Random Matrix Analysis of Protein Families

ECS Transactions, 107 (1) 18877-18891 (2022)
10.1149/10701.18877ecst ©The Electrochemical Society

Random Matrix Analysis of Protein Families

Rakhi Kumari[a], Pradeep Bhadola[b] and Nivedita Deo[a]
[a] Department of Physics & Astrophysics, University of Delhi, Delhi 110007, India
[b] Centre for Theoretical Physics & Natural Philosophy "Nakhonsawan Studiorum for Advanced Studies", Mahidol University, Nakhonsawan Campus, Nakhonsawan, 60130, Thailand.

Proteins are vital for almost all biochemical and cellular processes. Although there is an enormous growth in the protein sequence data, the statistical characterization, structure and function of many of these sequences are still unknown. The statistical and spectral analysis of the Pearson correlation matrices between positions based on physiochemical properties of amino acids of seven protein families is performed and compared with the random Wishart matrix model results. A detailed analysis shows that the protein families significantly diverge from the Marchenko-Pastur distribution with many eigenvalues (outliers) outside the Wishart lower and upper bound. It is shown that level spacing distribution of protein families is similar to the Gaussian orthogonal ensemble. Further, the number variance varies as log of the system size indicating the presence of long range correlations within the protein families.

1. Introduction

In molecular biology, there is one fundamental question that is how does an amino acid sequence determine the biological, functional, and structural properties of the protein sequence. In this context, we study the statistical and spectral properties of the cross-correlations between positions based on the physiochemical properties of amino acids. This study is a large-scale statistical analysis based on random matrix theory which uses physio-chemical properties-based correlation matrices of multiple protein families downloaded from the PFAM (1) database.

Random matrix theory proved its importance in diverse fields with a wide range of applications ranging from nuclear physics (2), biological physics (3,4), the stock market (5), and many more. We find that the results from the random matrix theory on Wishart matrices are of prime importance to understanding the statistical structure of the interaction and correlations between positions in the protein families. The properties of physio-chemical-based correlation matrices are compared with the analytical results of Wishart matrices (6,7). A Wishart matrix is computed as W=DD†, where D is a random matrix (dimension N x M,) with entries following a Gaussian distribution that has zero mean and unit variance. D† is the complex conjugate of D.

For each family, protein sequences are processed, filtered, and then create the multiple sequence alignment (MSA). The two-dimensional MSA is converted into a three-dimensional data matrix, by using the physio-chemical properties of amino acids in which the amino acid is replaced by its physiochemical property value. The physio-chemical property is represented by the third dimension of the data matrix. Firstly we study the statistical properties of independent off-diagonal elements of

the data correlation matrix and compare the results with Wishart matrices. We plot and analyze the probability density function (PDFs) of the system correlations and Wishart matrices. We find that for most of the protein families, the distribution differs significantly from Wishart on the upper as well as on the lower side, indicating the presence of high correlation and anti-correlations, which are the result of evolutionary coupling and natural selection. These correlations are not of random origin. The dependence of correlation between different physiochemical properties was checked as they were derived from the same data (MSA). We find the structure of correlation for a given protein family calculated using properties shows less similarity with other property indicating that each property reveals new information.

The eigenvalue distributions of protein families for all properties were compared with the analytical results for Wishart matrices. Wishart matrices are studied in great detail with the spectral eigenvalue density function which follows the Marchenko-Pastur distribution with well-defined upper and lower bounds on the eigenvalues. Protein families show significant divergence from the Marchenko-Pastur distribution with many eigenvalues outside the Wishart lower and upper bound. These eigenvalues contain significant information about the system.

Next, we study the nearest neighbor eigenvalue spacing distribution. We find that the presence of short-range correlations between eigenvalues for most of the protein families are in agreement with the level spacing of Gaussian orthogonal ensemble and represent a universal feature. To check for the long-range correlation between the eigenvalues we calculate the number variance for protein families for all properties. The analysis shows that for proteins the number variance varies approximately as log (L), where L is system size (number of positions in MSA). The variation of the number variance as log of the system size indicates the presence of long-range correlations within the system.

2. System and Data

For the analysis, we use 7 different protein families. The multiple sequence alignment (MSA) was downloaded from PFAM. The details of the protein families used are as follows:

1) **Mitochondrial Protein (PF00153):** Mitochondrial proteins resides within the mitochondria of cells. They are responsible carrying out reactions of the electron transport chain. The MSA of this protein family consists of 160 sequences each of length 101.

2) **Expanded EBP (EXPERA) Protein (PF05241):** EBP enzyme catalyzes the transposition of a double bond in the sterol B-ring. The MSA consists of 675 sequences each of 134 positions.

3) **CAP-Gly Protein (PF01302):** Cytoskeleton-associated proteins (CAPs) helps in transportation of vesicles and organelle along the cytoskeletal network as well as helps in the organization of microtubules. The MSA consists of 593 sequences each of 83 positions.

4) **Cadherin 4 Protein (PF17803):** Cadherin 4 Protein has many bacterial-like domains those are part of extracellular proteins. The MSA consists of 694 sequences with 77 positions.

5) **IPD Protein (PF00475)**: IPD enzyme involves in the cutting of the carbon-oxygen bond which participates in the histidine metabolism. The MSA consists of 487 sequences with 151 positions.

6) **Histidine Kinase Protein (PF02518)**: Histidine Kinase Proteins are the part of several ATP-binding proteins. The MSA consists of 685 sequences with 155 positions.

7) **HAM1 Protein (PF01725):** HAM1 family controls 6-N-hydroxylaminopurine in S. cerevisiae. It also protects the cell from HAP, either on the level of deoxynucleoside triphosphate or the DNA level. The MSA consists of 1061 sequences with 208 positions.

The multiple sequence alignment (MSA) of each family is downloaded from the PFAM database. The MSA is first processed and then transformed into a 3-dimensional data matrix using different physio-chemical properties. In the current article, we have used 4 different physiochemical properties namely hydrophobicity, polarity, volume, polarizability. The physio-chemical properties of each amino acid are downloaded and then we rescaled the values between -1 to 1, where 1 is the maximum value and -1 is the minimum value of a given physiochemical property. The rescaled values of amino acids are substituted in the MSA, which results in a 3-dimensional data matrix. The details of the process are given in (8,9). For each family we obtain a 3-dimensional data matrix $D^{\gamma}_{s,i}$, where i represents the position (column), s represents the sequence (row) in the MSA, and γ indicates the physiochemical property under consideration.

