Physiochemical property based approach for protein sequence analysis

Pradeep Bhadola\textsuperscript{1} and Nivedita Deo\textsuperscript{2}

\textsuperscript{1}The Institute for Fundamental Study, Naresuan University, Phitsanulok, Thailand
\textsuperscript{2}Department of Physics and Astrophysics, University of Delhi, India

E-mail: \textsuperscript{1}bhadola.pradeep@yahoo.co.in, \textsuperscript{2}ndeo@physics.du.ac.in

Abstract. We propose a physiochemical property based analysis, that represent a protein sequence as a multidimensional time-series from which residue positions controlling protein function can be extracted. We observe that the favorable substitutions at a position are the ones which preserve the property crucial for functioning of that site which is quantified by the entropy and Kullback-Leibler divergence. The entropic measures shows that during the evolutionary history of the protein family, it is the certain physiochemical properties that are conserved rather than the type of amino acids. For each physiochemical property, the correlation matrix between positions is calculated, and using an ensemble of Wishart matrices from the random matrix theory for the noise estimation and information filtering. The spectral properties of correlation matrices are calculated and compared with the analytical results for the Wishart matrices.

1. Introduction

The protein sequence space is growing at an exponential rate, it is extremely important to develop efficient and accurate techniques and tools to draw inferences from the sequence data. Even with existing theoretical models and computational techniques the function and structure encoded by a protein sequence is not completely established \cite{1-4}. Most of the existing methods mainly treat protein sequences as strings of characters and analyze the sequence based on the frequencies of amino acids (will be referred as string method) by estimating entropy or mutual information (MI) \cite{1, 2, 4}, without looking at the chemical, physical or biological properties of amino acid. There are studies that show that the evolution and functional properties greatly depends on the physiochemical properties, by influencing the mutation rate or determining the native state \cite{5}. To address the above problem we propose a physiochemical property based analysis that represent a protein sequence as a multidimensional time series, where each dimension represents a property value. The approach is mainly based on two conjectures: (1) structurally and functionally important sites show a high degree of conservation, if not in amino acid type, it will be reflected in the amino acid physiochemical properties, which is the outcome of the evolutionary constraint. (2) Functionality is a manifestation of structural interactions (direct or indirect mediated by series of intermediate amino acids) between positions, which is a result of the physical, chemical and biological coupling. The work involves information theorectic measure followed by the tools from the random matrix theory (RMT) which has been very successful in application to diverse areas of research including nuclear physics, disordered systems, information theory, string theory, statistical physics, biological sciences, finance, social systems \cite{3, 6}. RMT is used as a null hypothesis to to isolate non-random information from statistical and phylogenetic noise.
2. Data and Method

The system under study is class A β-lactamase enzyme family (Interpro entry IPR000871 consisting of 5447 proteins) which are the enzymes secreted by bacteria in response to β-lactam antibiotics like penicillin [7, 8]. These sequence are first filtered and then aligned in a multiple sequence alignment. Different physiochemical properties of amino acid are extracted from various sources and then rescaled between $-1$ to $1$. Each amino acid in MSA is replaced by the physiochemical properties to get a 3 dimensional matrix $D_{\alpha,s,i}$, where $s$ represents the sequence in MSA, $i$ represents the amino acid position, and $\alpha$ represents the property (see [3] for details). We define a sequence vector as $U_{\alpha,s}(i) = \sum_{k=1}^{\infty} D_{\alpha,s,k}$ to see the trend in the change of values of a property across a sequence. The plot of sequence vector for 4 different protein sequence is using hydrophobicity and polarity is show in figure 1(a). The evolutionary divergence between two protein sequence can be estimated by comparing the two sequence vector using any metric distance measure. This evolutionary divergence score not only depends on the type of amino acid but also takes the physical, chemical and biological properties into account unlike the string based method which relies solely on the amino acid type.

3. Entropy

The diversity of amino acid at a position for a given family is measured by the Boltzmann-Shannon entropy defined as

$$E(i) = -\sum_{j=1}^{n} P_j(i) \log_n (P_j(i))$$

The Boltzmann-Shannon entropy quantifies the expected value of the information contained at position $i$ with the probability distribution $P_j(i)$ as the frequency of occurrence ($j = 1 \cdots n$ with $j$ labeling the type of amino acids and $n = 20$ the total number of distinct residues). For a fixed property $\alpha$, the frequency...
distribution of physiochemical values is constructed using $m$ equal sized bins in the interval $[-1,1]$ and binning the physiochemical property value (column of data matrix $D^\alpha$) into them ($m=21$ for current analysis). The normalized probability distribution ($P^\alpha_b$) at each column $j$ is estimated from this frequency distribution (normalized). $P^\alpha_b$ gives the probability that a amino acid at a given position $j$ has a value in the specified bin for a given property $\alpha$. The entropy at position $i$ considering property $\alpha$ with normalized frequency distribution $P^\alpha_b(i)$ is given by

