Functional Motifs in a Protein Family

Dr. Nivedita Deo

Department of Physics & Astrophysics
University of Delhi,
Delhi, India.

Collaborator: *Pradeep Bhadola*

Department of Physics & Astrophysics
University of Delhi, Delhi, India.

Indian statistical physics community meeting
18/02/2017

System : Beta-Lactamase Family

- Beta-lactamases: Enzymes secreted by bacteria in response to beta-lactam antibiotics like Pencillin & Bacterial Cephalosporins.
- Beta lactamase enzymes irreversibly hydrolyzes the amide bond of beta-lactam ring making beta-lactam antibiotics inactive.

New Antibiotics

Evolution of Enzyme (simple Mutation)

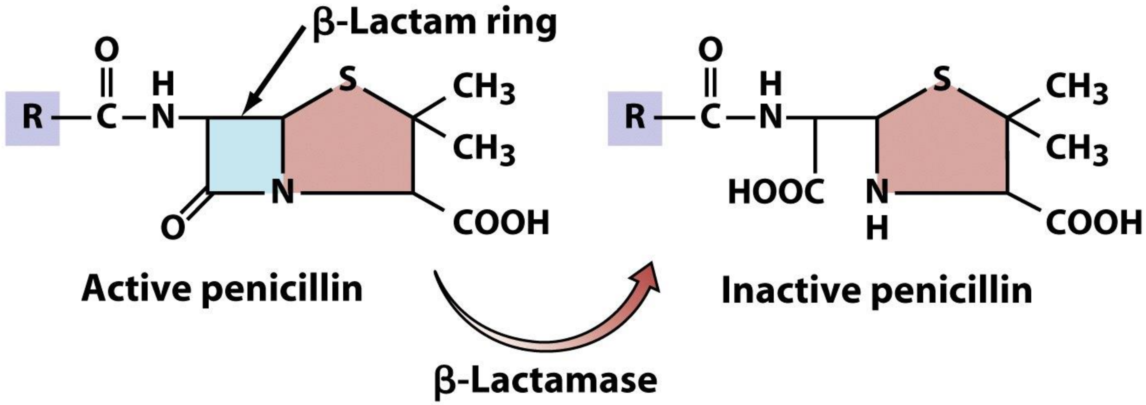


Figure 13-7 Microbiology, 6/e
© 2005 John Wiley & Sons

Antibiotics Resistance

increased spectrum of action on antibiotics.

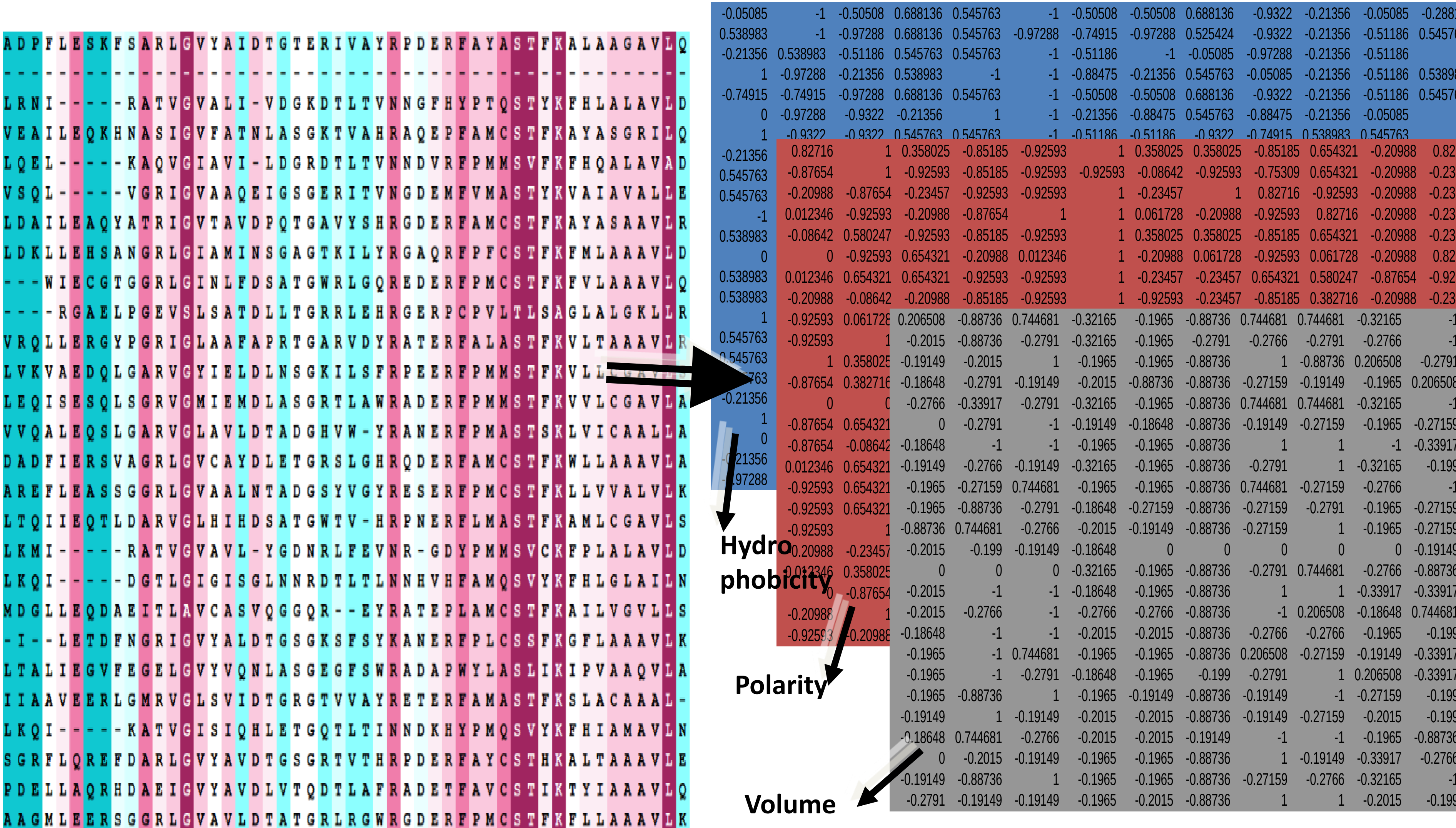
Challenge to medicinal future of antibiotics

❑ To identify the important motifs or sectors in the protein family for targets to control (deactivate or activate) the enzymatic actions.

- **Interpro entry IPR000871 comprising 5447 proteins for class A/D beta-lactamases family .** Goal: Bring the greatest number of similar characters into the same column of the alignment S=559 L=248

Multiple Sequence Alignment

Physiochemical Based Datamatrices



Co-evolution Matrix: Frequency or String based

Covariance between amino acids (a & b) at positions i & j in a multiple sequence alignment is defined as

$$\boxed{C_{ij}^{(ab)} = f_{ij}^{(ab)} - f_i^{(a)} f_j^{(b)}}$$

Where $f_i^{(a)}$ is the frequency of having amino acid ‘a’ at position i

and $f_{ij}^{(ab)}$ is the joint frequency of having ‘a’ at position i & ’b’ at j.

The final co-evolution matrix is defined as

$$\boxed{\bar{C}_{ij} = \sqrt{\left(\sum_{a,b} \left(C_{ij}^{(ab)}\right)^2\right)}}$$

N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan; Cell 138, 774–786, (2009)

Correlation Matrix

The co-evolution between positions of the data matrix D is given by

$$C_{i,j}^{\alpha} = \frac{Cov(d_i^{\alpha}, d_j^{\alpha})}{\sigma_{d_i^{\alpha}} \sigma_{d_j^{\alpha}}}$$

Where d_i^{α} is the i^{th} column of data matrix D^{α} having standard deviation $\sigma_{d_i^{\alpha}}$

$Cov(d_i^{\alpha}, d_j^{\alpha})$ is the covariance between i^{th} and j^{th} column of data matrix D^{α} given by

$$Cov(d_i^{\alpha}, d_j^{\alpha}) = \langle (d_{s,i}^{\alpha} - \langle d_{s,i}^{\alpha} \rangle)(d_{s,j}^{\alpha} - \langle d_{s,j}^{\alpha} \rangle) \rangle$$

<..> implies average over sequences.

Noise Dressing

The resulting correlation matrix from random data matrix of dimension $S \times L$ with elements drawn from a distribution with mean zero and standard deviation sigma is a Wishart matrix with well studied properties.