Methods

Sequence Vector

Using the data matrix, we create a sequence vector to check the trends in the property values. For each protein sequence we create sequence vector W^{γ}_{s} is defined as

$$W^{\gamma}_{s}(i) = \sum_{m=1}^{i} D^{\gamma}_{s,m} \qquad [1]$$

The sequence vector W^{γ}_{s} , is the cumulative sum of the properties values of all previous amino acids. The sequence vector is a graphical way to visualize and compare the protein sequences. The sequence vector representation depends on the physio-chemical properties with different properties resulting in different representations of the same protein sequence.

The sequence vector is also helpful in determining the key properties responsible for the function and working of a given protein family. The underlying argument to identify the important physicochemical properties is that since all sequence belongs to the same family, they have an identical function, therefore the properties responsible for the function of the family will show an identical trend. Since all sequences are different at the sequence level (amino acids), and they show similar trends in the given physio-chemical property, this implies that change in the amino acid type does not make significant changes in the sequence vector. This similarity in the trend of sequence vector indicates that the property is conserved for the family and may be responsible for its structure and function. We analyze the sequence vectors for all the seven families and the results are discussed in the Results section.

Correlation Matrix

Using the data matrix, we construct the correlation matrix for each physio-chemical property and for each family. The element of the correlation matrix is defined as the Pearson's correlation coefficient between the positions in the MSA for a given property. For a property γ, the correlation is calculated between columns of the data matrix D^{γ}, which results in a correlation matrix C^{γ} given by

$$C_{i,j}^{\gamma} = \frac{< \left(d_i^{\gamma} - < d_i^{\gamma} >\right) \left(d_j^{\gamma} - < d_j^{\gamma} >\right) >}{\sigma_{d_i^{\gamma}} \sigma_{d_j^{\gamma}}} \qquad [2]$$

where the i[th] column of the data matrix D^{γ} is given by d_i^{γ} with $\sigma_{d_i^{\gamma}}$ as the standard deviation. The average ($< \cdot \cdot \cdot >$) is defined over the sequences. The correlation matrix is calculated for all properties, and can be represented as a 3-dimensional matrix C^{γ}, here the third dimension is the physiochemical property dimension.

The correlation matrix depends on the coevolution of the physiochemical properties. If there is a change in amino acid type at one position, its correlation with another position for a given physiochemical property is not affected if the new amino acid has the same value as the physiochemical property. The property-based correlation results in the correlation between the properties values at the different positions during the course of evolution. The correlation matrix of each family for 4 different properties is calculated and discussed in the Results sections.

Spectral Analysis of Correlation Matrix:

For a given family, the eigenvalues and eigenvectors of the correlation matrix for each physiochemical property are calculated. The eigenvalue distribution of the correlation matrix is created. In order to quantify the statistical noise representing the phylogenetic effects in the data and estimate the random noise, we make use of the Wishart matrices (6,7) Wishart matrices are sample correlation matrices from a multivariate normal distribution. For a random matrix Z of dimension N × M with entries distributed as Gaussian random variable having mean as zero and variance as one. The correlation matrix resulting from the random matrix Z is a Wishart matrix. The spectral properties of Wishart matrices are studied in detail (6,7). The probability density function $P_{ran}(\lambda)$ of the Wishart matrices follow the Marcenko-Pastur distribution distribution

$$P_{ran}(\lambda) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \qquad [3]$$

where Q is defined as ratio of number of observations (sequences) and number of variables (Positions) ie Q= S/L ≥ 1. The upper and lower bounds of the eigenvalue distribution is given by

$$\lambda_{\pm} = \sigma^2 \left(1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}\right) \qquad [4]$$

The eigenvalue bounds for all the families are estimated and compared with the largest and the smallest eigenvalues of the property based correlation matrices.

Universal properties

Nearest-neighbor eigenvalue spacing distribution

Next, we create the level spacing distribution for the protein families. The level spacing distribution also known as nearest-neighbor spacing distribution is estimated for all families with different physiochemical properties. To construct the level spacing distribution, first, an unfolding process is done on the eigenvalue spectra. The unfolding transforms the eigenvalues in terms of the local average spacings, the details of the process can be found in (5, 10,11). In the new system, the modified eigenvalues χ_i has a unit local density. The level spacing distribution of a protein family is compared with the level spacing distribution of the correlation matrix created from a matrix X with the same dimension as the protein family but belongs to the Gaussian orthogonal ensemble (GOE). The GOE correlation matrix is given by

$$C_{GOE} = \frac{1}{S} X X^T \qquad [5]$$

X^T is the transpose of the random matrix X. The Wishart matrix and C_{GOE} are the same. The nearest-neighbor spacing between two levels is given by s= $\chi_{i+1} - \chi_i$ where χ_{i+1} and χ_i are two consecutive eigenvalues arranged in the ascending order. The nearest-neighbor spacing distribution of the C_{GOE} is given by

$$P_{GOE}(s) = \frac{\pi s}{2} exp\left(\frac{-\pi s^2}{4}\right) \qquad [6]$$

The nearest-neighbor spacing distribution of each protein families with different physio-chemical properties is created and compared with P_{GOE}. The next nearest-neighbor spacing distribution is also compared for the P_{GOE} and protein families.

Long-range eigenvalue correlations:

Level spacing distribution (nearest-neighbor) captures the short range correlation within the eigenvalue spectrum. To study the long range correlations, we calculate the number variance $\Sigma^2(\mathrm{L})$, and study its variation for different physiochemical properties. The number variance is given as

$$\Sigma^2(\mathrm{L}) = < \left[N\left(\chi + \frac{l}{2}\right) - N\left(\chi - \frac{l}{2}\right) - l \right]^2 >_\chi \qquad [7]$$

The quantity $N(\chi)$ defined as $N(\chi) = \sum_i \Theta(\chi - \chi_i)$ is the integrated unfolded density. The average is done over the unfolded eigenvalues. The number variance variation with l is studied. For the a uncorrelated spectrum the number variance varies as l. For the GOE ensemble the number variance shown ln(l) (12). We check the number variance all seven families are estimated and analyzed and is shown in the Results section.