$$E^\alpha(i) = -\sum_{b=1}^{m} P^\alpha_b(i) \log_m(P^\alpha_b(i))$$

The string based entropy $[E(i)]$ results in larger values for less conserved positions (large variations) as changes do not take into account the similarity of some of the amino acids imposing a severe limitation to the method. The amino acids with identical values for a given property will be placed in the same bin for property based entropy $[E^\alpha(i)]$ even the type of amino acid is different. This sets a range for which such substitutions can be treated as identical by taking the similarity between amino acid with respect to the property into account. The estimated entropy is lower, with increased accuracy for the property based approach as compared to string based approach as shown in figure 1(b) which gives the difference between the values of hydrophobicity and string based entropy ($E^\alpha(i) - E(i)$) where $\alpha$ is hydrophobicity. For all position the hydrophobicity based entropy is lower implying property conservation is more than the amino acid type conservation. For the given enzyme family the analysis is robust if the number of sequence exceeds 200. The entropy estimate is better if the number of bins are close to 20, increasing the number of bins ($\approx 50$) makes the property based entropy similar to the string based entropy. Th positions with the lowest entropy for hydrophobicity figure 1(c) estimated as (38, 41, 97, 98, 99, 199, 200, 202), are the conserved motifs for the family. These are most functional positions in the protein family which are the part of the active and catalytic sites as specified by the experimental studies [9, 10].

4. Kullback-Leibler (KL) Divergence

The information content of a position is quantified by the Kullback-Leibler (KL) divergence or relative entropy defined as the divergence of a probability distribution at a position from the background distribution. The normalized probability distribution ($P^\alpha_j$) for the property is used and compared with the background probability to calculate the divergence in the properties values. The background probability estimated in [4] (that is the mean frequency of amino acids constructed from all proteins) is used to construct a background column of amino acids having frequencies equivalent to the background frequencies of occurrence. Then substituting each amino acid by the rescaled physiochemical property value and using the same binning procedure as discussed above (in the Entropy section) for the background column, the background property based distribution $Q^\alpha$ is estimated. The Kullback-Leibler (KL) divergence or relative entropy between two probability distribution $P^\alpha_j$ and $Q^\alpha$ is defined as

$$K(P^\alpha_j || Q^\alpha) = \sum_{i=1}^{n} P^\alpha_j(i) \log_a \left( \frac{P^\alpha_j(i)}{Q^\alpha(i)} \right)$$

The figure 1(d) shows the KL divergence score for the hydrophobicity and polarity for the betalactamase family where some positions show high divergence for hydrophobicity while other positions for polarity. We propose that each position is characterized by one or more physiochemical property, i.e. the property which has a high divergence score at a given position is responsible for the function of that position. Larger the KL divergence value for a property the more important is the property at that position for functioning. The divergence score can be used to theoretically and numerically assign properties responsible for the normal working of the positions.
Figure 2. (a) Distribution of correlation coefficient for different physiochemical properties. (b) Comparison of the Eigenvalue distribution for hydrophobicity with RMT bounds (red color). Insets show eigenvalues outside the lower bound.