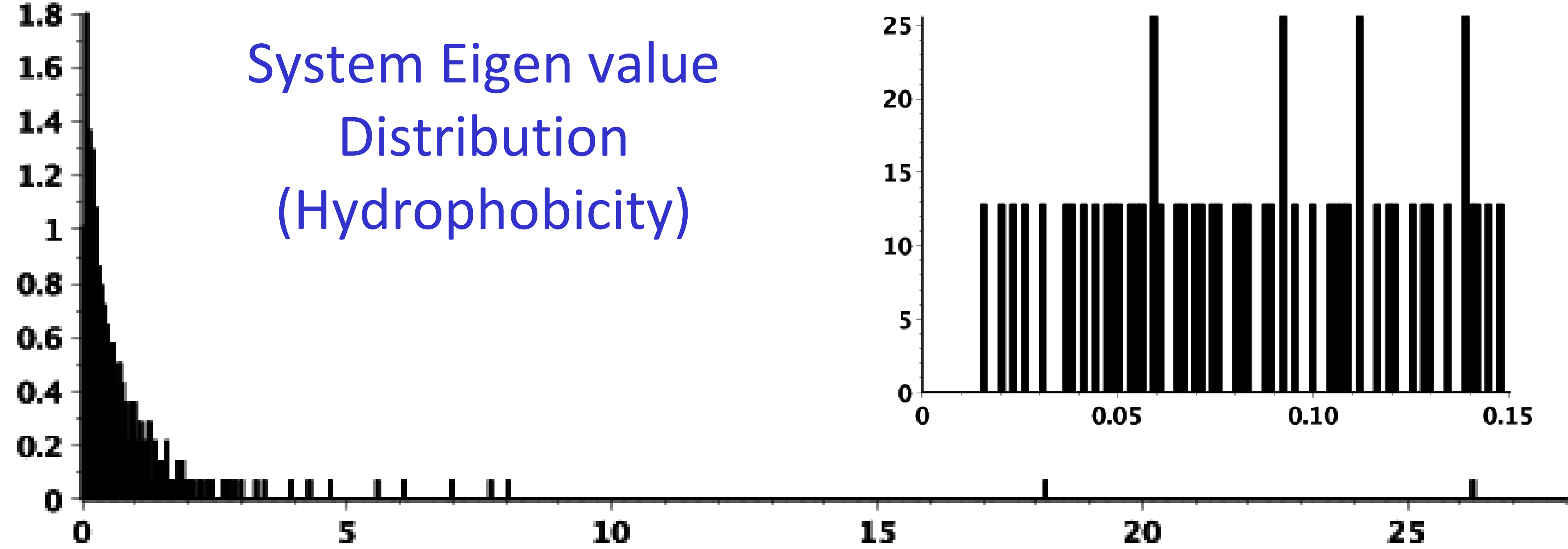
The eigenvalue bounds are given by

$$\lambda_{\pm} = \sigma^2 \left(1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}\right) \quad \text{with} \quad Q = \frac{S}{L} \geq 1$$

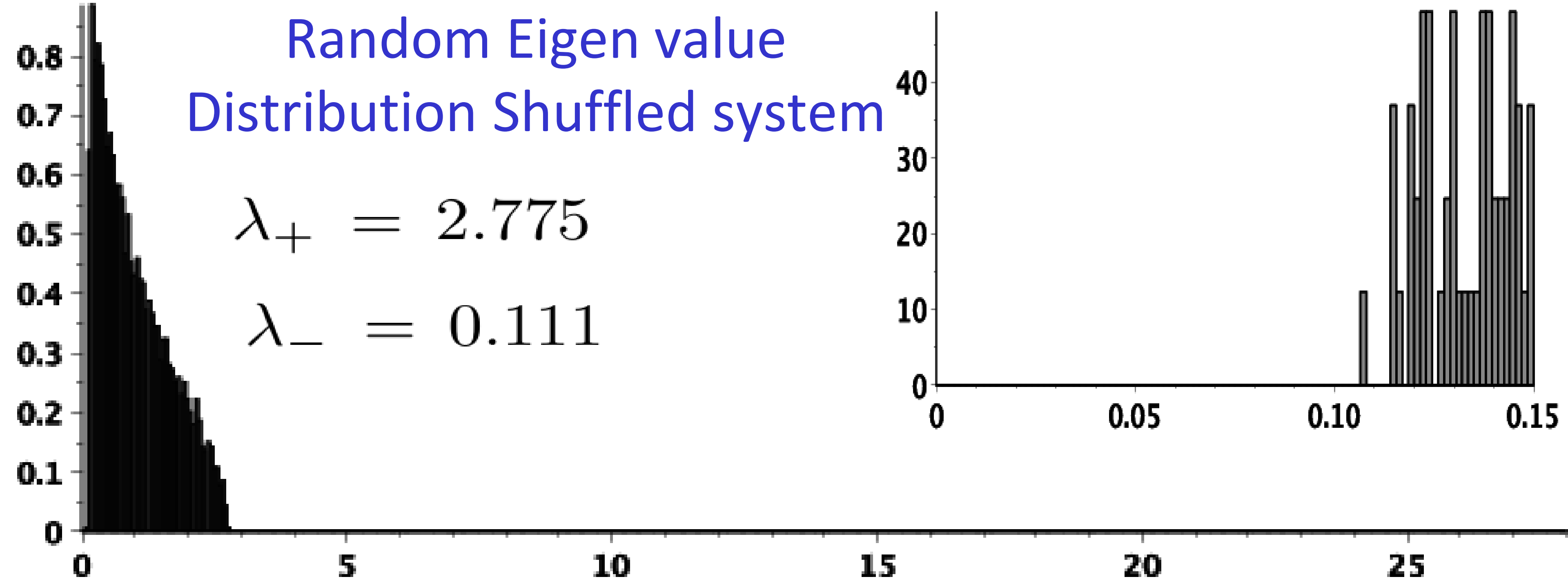
For Beta-Lactamase family S=559 and L=248, giving

$\lambda_{+} = 2.775$ and $\lambda_{-} = 0.111$

Computationally verified the bounds.



System Eigen value Distribution (Hydrophobicity)



Random Eigen value Distribution Shuffled system

$$\lambda_+ = 2.775$$

$$\lambda_- = 0.111$$

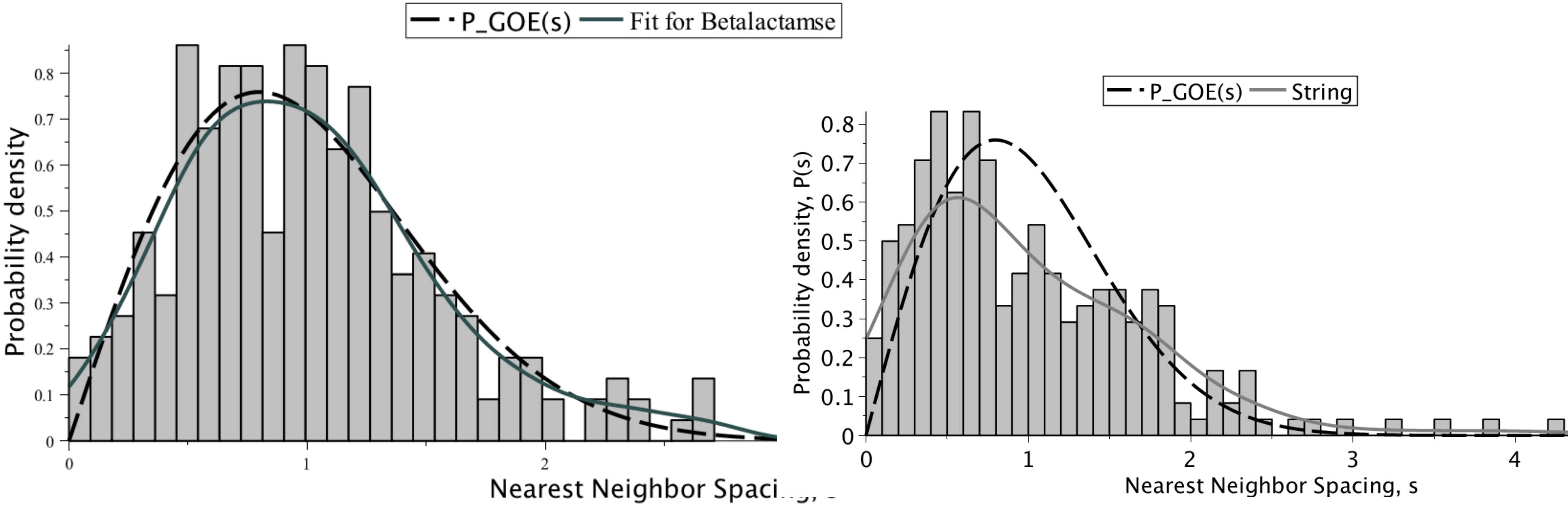
Nearest Neighbor eigenvalue spacing distribution

For Wishart matrices resulting from the GOE the eigenvalue spacing distribution is

$$P_{GOE}(s) = \frac{\pi s}{2} exp\Big(- \frac{\pi^2 s}{4}\Big).$$

Where $s = \chi_{i+1} - \chi_i$ with χ_i as the unfolded eigenvalue

Created by mapping λ_i to χ_i such that the local density of the new eigenvalue is one

