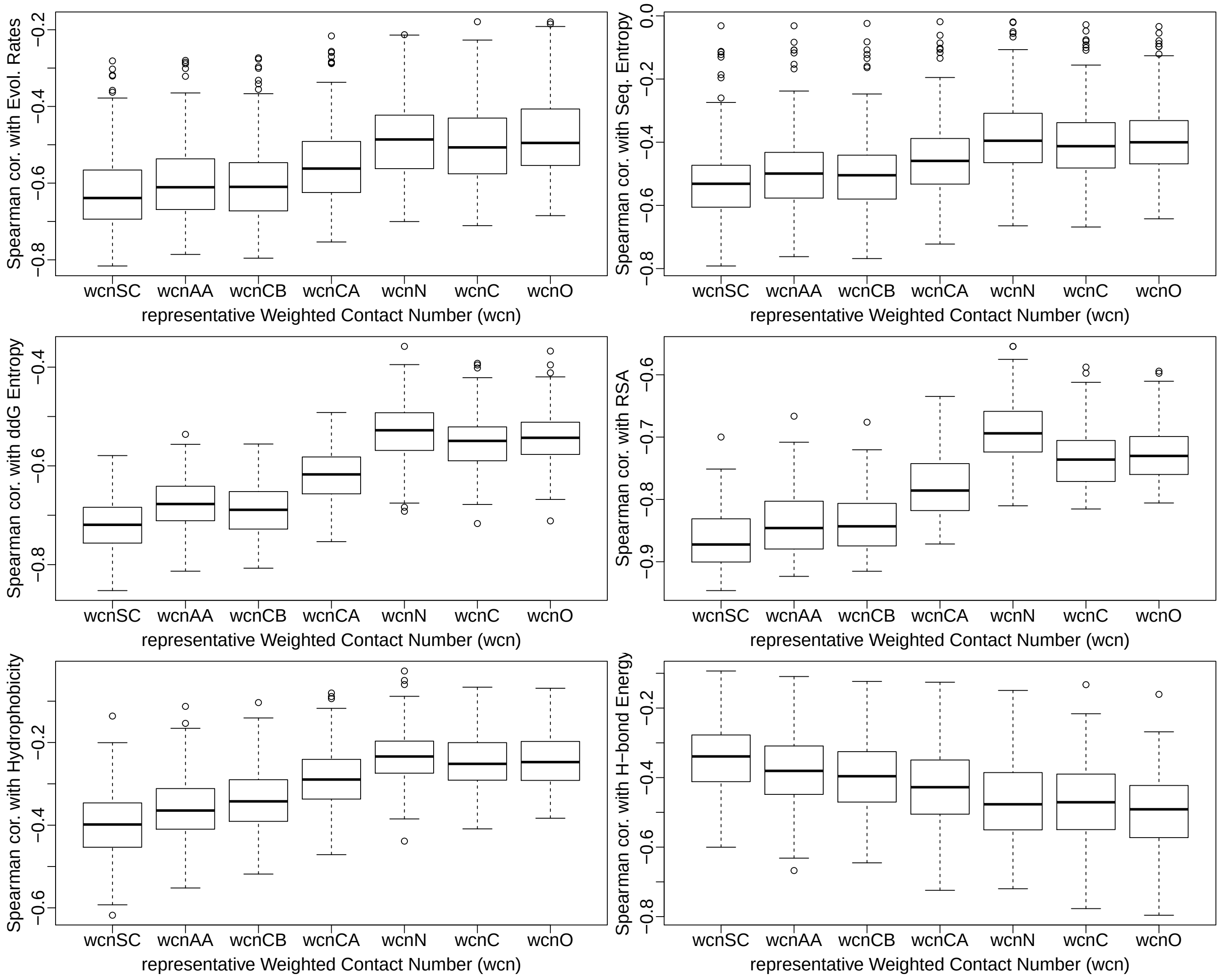

Figure 2: A comparison of the predictive power of different measures of Weighted Contact Number (WCN) about different structure or sequence properties for 209 proteins in dataset. In each plot, the Spearman correlation strengths of a given structure or sequence property on the vertical axis with different measures of WCN on the horizontal axis are compared against each other. The capital letters in the variable names on the horizontal axis denote the set of atomic coordinates used to calculate WCN. The variable *wcnSC* denotes WCN measure based on the average coordinates over all heavy atoms in the side chain and the variable *wcnAA* denotes WCN measure based on the average coordinate over all heavy atoms in the side chain and backbone of the amino acid.