Results

The proposed method is applied to all seven protein families which were taken from the PFAM database. The details of the eigenvalue statistics for all families are given in Table 1. Protein families show significant divergence from the Marchenko-Pastur distribution with many eigenvalues

outside the Wishart lower and upper bound. The number of eigenvalues outside the Wishart bounds or the outliers changes with physiochemical properties even within a family although they are derived from the same MSA. The details of each family analyzed are given below.

a) **Mitochondrial Protein (PF00153):** The MSA consists of 160 sequences each of length L=101. The MSA is first converted into the 3-dimensional data matrix. We then create the sequence vector from the data matrix. The sequence vector of the first 3 sequences for four physicochemical properties is shown in Figure 1. We observe that all the sequences follow the same trend for the hydrophobicity property. This indicates that the hydrophobicity property is conserved for the Mitochondrial Protein family. The three other properties show significant differences in the sequence vector.

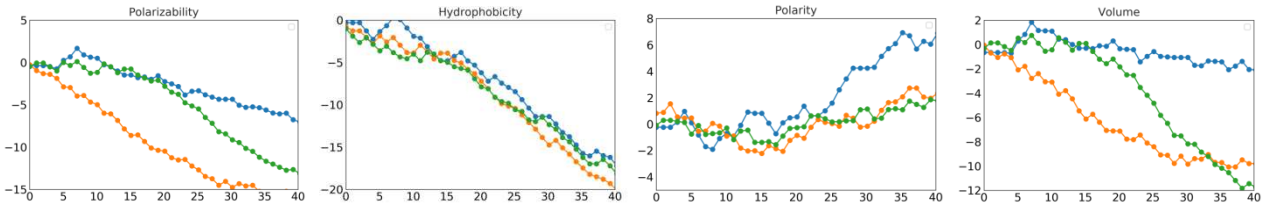


Figure 1. Plot of the three sequence vectors with first 40 positions for 4 physiochemical properties for Mitochondrial Protein family.

For the mitochondrial protein family, the correlation matrix is constructed for all four physiochemical properties and the eigenvalues and eigenvectors are estimated. The details of the eigenvalues, as well as the comparison with Wishart matrices, are given in Table 1. The eigenvalue bounds are the same for all properties as they depend only on the size of MSA and are independent of the property under consideration. The eigenvalue distribution is shown in Figure 2. For the eigenvalue spectra, the nearest neighbor spacing distribution is created and is shown in Figure 3(a) for hydrophobicity. The spacing distribution for other properties is identical to hydrophobicity. The spacing distribution shows resemblance to GOE and hence is from the same universality class. The next nearest spacing distribution Figure 3(b) and the number variance Figure 3(c), also suggest that the family closely follow the GOE universal class. The number variance varies as ln(l) which is the same as observed in GOE matrices.

Table 1. Details of the eigenvalue statistics of each family for different phyochemical properties with the theoretical eigenvalue RMT bounds on eigenvalues.

Protein Family	**Property**	**Size**	**EV Bounds**	$\lambda < \lambda_-$	$\lambda > \lambda_+$	**Smallest λ**	**Largest λ**
Mitochondrial	Hdrophobicity	S=160 L=101	0.05-3.21	6	5	0.025	6.37
	Polarity			5	5	0.024	7.15
	Polarizability			6	5	0.021	5.72
	Volume			7	6	0.033	5.63
Expanded EBP	Hdrophobicity	S=675 L=134	0.29-2.09	36	10	0.039	25.47
	Polarity			32	8	0.026	27.09
	Polarizability			28	9	0.039	25.65
	Volume			38	11	0.039	25.64
Cap-Gly	Hdrophobicity	S=593 L=83	0.4-1.88	12	8	0.172	5.75
	Polarity			14	9	0.064	6.98
	Polarizability			18	6	0.238	4.07
	Volume			13	5	0.784	7.89
Cadherin 4	Hdrophobicity	S=694 L=77	0.45-1.77	7	5	0.344	3.47
	Polarity			6	5	0.267	3.70
	Polarizability			7	6	0.280	4.35
	Volume			5	4	0.322	3.96
IPD Protein	Hdrophobicity	S=487 L=151	0.19-2.43	23	12	0.001	7.19
	Polarity			29	9	0.001	8.32
	Polarizability			32	8	0.005	6.721
	Volume			26	9	0.0004	6.87
Histidine Kinase	Hdrophobicity	S= 658 L=155	0.28-2.2	25	10	0.127	10.53
	Polarity			23	6	0.142	9.245
	Polarizability			27	8	0.100	11.17
	Volume			24	5	0.094	10.65
HAM1	Hdrophobicity	S=1061 L=208	0.31-2.07	50	15	0.018	25.81
	Polarity			48	12	0.017	24.58
	Polarizability			45	14	0.022	25.34
	Volume			47	9	0.051	18.59

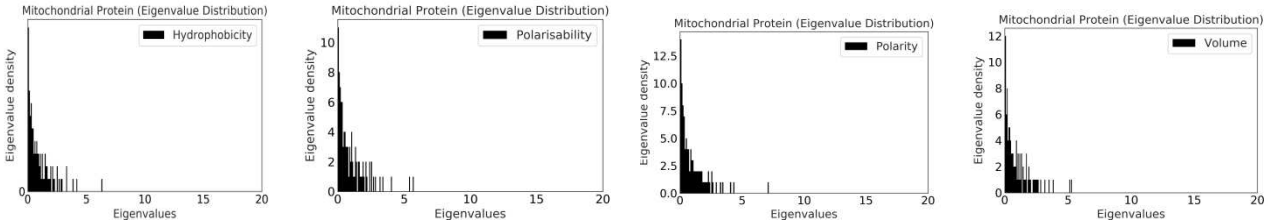
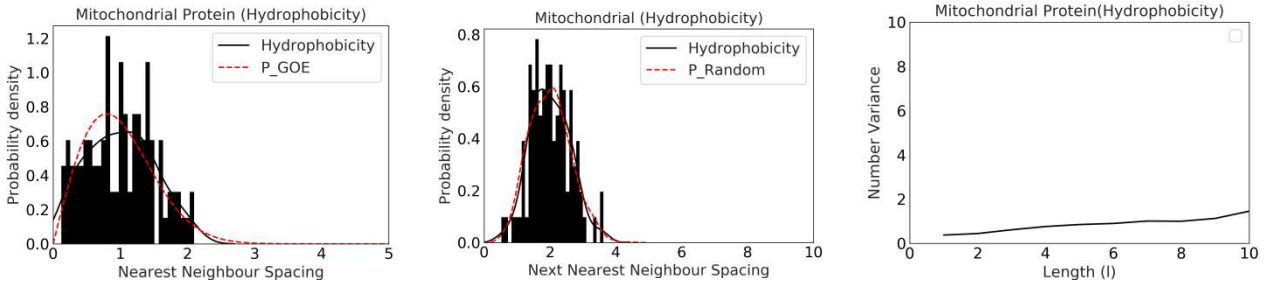