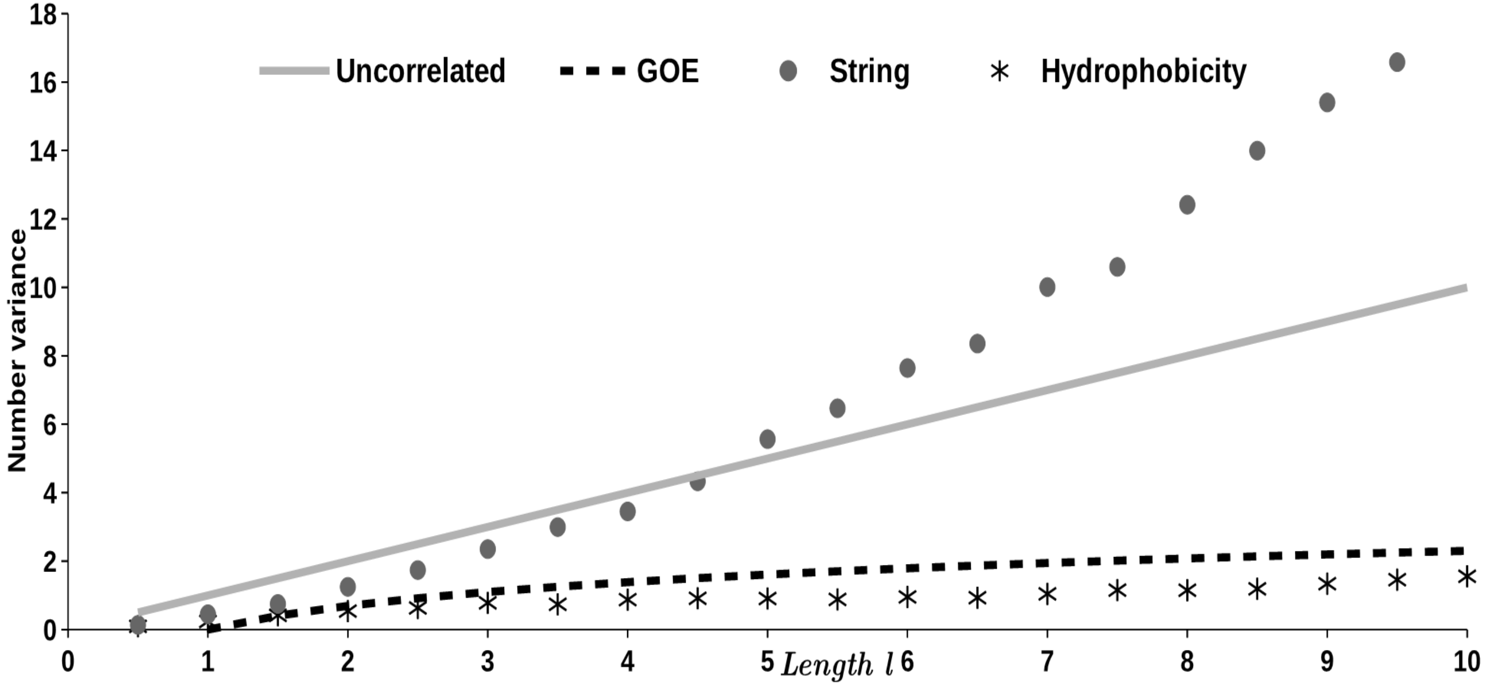


Long range eigenvalue correlations

The number variance $\sum^2$ gives the long range pair correlations present in the Eigen value Spectrum, defined as

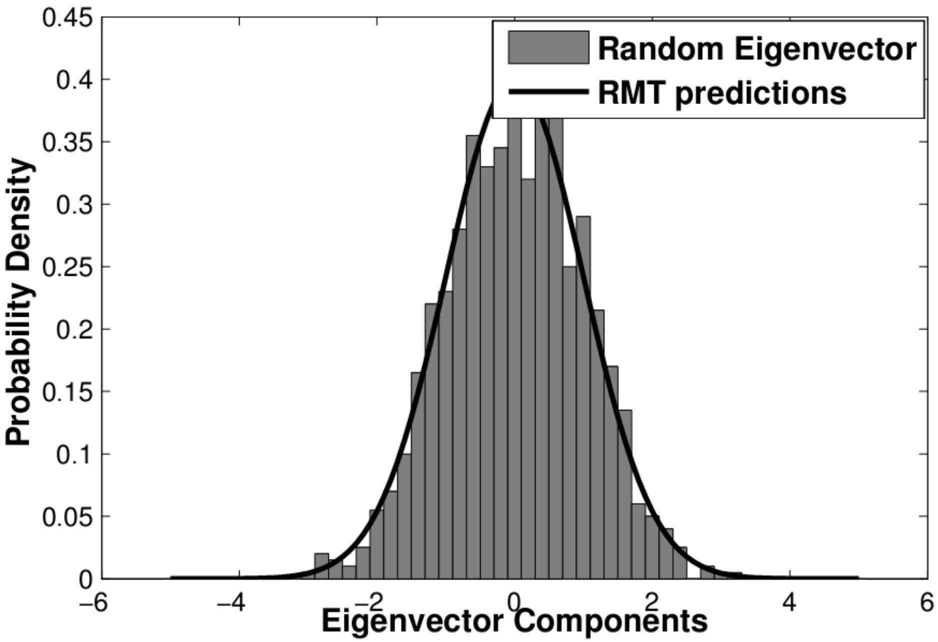
$$\Sigma^2(L) = \left\langle \left[N(\chi + \frac{l}{2}) - N(\chi - \frac{l}{2}) - l \right]^2 \right\rangle_\chi$$

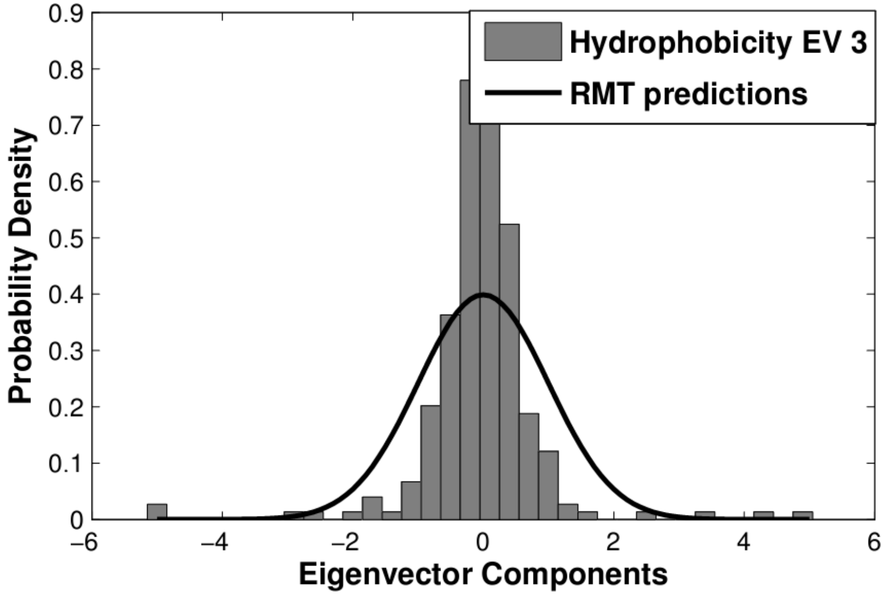
is the variance in the number of unfolded eigenvalues contained in the interval of length l around each un-folded Eigen value . Where $N(\chi) = \sum_i \theta(\chi - \chi_i)$

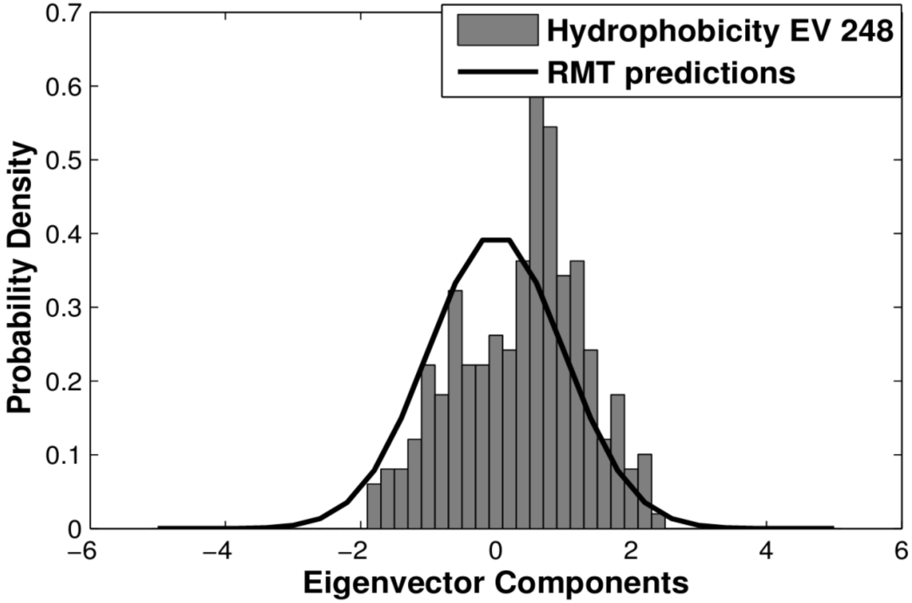


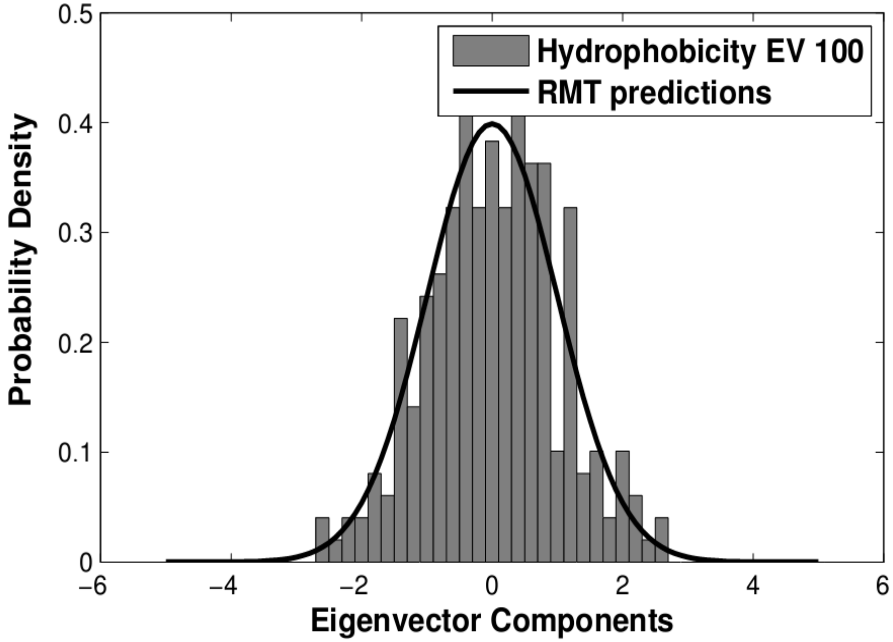
Number variance for the betalactamase family calculated from the unfolded eigenvalues of the hydrophobicity based correlation matrix and string based coevolution matrix. The observations are compared with the results of the GOE matrices and the uncorrelated series