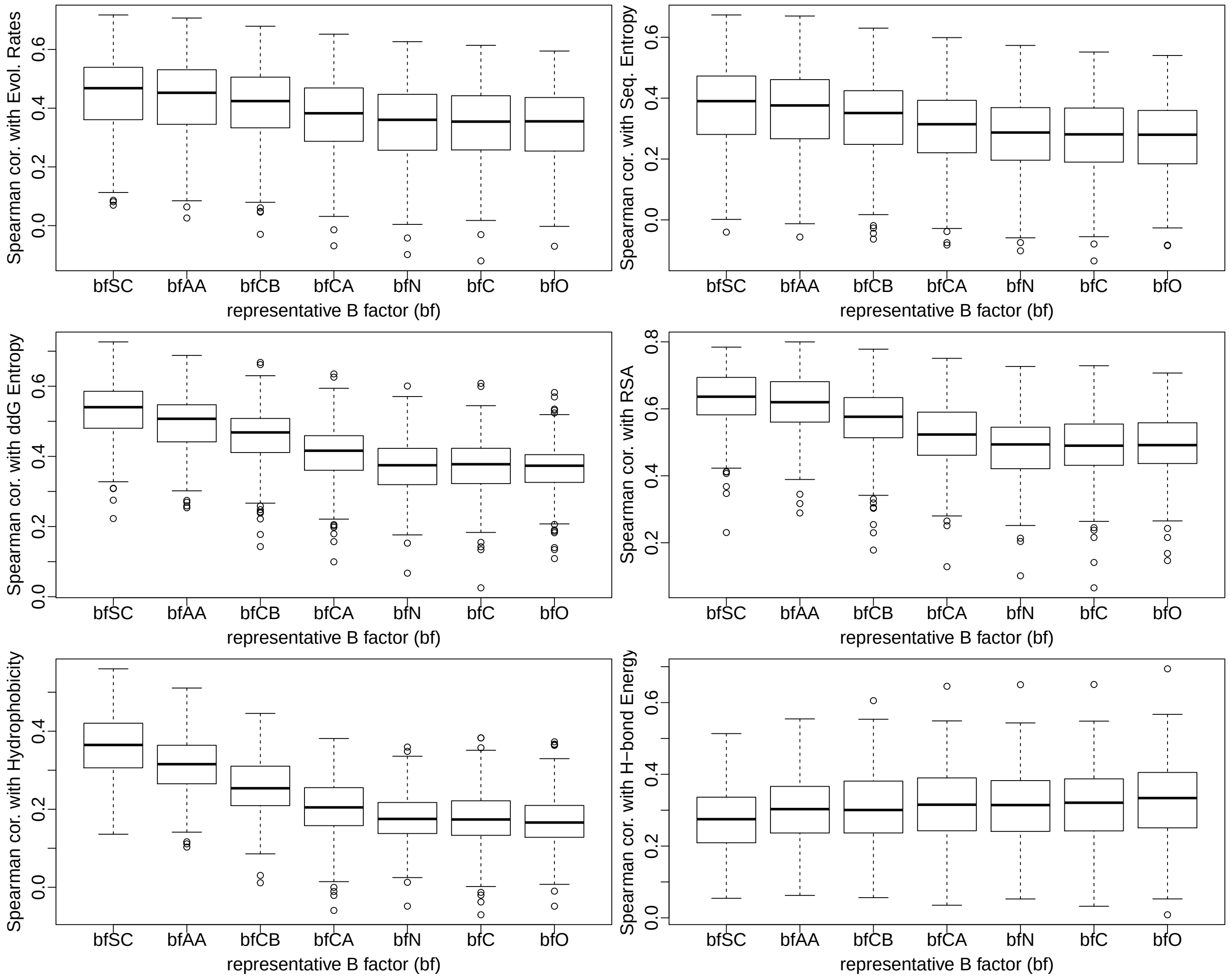

Figure 3: A comparison of the predictive power of different measures of Bfactor (BF) with different structure or sequence properties for 209 proteins in dataset. In each plot, the Spearman correlation strengths of a given structure or sequence property on the vertical axis with different measures of atomic Bfactors on the horizontal axis are compared against each other. The capital letters in the variable names on the horizontal axis denote the representative atomic Bfactors. The variable $bfSC$ denotes the average coordinate over all heavy atoms in the side chain for each site in protein and the variable $bfAA$ denotes average coordinate over all heavy atoms in the side chain and backbone of the amino acid.