Figure 2. Eigenvalue distribution of Mitochondrial Protein family for all properties.



(a) **(b)** **(c)**

Figure 3. **(a)** Nearest neighbor eigenvalue spacing distribution **(b)** Next Nearest neighbor eigenvalue spacing distribution for the unfolded eigenvalues of the correlation matrix based on the hydrophobicity property for the Mitochondrial protein. The solid line is the density plot for the Mitochondrial Protein family and the dashed line is the nearest and next nearest neighbor spacing distribution for correlation matrix created for the GOE ensemble. **(c)** Number Variance calculated from unfolded eigenvalues.

b).Expanded EBP (EXPERA) Protein (PF05241): EBP enzyme family consists of MSA with 675 sequences and 134 positions giving the theoretical estimates for the random bounds on the eigenvalues as $\lambda_+ = 2.098$ and $\lambda_- = 0.298$. The sequence vector for three sequences and the first 40 residues is shown in Figure 4. We find that there is no conserved property for this family. The eigenvalue statistics is given in Table 1. The family also belongs to the GOE universal class as shown by the nearest-neighbor spacing distribution Figure 6(a), next nearest-spacing distribution Figure 6(b) and number variance Figure 6(c).

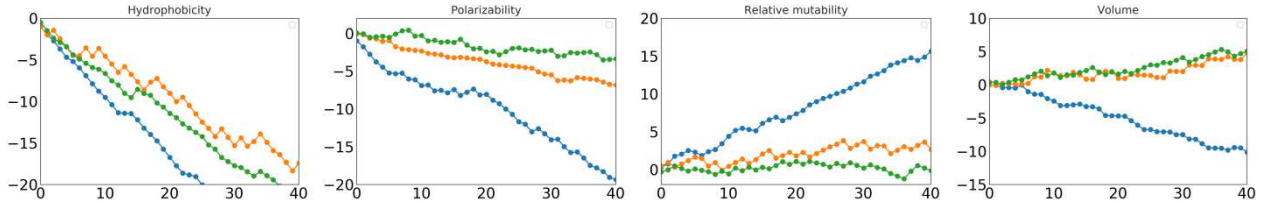


Figure 4. Plot of the three sequences with first 40 positions for different properties for the Expanded EBP Protein family.

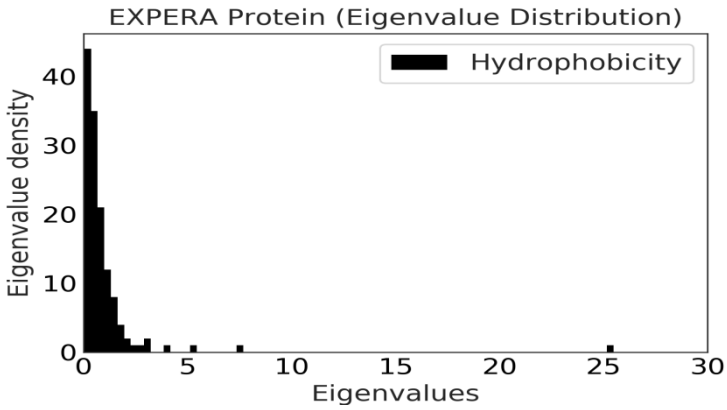
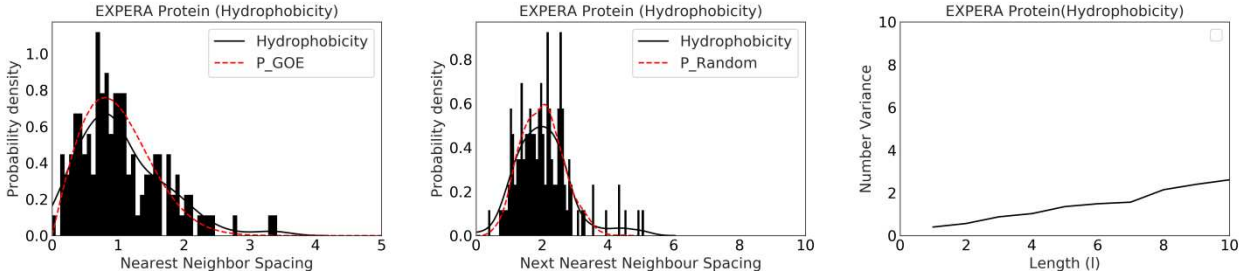


Figure 5. Eigenvalue distribution of Expanded (EBP) Protein family for Hydrophobicity



(a) **(b)** **(c)**

Figure 6. (a) Nearest neighbor eigenvalue spacing distribution **(b)** Next Nearest neighbor eigenvalue spacing distribution for the unfolded eigenvalues of the correlation matrix based on the hydrophobicity property for the EXPERA family. The solid line is the density plot for the EXPERA Protein family and the dash line is the nearest and next nearest neighbor spacing distribution for correlation matrix created GOE ensemble.**(c)** Number Variance for family created from unfolded eigenvalues

c). CAP-Gly Protein (PF01302): The MSA of this family consist of 593 sequences with 83 positions giving the RMT bounds as $\boldsymbol{\lambda_+ = 1.88}$ and $\boldsymbol{\lambda_- = 0.4}$. The sequence vector [Figure 7] shows that the there is a slight conservation hydrophobicity, there is a significant divergence in sequence vector for all other properties. This family strongly belongs to the GOE universal class as shown by the nearest-neighbor spacing distribution Figure 9(a), next nearest-spacing distribution Figure 9(b) and number variance Figure 9(c) with very strong resemblance to GOE ensemble.