Distribution of eigenvector components







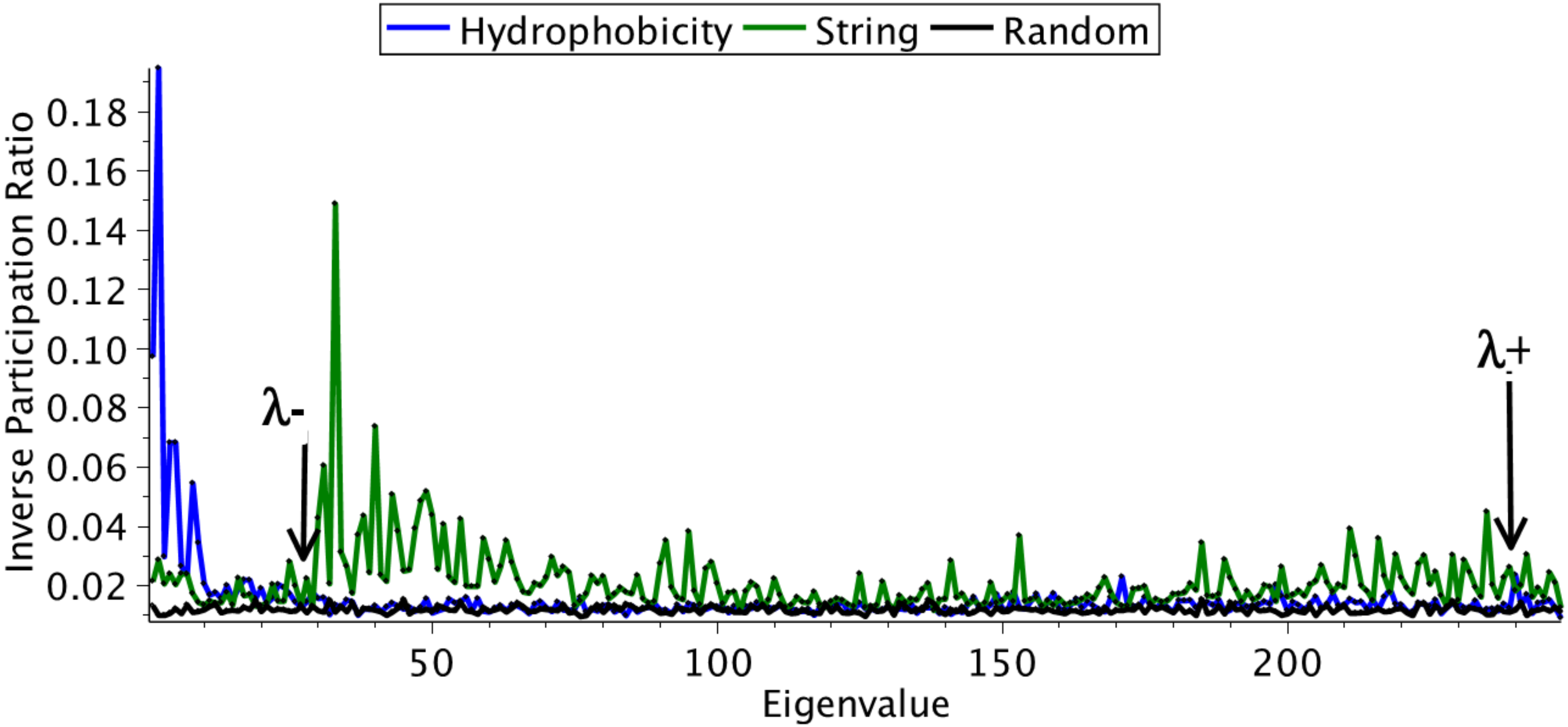


Inverse participation ratio (IPR)

The number of components that contributes significantly (unequal) will quantify the amount of deviation from the RMT predictions. To probe this we use the inverse par-ticipation ratio defined as

$$IPR(i) = \sum_{k=1}^{L} [v_i(k)]^4$$

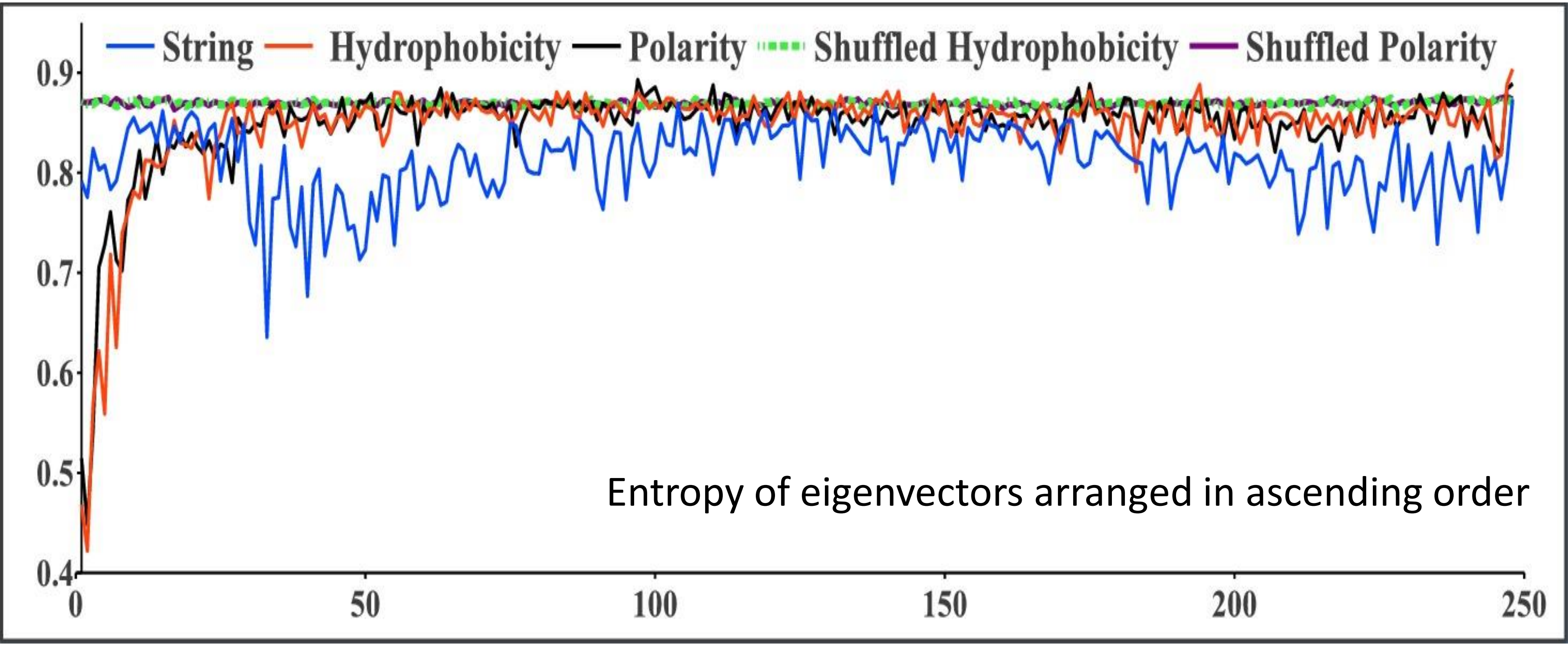
Where v_i is the normalized i^{th} eigenvector with the k^{th} component as $v_i(k)$