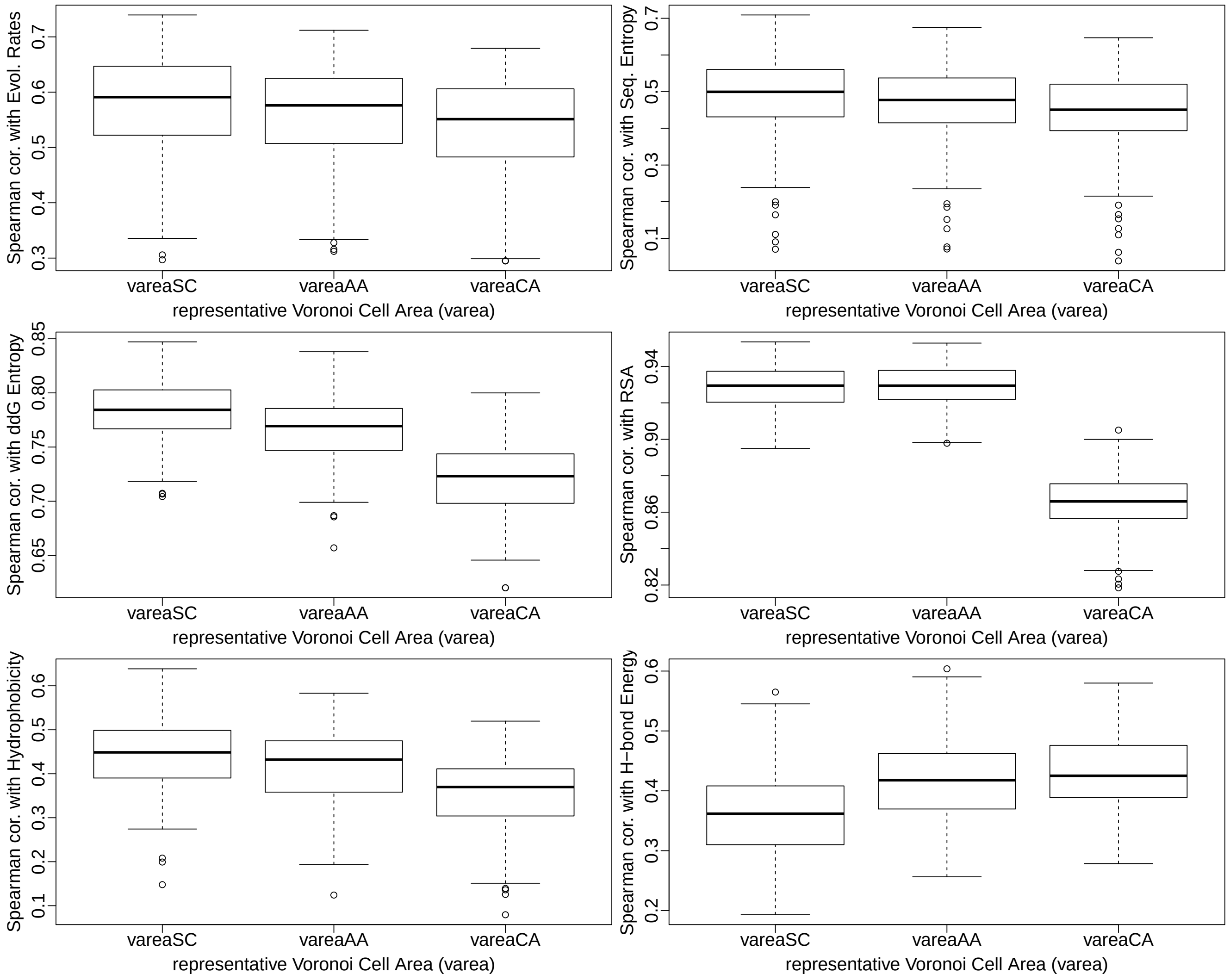

Figure 4: Should be likely modified to volume-area-length-sphericity measures instead.