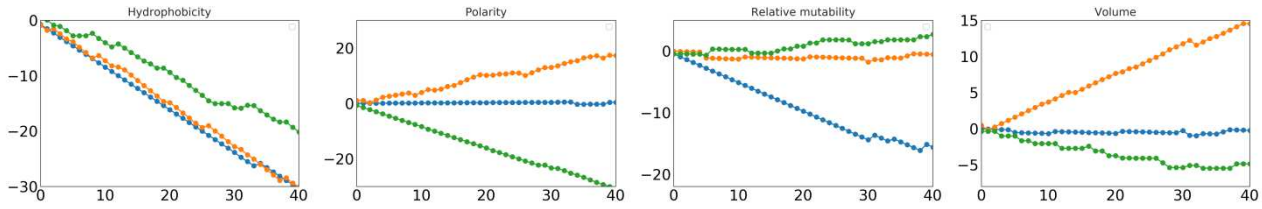


Figure 7. Plot of the three sequence vectors with first 40 positions for 4 physiochemical properties for CAP-Gly Protein family.

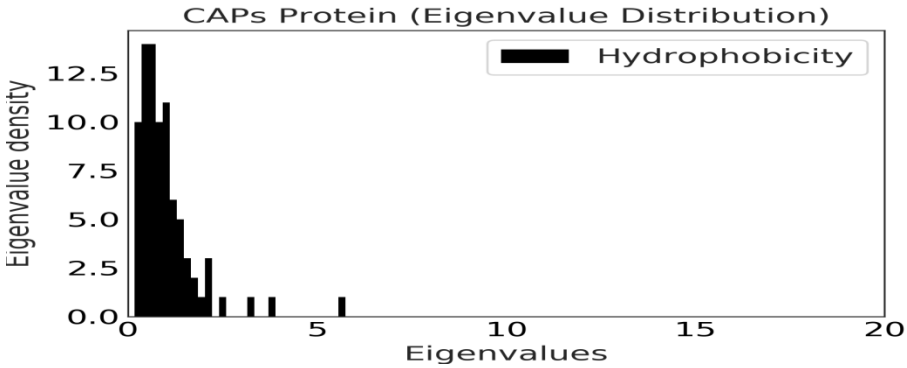
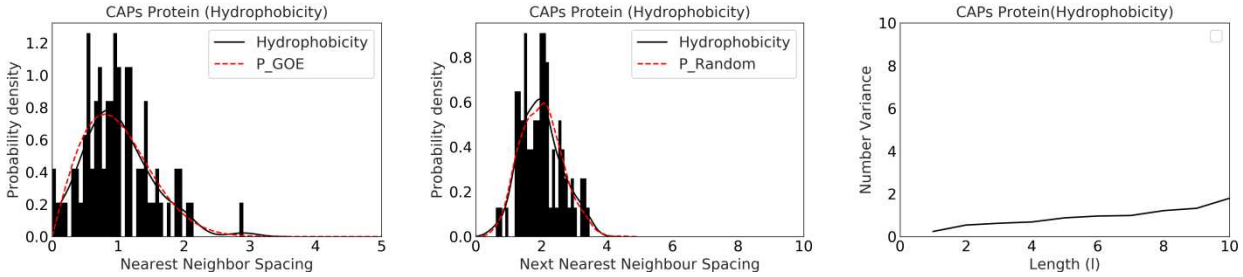


Figure 8. Eigenvalue distribution of CAPs Protein family for Hydrophobicity



(a) **(b)** **(c)**

Figure 9. (a) Nearest neighbor eigenvalue spacing distribution **(b)** Next Nearest neighbor eigenvalue spacing distribution for the unfolded eigenvalues of the correlation matrix based on the hydrophobicity property for the CAPs Protein family. The solid line is the density plot for the CAPs Protein family and the dash line is the nearest and next nearest neighbor spacing distribution for correlation matrix created GOE ensemble.**(c)** Number Variance calculated from unfolded eigenvalues.

d). Cadherin 4 Protein (PF17803): MSA of this family consist of 694 sequences with 77 positions giving the RMT bounds as $\boldsymbol{\lambda_+ = 1.77}$ and $\boldsymbol{\lambda_- = 0.45}$. The sequence vector [Figure 10] shows that the there is a slight conservation in polarizability and hydrophobicity. The spacing distribution [Figure 12(a)] shows close resemblance to GOE and hence is from the same universality class. The next nearest spacing distribution [Figure 12(b)] and the number variance [Figure 12(c)], also suggests that the family closely follow the GOE universal class.

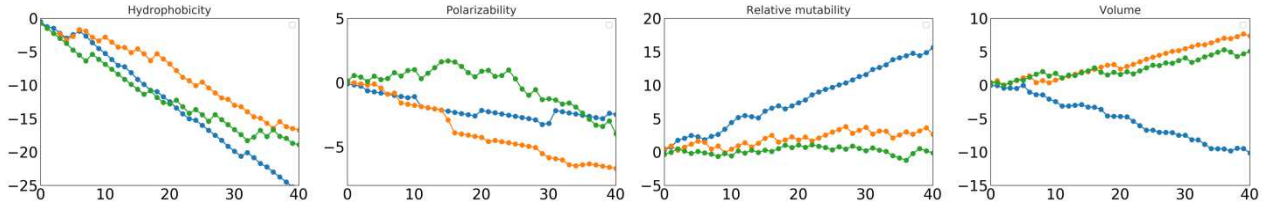


Figure 10. Plot of the three sequence vectors with first 40 positions for 4 physiochemical properties for Cadherin 4 Protein family.