Spectral Analysis of Correlation Matrix

Entropy of Eigenvector ν_i

$$H_i = -\sum_{j=1}^{L}[v_i(j)]^2 log_L([v_i(j)]^2)$$

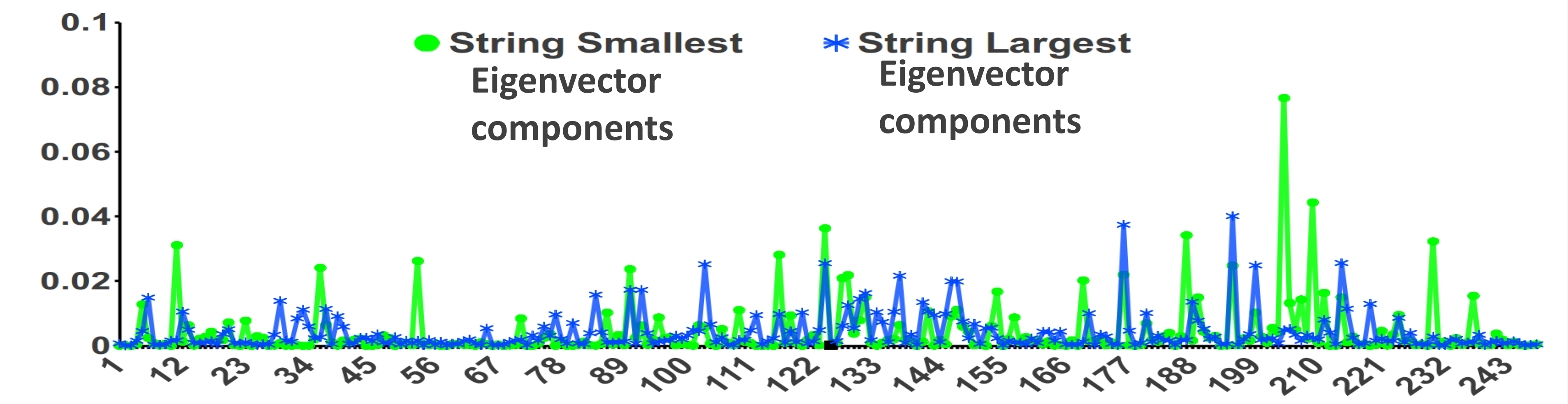


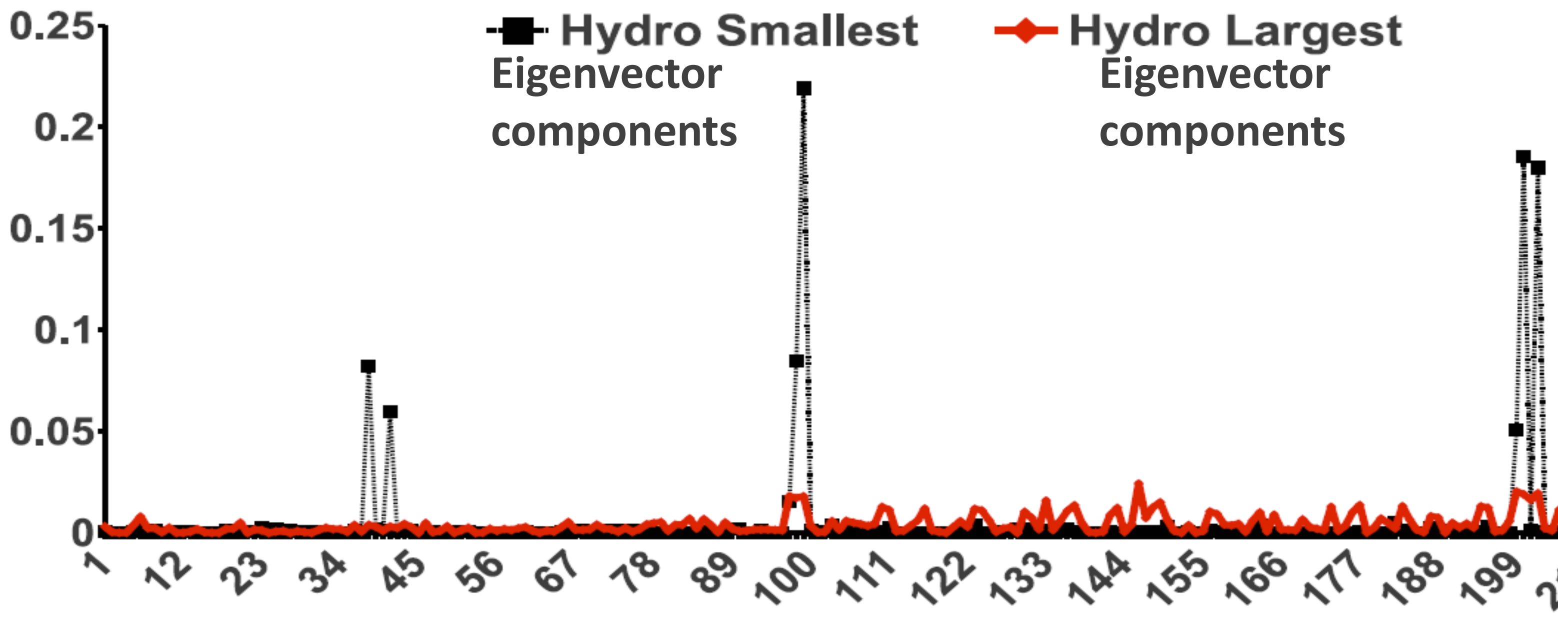
Entropy of eigenvectors arranged in ascending order

Smallest Eigenvectors are highly localized than Eigenvectors corresponding to large Eigen values.

Estimating sectors and interactions

Square of Eigenvector Components with low entropy gives the sites having significant contributions with a collective structural and functional role





Sector 1:
38, 41, 97, 98, 99, 199, 200, 202

Residues involved in Sectors

Using hydrophobicity scale, 248 positions of MSA are reduced to the following positions that shows high pattern of correlations

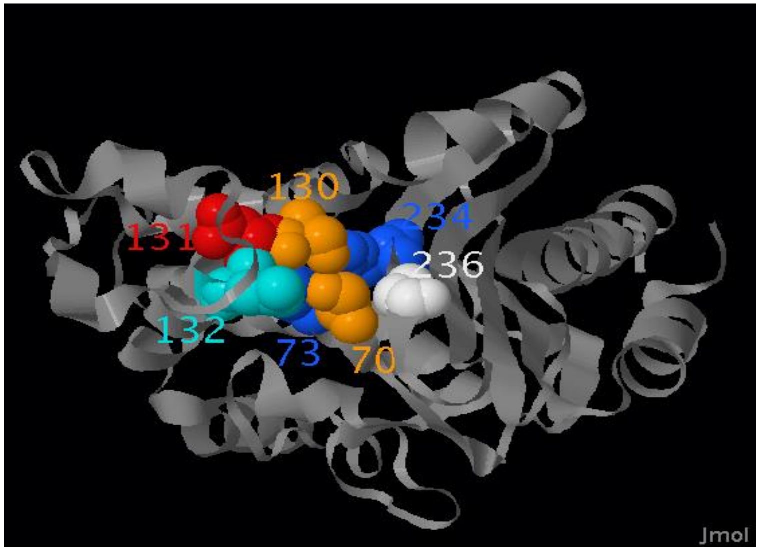
Eigen Vector	Sector Role	Residues Involved (number in Multiple Sequence Alignment)	Residues Involved as number in 1SHV.pdb
1	Conserved Motifs	38, 41, 97, 98, 99, 199, 200, 202	70 , 73, 130, 131, 132, 233, 234, 236
2	Catalytic & Ligand Binding Site	38, 41, 98, 124, 133, 147, 150, 199, 202	70, 73, 131, 157, 166 , 180, 183, 233, 236
4	Boundary Active site	38, 41, 97, 98, 99,124, 133, 137, 147, 200, 202	70, 73, 130, 131, 132, 157, 166, 170 , 180, 234 236
3	No identified Role (Individually important)	38, 49, 98, 99, 124, 133, 137, 146, 147, 150, 200, 202	70, 81, 131, 132, 157, 166, 170, 179, 180, 183, 234, 236

Analysis dealing with the polarity gives sector that characterizes substrate specificity and mutational stability. Positions contributing to that sector are **36, 103, 106, 129, 130, 131, 136, 137, 200, 245**.

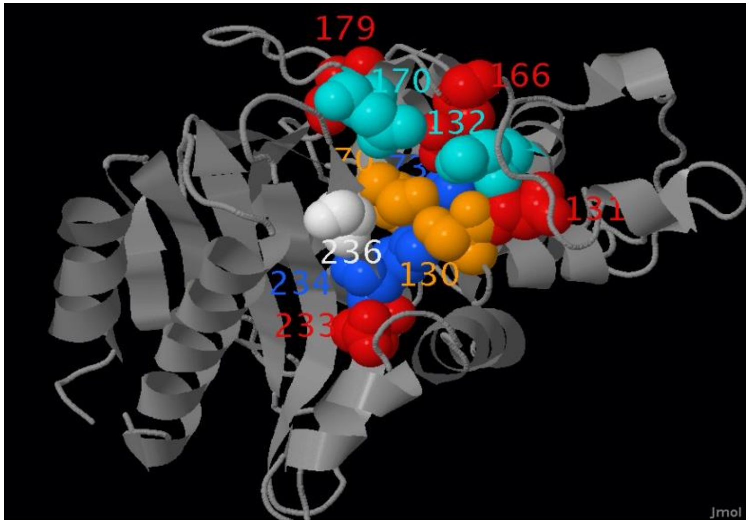
- I. Kather, et. al., J. Mol. Biol. 383, 238 (2008)
- A. Matagne et. Al., Biochem. J. 330, 581-598 (1998)
- X. Raquet et. al., Biophysical Journal 73, 2416 (1997).
- E. J. Lietz et. al., Biochemistry, 39, 4971-4981 (2000)
- Johannes C. Hermann et. al., J. AM. CHEM. SOC. 125, 9590 (2003).