5. Correlation Matrix and eigenvalues distribution

Pearson’s correlation coefficient is used to calculate the correlations between different positions of the data matrix $D^\alpha$ for a given property (fixing $\alpha$). This results in a three dimensional correlation matrix $C^\alpha$, which for fixed property gives cross correlation between positions in the MSA. The figure 2(a) compares the distribution of the elements of correlation matrix for different physiochemical properties, with the random correlation. The noise in the system implies the presence of unimportant mutation or substitution in the course of evolution leading to the random connection between positions which may not have any functional importance. A null model is constructed by shuffling each column of the data matrix to break existing correlations. In current analysis random shuffling is numerically equivalent to the analytical results for the Wishart matrices. For the Wishart matrices the density function for the eigenvalues is given by Marcenko-Pastur distribution [11], where the the upper and lower bounds for the eigenvalue $\lambda$ are defined as $\lambda_{\pm} = \sigma^2 \left( 1 + \frac{1}{Q} \pm 2 \sqrt{\frac{1}{Q}} \right)$ with $Q = \frac{S}{L} \geq 1$, where $S$ is the sequences and $L$ is the positions in MSA. Giving $\lambda_+ = 2.775$ and $\lambda_- = 0.111$, which is applicable for all properties and are taken as benchmarks for finding eigenvalues of the system having significant information. Eigenvalues deviating from the RMT bounds on both sides of the spectrum contain significant information about the system. The eigenvalue distribution for the hydrophobicity based correlation matrix is show in figure 2(b), red line indicates the RMT bounds. There are a total of 44 eigenvalues out of 204, that are outside the RMT bounds, (14 eigenvalues $> \lambda_+$ and 30 eigenvalues $< \lambda_-$) The number of eigenvalues on the lower side of the spectra which are outside the RMT bounds are higher in number as compared to the eigenvalues on the higher side of the spectrum. Analyzing the eigenvectors corresponding eigenvalues on the lower side of spectra it is observed that components with the largest components (found from the small eigenvectors) forms a highly interacting set of residues This set has a well defined collective structural and functional role. For example, for hydrophobicity, conserved motifs are given by smallest eigenvector (EV1), Catalytic and ligand binding by EV2 and Boundary of the active site by EV4 [3]. Each property gives additional information about the family for example the analysis based on Polarity we find one eigenvector which corresponds to position that take part in characterization of substrate specificity and mutational stability for the $\beta$-lactamase family. For each property the eigenvector of the correlation matrix have different interpretations and information.

6. Conclusion

A method to represent protein sequences based on the physiochemical properties is devised. Using this method, each protein sequence in MSA is represented as a multidimensional time series, where evolutionary divergence between sequences can be estimated taking into account for the physiochemical
properties of amino acids. The entropic measure shows that at functional position within the protein sequence some of the physiochemical properties are more conserved than the type of amino acid. The entropic and the KL divergence based on the physiochemical properties identifies positions that have important structural and functional role. The correlation based on the physiochemical properties, are calculated and compared with the random system. The eigenvalue distribution of the system is compared with null model and analytic results of the RMT. The Comparison reveals that the eigenvalues on both sides of the spectrum deviates significantly from the RMT results. The eigenvalues outside the RMT bound contains useful information. When the eigenvector corresponding to these eigenvalues are analyzed it was found that eigenvector on lower side are more localized and informative, which can group positions into a collective functional role. The eigenvalue spectra of different properties uncovers different set of information.

Acknowledgments
We acknowledge the Department of Science & Technology (DST), India, (SERB-DST No-EMR/2016/006536) for financial support.

References
[1] Chakraborty A and Chakrabarti S 2014 Brief. Bioinfo. 16(1) 71 – 88.
[2] Radivojac P, Clark W T, Oron T R, Schnoes A M, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A and Pandey G 2013 Nat. methods 10(3) p.221. ; Zakeri P, Jeuris B, Vandebriel R and Moreau Y 2014 Bioinfo. 30 1850 – 1857 (2014).
[3] Bhadola P and Deo N 2016 Phys. Rev. E 94(4) 042409.
[4] Halabi N, Rivoire O, Leibler S and Ranganathan R 2009 Cell 138(4) 774.
[5] Xia X and Li W H 1998 J. Mol. Evol. 47 (5) 557; Yang Z, Nielsen R and Hasegawa M 1998 Mol. Biol. Evol. 15(12) 1600; Anfinsen C 1973 Science 181 223.
[6] Efetov K Supersymmetry in Disorder and Chaos (Cambridge University Press 1997 Cambridge); Guhr T, Muller A and Weidenmuller H A, 1998 Phys. Rev. Lett. 83 1467; Bhadola P and Deo N 2013, Phys. Rev. E 88(3) 032706; Bhadola P, Garg I and Deo N 2013 Nucl. Phys. B 870 384.
[7] Freer J M 1995 Mol. Microbiol. 16 385; Massova I and Mobashery S 1998 Antimicrob. Agents Chemother. 42 (1) 1 – 17.
[8] Bush K and Fisher J F 2011 Ann. Rev. Microbio. 65 455 – 478.
[9] Brasseur J L, Dive G, Dideberg O, Charlier P, Frere J M and Ghysens J M 1991 Biochem. J. 279 213.
[10] Hermann J C, Ridder L, Mulholland A H and Holtje H D 2003 J. AM. CHEM. SOC. 125 9590.
[11] Bowick M J and Brezin E 1991 Phys. Lett. B 268 21.; Feinberg J and Zee A 1997, J. Stat.Phys. 87 473.