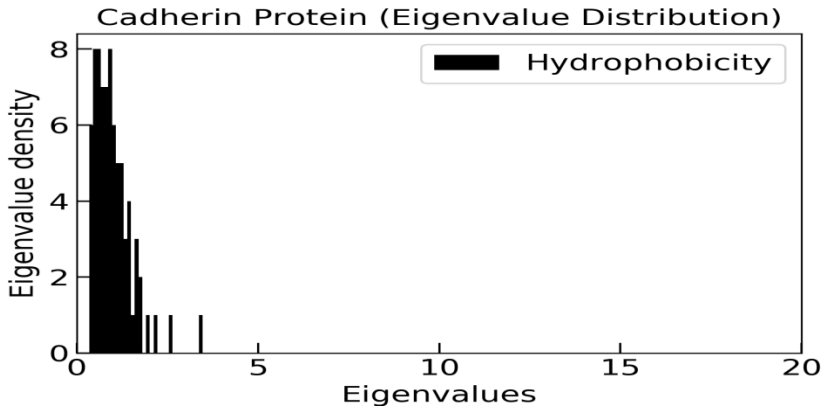
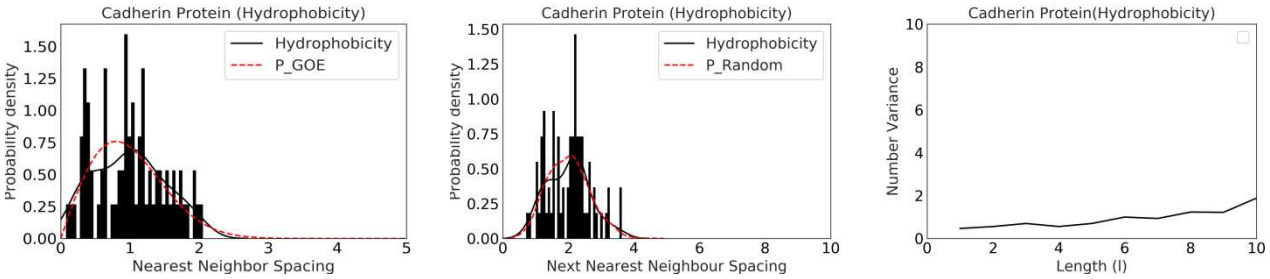


Figure 11. Eigenvalue distribution of Cadherin Protein family for Hydrophobicity



(a) **(b)** **(c)**

Figure 12. **(a)** Nearest neighbor eigenvalue spacing distribution **(b)** Next Nearest neighbor eigenvalue spacing distribution for the unfolded eigenvalues of the correlation matrix based on the hydrophobicity property for the Cadherin Protein family. The solid line is the density plot for the Cadherin Protein family and the dash line is the nearest and next nearest neighbor spacing distribution for correlation matrix created GOE ensemble.**(c)** Number Variance calculated from unfolded eigenvalues.

e). IPD Protein (PF00475): IPD protein family has 487 sequences and 151 positions giving the theoretical estimates for RMT bound on eigenvalues distribution as $\lambda_+ = 2.43$ and $\lambda_- = 0.19$. The sequence vector for three sequences and the first 40 residues is shown in Figure 13. . The family also belongs to the GOE universal class as shown by the nearest-neighbor spacing distribution Figure 15(a), next nearest-spacing distribution Figure 15(b) and number variance Figure 15(c).

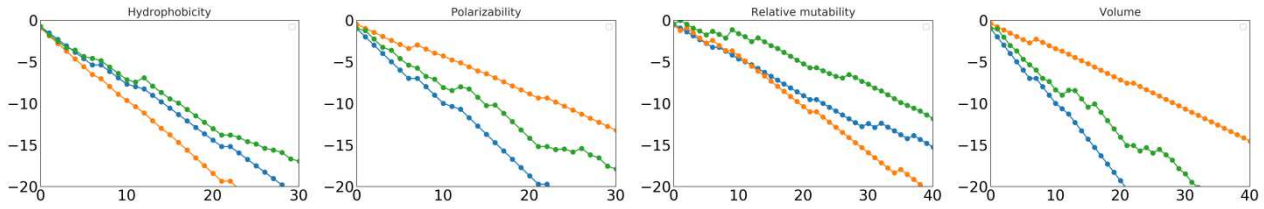


Figure 13. Plot of the three sequence vectors with first 40 positions for 4 physiochemical properties for IPD Protein family

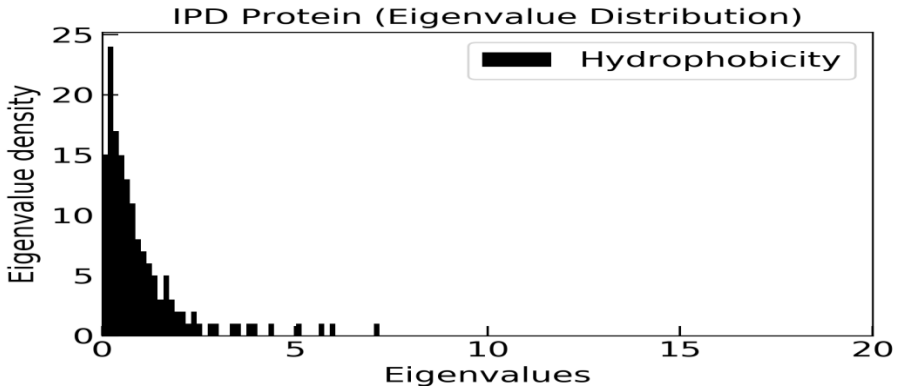
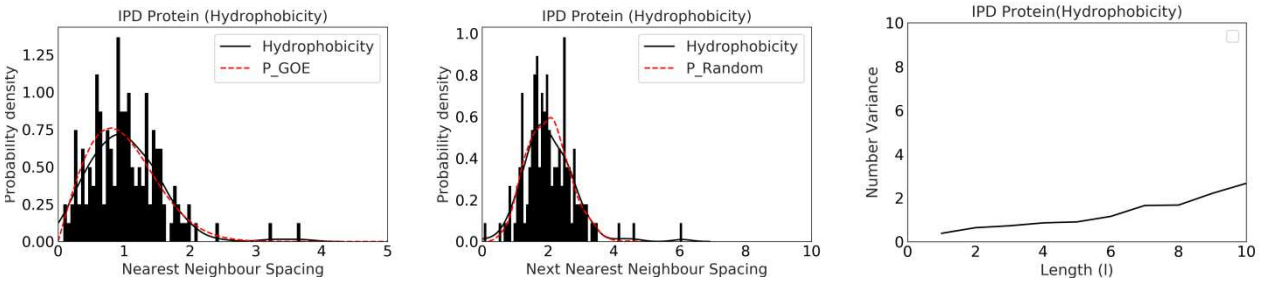


Figure 14. Eigenvalue distribution of IPD Protein family for Hydrophobicity.