Sectors Mapped on 1SHV.pdb

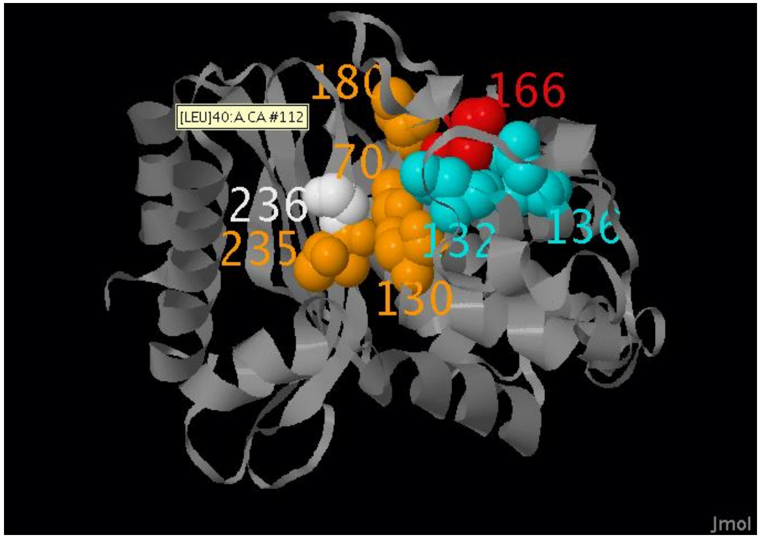
Sector 1



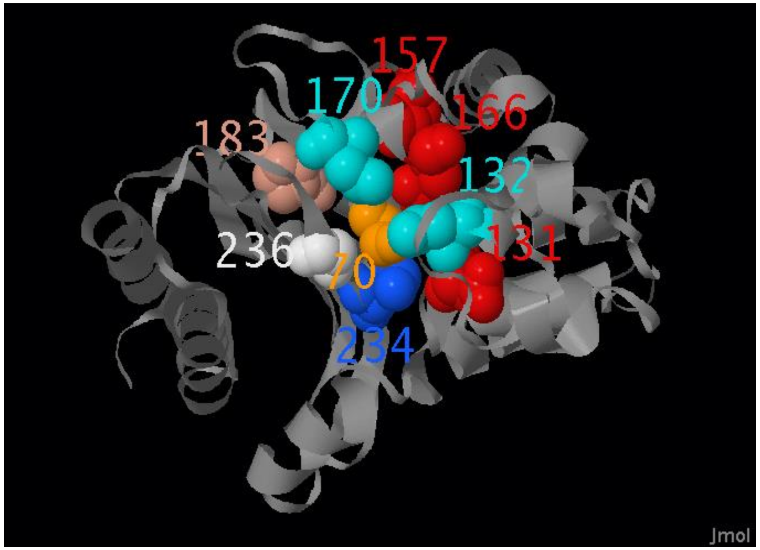
Sector 2



Sector 3



Sector 4



Positions contributing to sectors are close in the 3D structure

Network

The system can be represented by a graph G(N,E) with nodes (N) given by the positions in the multiple sequence alignment and edges

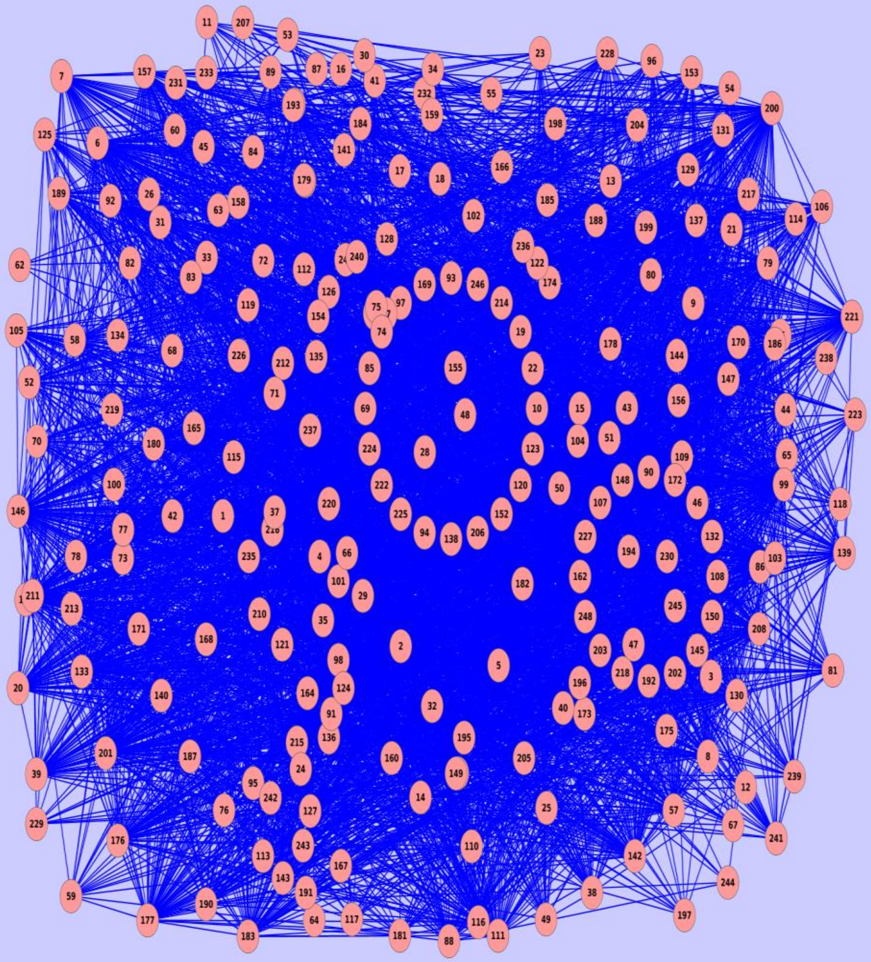
$$E_{i,j} = \begin{cases} 1 & if \quad i \neq j, C_{i,j} \geq \theta \\ 0 & otherwise. \end{cases}$$

where θ is the threshold value.

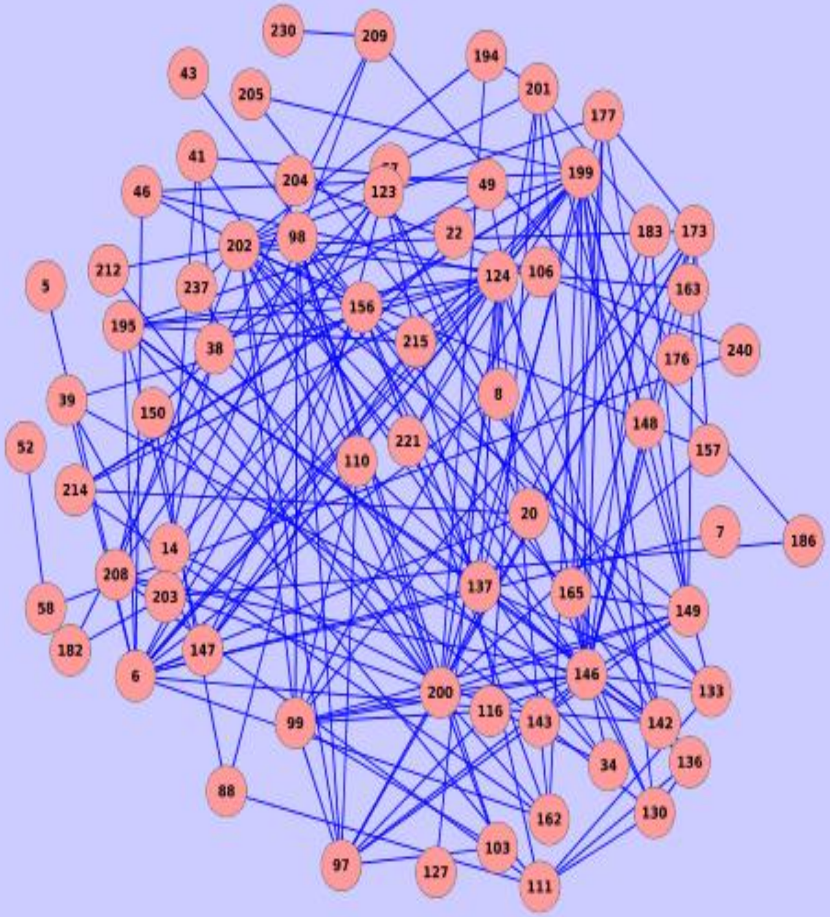
Connected components for different threshold are extracted which shows both structural and functional significance

Correlations as Network

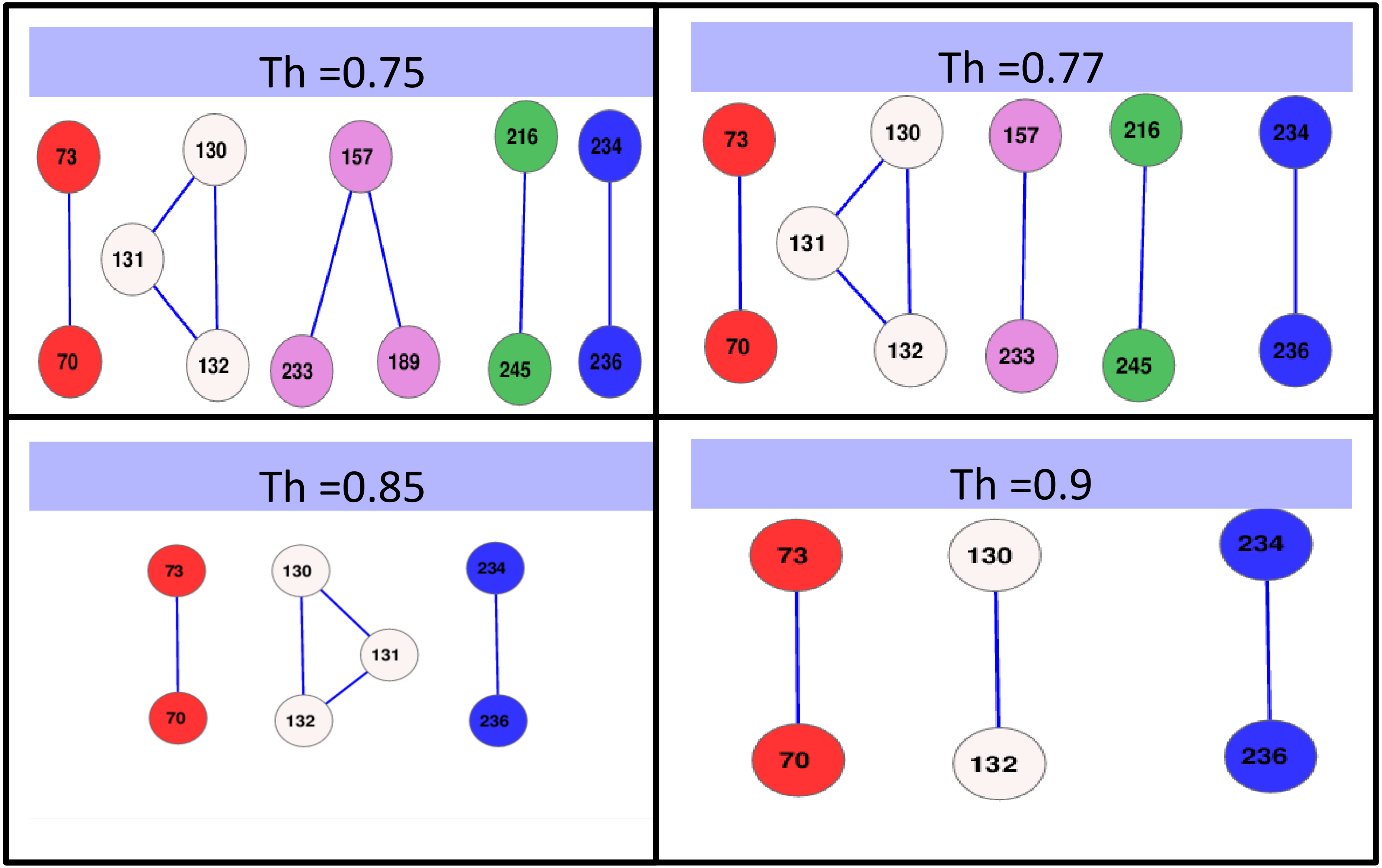
Threshold =0.2



Threshold =0.5



Connected Components


Th =0.75

Th =0.77

Th =0.85

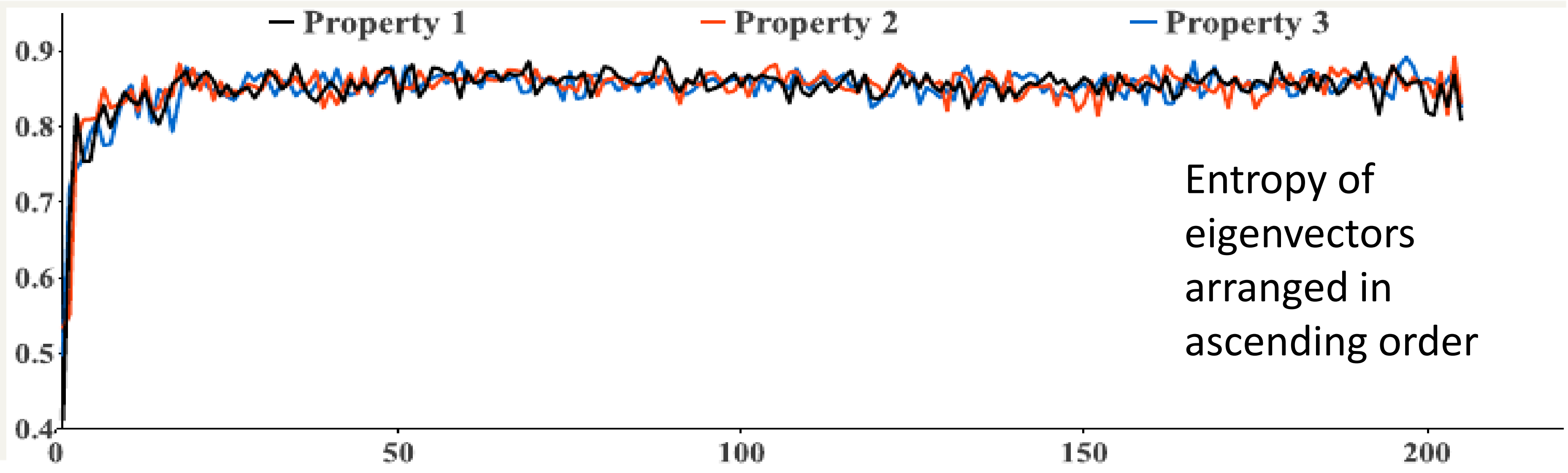
Th =0.9

Th=0.85 is same as sector 1 (conserved domain). Network analysis gives linking between positions

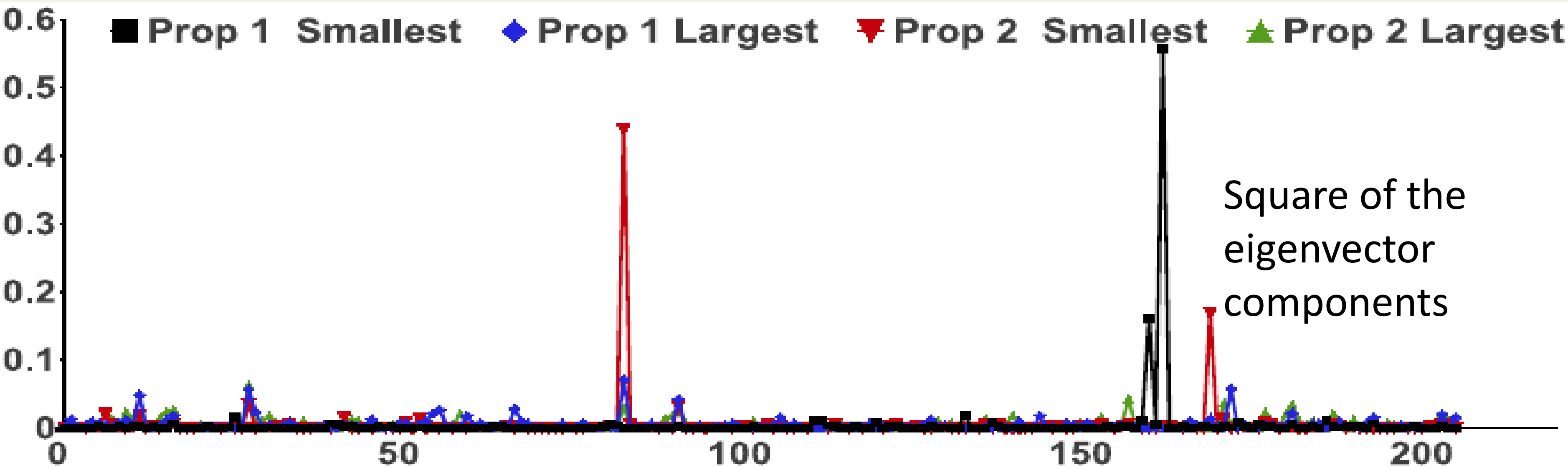
Other Protein families

Serine Protease

Protein family serine protease responsible for a wide variety of function which includes immune response, digestion, reproduction and blood coagulation.



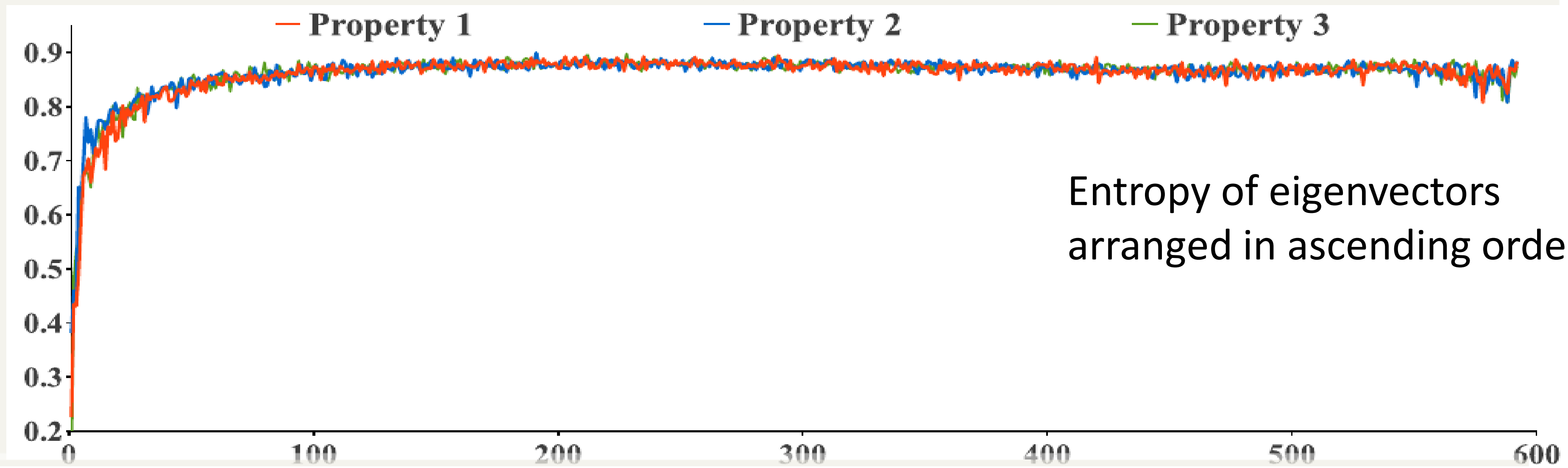
Entropy of eigenvectors arranged in ascending order