(a) **(b)** **(c)**

Figure 15. **(a)** Nearest neighbor eigenvalue spacing distribution (b) Next Nearest neighbor eigenvalue spacing distribution for the unfolded eigenvalues of the correlation matrix based on the hydrophobicity property for the IPD Protein family. The solid line is the density plot for the IPD Protein family and the dash line is the nearest and next nearest neighbor spacing distribution for correlation matrix created GOE ensemble.**(c)** Number Variance calculated from unfolded eigenvalues.

f) Histidine Kinase Protein (PF02518): The MSA of this family has 658 sequences and 155 positions giving the theoretical RMT estimates on the eigenvalues as $\lambda_+ = 2.20$ and $\lambda_- = 0.28$. The sequence vector for [Figure 16] shows that there is no conserved property for this family. The family also belongs to the GOE universal class as shown by the nearest-neighbor spacing distribution Figure 18(a), next nearest-spacing distribution Figure 18(b) and number variance Figure 18(c).

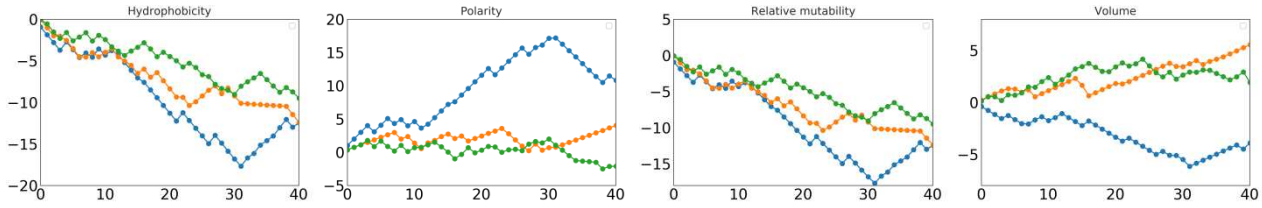


Figure 16. Plot of the three sequence vectors with first 40 positions for 4 physiochemical properties for Histidine Kinase Protein family.

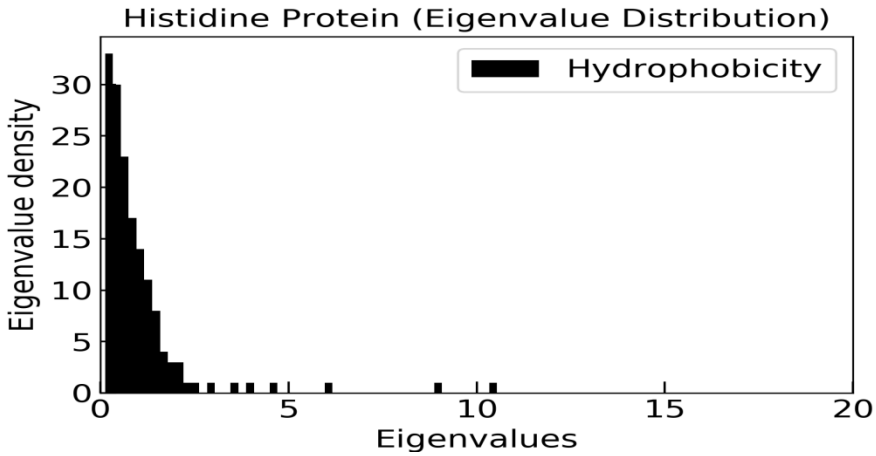
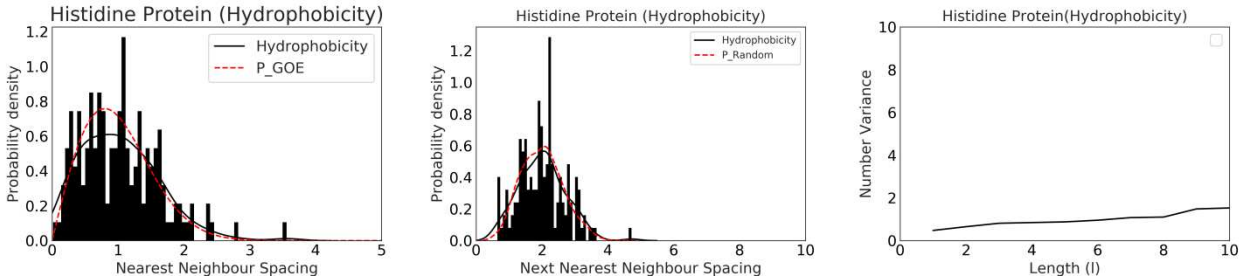


Figure 17. Eigenvalue distribution of Histidine Protein family for Hydrophobicity.



(a) **(b)** **(c)**

Figure 18. **(a)** Nearest neighbor eigenvalue spacing distribution (b) Next Nearest neighbor eigenvalue spacing distribution for the unfolded eigenvalues of the correlation matrix based on the hydrophobicity property for the Histidine Protein family. The solid line is the density plot for the Histidine Protein family and the dash line is the nearest and next nearest neighbor spacing distribution for correlation matrix created GOE ensemble.**(c)** Number Variance calculated from unfolded eigenvalues.

g) HAM1 Protein (PF01725): HAM1 family consists of 1061 sequences and 208 positions giving the theoretical RMT bounds as $\lambda_+ = 2.07$ and $\lambda_- = 0.31$. The sequence vector for [Figure 19] shows that there is no conserved property for this family. The family also belongs to the GOE universal class as shown by the nearest-neighbor spacing distribution [Figure 21(a)], next nearest-spacing distribution [Figure 21(b)] and number variance [Figure 21(c)].

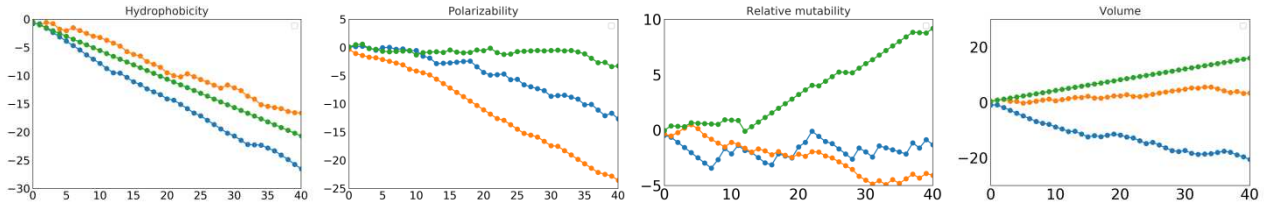


Figure 19. Plot of the three sequence vectors with first 40 positions for 4 physiochemical properties for HAM1 Protein family

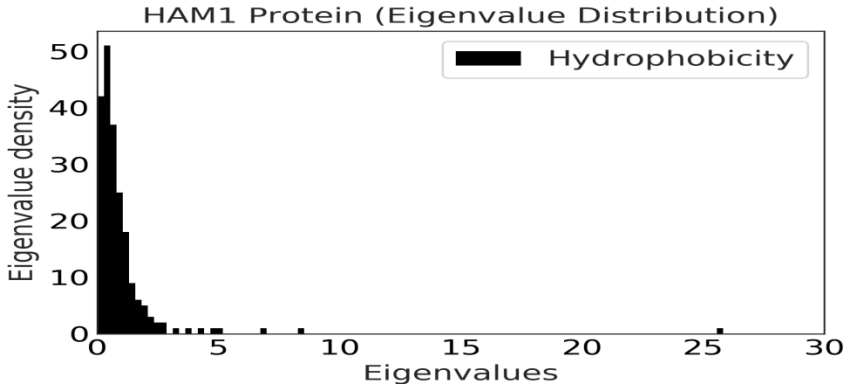
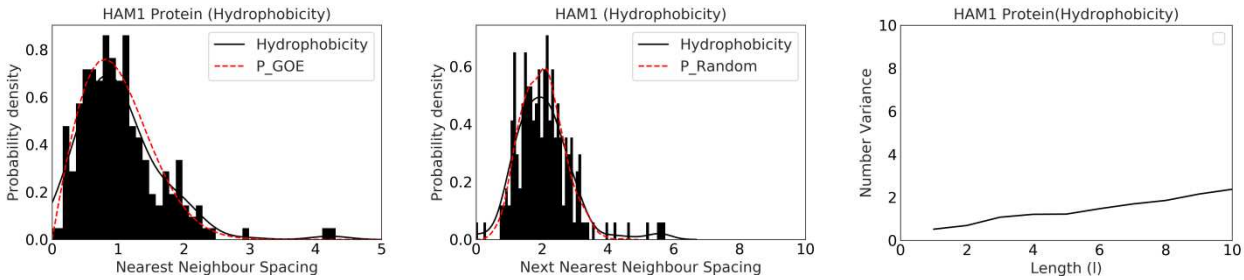


Figure 20. Eigenvalue distribution of HAM1 Protein family for Hydrophobicity



(a) **(b)** **(c)**

Figure 21. **(a)** Nearest neighbor eigenvalue spacing distribution (b) Next Nearest neighbor eigenvalue spacing distribution for the unfolded eigenvalues of the correlation matrix based on the hydrophobicity property for the HAM1 Protein family. The solid line is the density plot for the HAM1 Protein family and the dash line is the nearest and next nearest neighbor spacing distribution for correlation matrix created GOE ensemble.**(c)** Number Variance calculated from unfolded eigenvalues

Conclusions

We study the spectral properties of the correlation matrix created with 4 different physicochemical properties for multiple protein families. The eigenvalues statistics are compared with the RMT bounds from the Marchenko-Pastur distribution and the number of eigenvalues outside the bounds are estimated containing significant information about the system. All families show significant divergence from the RMT results. The protein families follow the GOE universality class as tested with the level spacing distribution and number variance which varies as the log of the size of the system similar to the GOE ensemble.

Acknowledgments

PB wants to thank Mahidol University for research support "MRCMGR 04/2565". ND wants to thank Science and Engineering Research Board, Department of Science and Technology, India (SERB-DST No. EMR/2016/006536) for research support.

References

1. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. Sonnhammer, S. C. Tosatto, L. Paladin, S. Raj, L. J. Richardson and R. D. Finn, *Nucleic Acids Research*, **49**(D1), D412 (2021). doi: 10.1093/nar/gkaa913.
2. V.V. Sokolov, and V.G. Zelevinsky, *Nucl. Phys. A* **504,** 562 (1989). doi: 10.1016/0375-9474(89)90558-7.
3. P. Bhadola, I. Garg, and N. Deo, *Nucl. Phys. B* **870**, 384 (2013). doi: 10.1016/j.nuclphysb.2013.01.010
4. P. Bhadola, and N. Deo, *Physical Review E*, **83**(3), 032706 (2013). doi: 10.1103/PhysRevE.88.032706.
5. L. Laloux, P. Cizeau, J.P. Bouchaud,and M. Potters, *Phys. Rev. Lett.*, **83,** 1467 (1999). doi: 10.1103/PhysRevLett.83.1467
6. M. J. Bowick and E. Brezin, *Phys. Lett. B*, **268**, 21 (1991). doi:10.1016/0370-2693(91)90916-E
7. J. Feinberg and A. Zee, *J. Stat. Phys.,* **87**, 473 (1997). doi: 10.1007/BF02181233
8. P. Bhadola, and N. Deo, *Physical Review E*, **94**(4), 042409 (2016). doi: 10.1103/PhysRevE.94.042409
9. P. Bhadola, and N. Deo, *Journal of Physics: Conference Series*, **1144**(1), 012083 (2018). doi: 10.1088/1742-6596/1144/1/012083.
10. T. A. Brody, J. Flores, J. B. French, P. A. Mello, A. Pandey, and S. S. M. Wong, Rev. Mod. Phys., **53**, 385 (1981). doi: 10.1103/RevModPhys.53.385.
11. T. Guhr, A. Muller-Groeling, and H. A. Weidenmuller, *Phys. Rep.*, **299**, 189 (1998). doi: 10.1016/S0370-1573(97)00088-4
12. V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Nunes Amaral, and H. E. Stanley, *Phys. Rev. Lett.*, **83**, 1471 (1999), doi: 10.1103/PhysRevLett.83.1471.