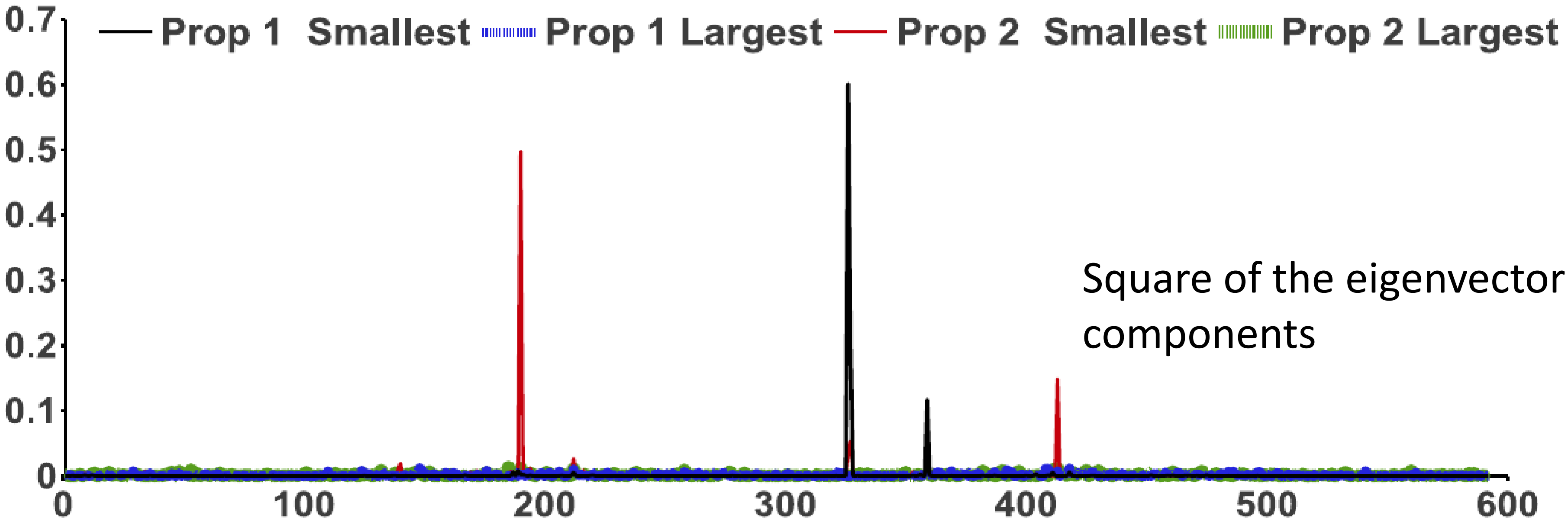
Square of the eigenvector components

HSP70 protein family

HSP70 protein family also known as the 70 kilodalton heat shock proteins plays a wide variety of function including assisting the protein folding process, refold the misfolded protein, inhibits apoptosis etc. Analyzing the family using the

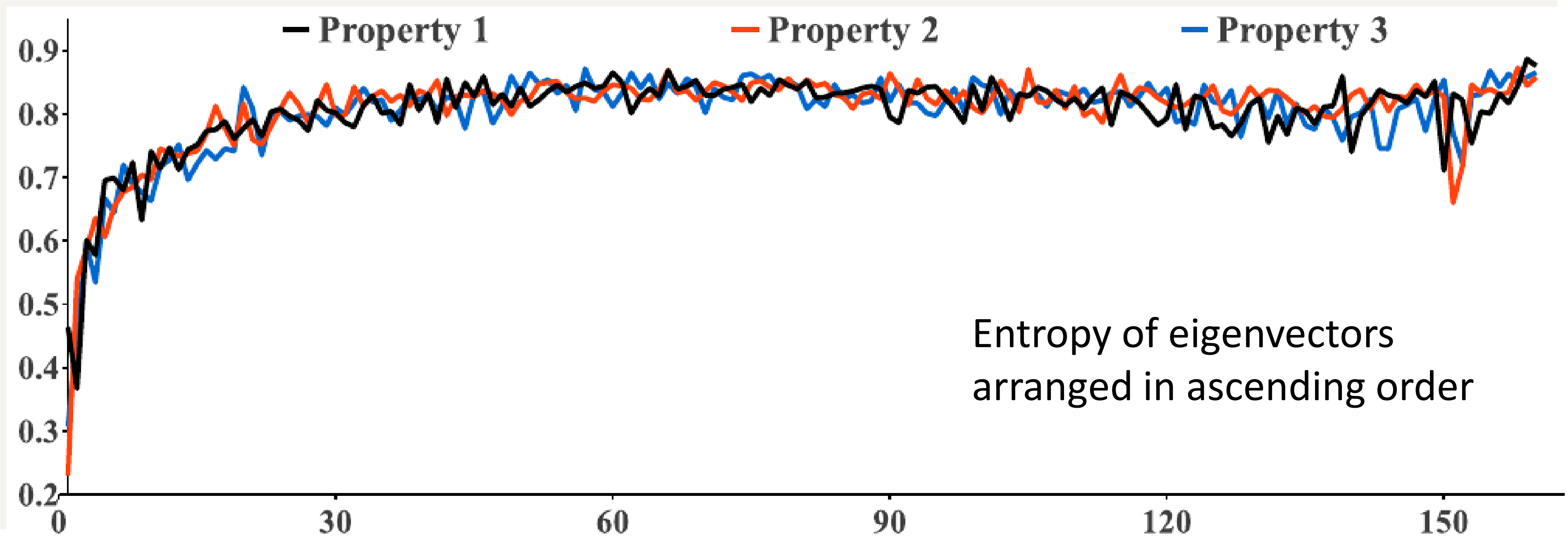


Entropy of eigenvectors arranged in ascending order

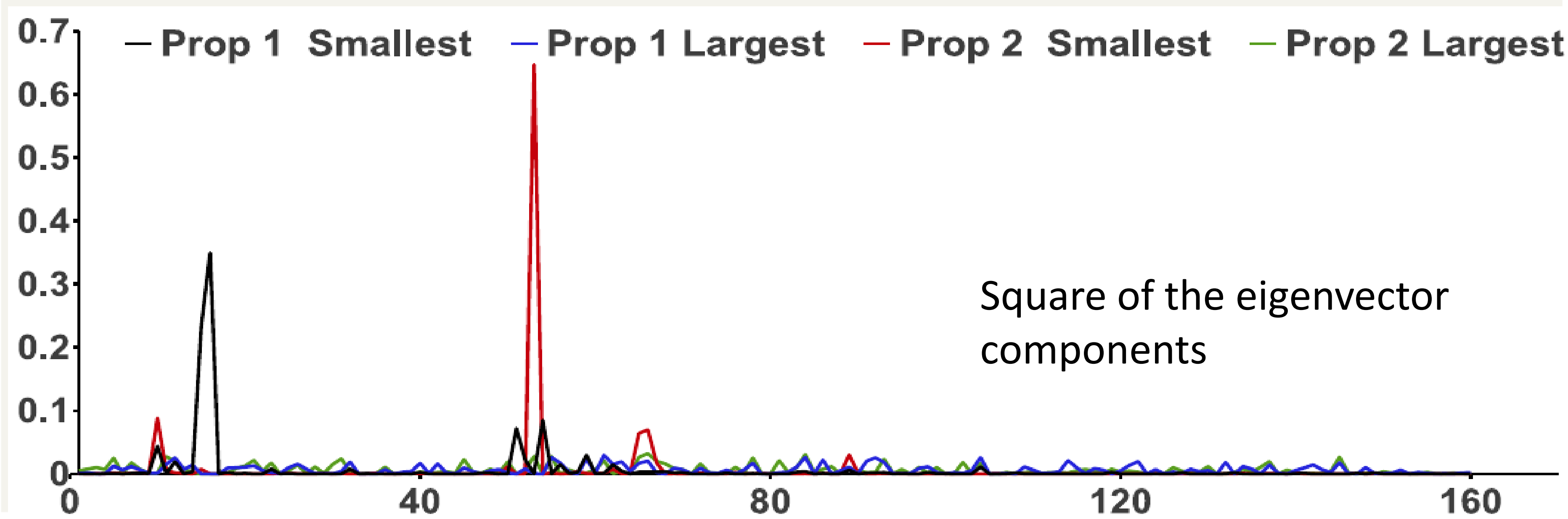


Square of the eigenvector components

G-protein protein family



Entropy of eigenvectors arranged in ascending order



Square of the eigenvector components

Conclusions

- Each protein sequence in MSA is represented as a multidimensional time series using different physiochemical properties.
- For protein families, the lower eigenmodes of the correlation matrix are more informative than the higher eigenmodes. These eigenmodes are exploited to extract structural and functional sectors.
- The graph theoretical analysis is used to visualize the property based interactions among positions and in the recognition of motifs by extracting the components in the graph.
- The property based correlation matrix is memory efficient and fast to compute as well as shows greater robustness to the sampling bias therefore it can be used with the existing MI based methods to enhance the speed as well as prediction accuracy.
- The spacing distribution reveals that there are short –range correlations between states present within the β-lactamase family , which shows a universal behavior. This universal behavior is clear in the property-based analysis while the string-based analysis deviates substantially from this universal behavior.

Acknowledgements

Financial Support:

1. CSIR-UGC for Senior Research fellowships

2. Delhi University R&D Scheme for financial support.

References

1. N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Cell **138**, 774 (2009).
2. S. Cocco, R Monsasson and M.Weigt, Plos Comput. Biol. 9, e1003176(2013).
3. I. Kather *et. al*., J. Mol. Biol. 383, 238 (2008).
4. A. Matagne et. al., Biochem. J. 330, 581-598 (1998).
5. X. Raquet et. al., Biophysical Journal 73, 2416 (1997).
6. E. J. Lietz et. al., Biochemistry, 39, 4971-4981 (2000).
7. Johannes C. Hermann et. al., J. AM. CHEM. SOC. 125, 9590 (2003).
8. A. Chakraborty et. al, Brief. Bioinform 16, bbt092 (2014).
9. Pradeep Bhadola and Nivedita Deo , Physical Rev. E 94, 042409 (2016).

Thank You