Graph Signal Processing on protein residue networks helps in studying its biophysical properties

Divyanshu Srivastava[1], Ganesh Bagler[1], Vibhor Kumar[1]

[a] *Department for Computational Biology, Indraprastha Institute of Information Technology, Delhi*
[b] *vibhor@iiitd.ac.in*

Abstract

Understanding the physical and chemical properties of proteins is vital, and many efforts have been made to study the emergent properties of the macromolecules as a combination of long chains of amino acids. Here, we present a graph signal processing based approach to model the biophysical property of proteins. For each protein inter-residue proximity-based network is used as basis graph and the respective amino acid properties are used as node-signals. Signals on node are decomposed on network's Laplacian eigenbasis using graph Fourier transformations. We found that the intensity in low-frequency components of graph signals of residue features could be used to model few biophysical properties of proteins. Specifically, using our approach, we could model protein folding-rate, globularity and fraction of alpha-helices and beta-sheets. Our approach also allows amalgamation of different types of chemical and graph theoretic properties of residue to be used together in a multi-variable regression model to predict biophysical properties.

Keywords: **Keywords.** Graph Signal Processing, Residue Interaction Graph, Graph Fourier Transform

1. Introduction

Proteins are the fundamental building blocks of a cell. The sequence of amino-acids (or residues) is stabilized into a native, functional three-dimensional state of the protein. Thus, the smaller building blocks of the protein emerge

as functional only when they are arranged in a defined arrangement in a three-dimensional space [15]. It is a rather well-known fact that protein's overall biophysical behaviour is governed by this peculiar order and residues arrangement. Therefore in addition to protein sequence information and secondary structure, researchers are also discovering new ways to utilize tertiary structure to model properties of proteins.

Multiple attempts have been made to model a protein molecule as network of residues as nodes connected with each other based on the distance between them in 3D structure of a protein [2, 5, 21]. Network properties of nodes (residues) based on graph theory such as centrality, betweenness and clustering coefficient have been used to predict biophysical properties using protein structures [5]. Another instance is community network analysis (CNA) which is used to study the dynamics of enzymes and protein/DNA (and/or RNA) complexes for understanding their allosteric mechanisms[20] [12]. Similarly, the folding rate of protein has also been modelled using network properties based on only graph-theoretic approach [1]. Though such approaches highlight the importance of residue network properties, however ignoring the biophysical properties of amino acids could lead to under-utilization of previously available information about the residues. Thus the question arises how to efficiently amalgamate different kinds of signals due to amino acid properties and 3D proximity neighbourhood information.

Here we propose a conceptually different approach using graph signal processing, which amalgamates residue properties and residue network structure to model proteins' biophysical properties. Notably, we show how a protein's structure information and residue's biophysical properties influence folding rate of protein. Further, we show how different kinds of features of amino acid residues can be combined with structural information to apply regression models to predict 3 types of properties of protein.

2. Materials and Methods

Throughout this study, protein molecules are processed as graphs. A graph $G(V, E)$ is a collection of V vertices and E edges. The entire framework from constructing these graphs to the processing's mathematical foundations is described in this section.

2.1. Protein Network Models

2.1.1. Residue Interaction Graph

A residue interaction graph (RIG) model of a protein is a simple network model of a protein. A RIG model for a protein is a graph, in which each vertex corresponds to a residue, and two residues are connected with an edge, if they lie in close proximity in their native-state structures. Given the three dimensional coordinates of each atom in a protein (available in the Protein Data Bank [4]), inter-residue distances are calculated. For consistency, center of a residue is considered to be its alpha carbon (C_α) atom [10]. Two vertices v_1 and v_2, with their corresponding C_α coordinates (x_1, y_1, z_1) and (x_2, y_2, z_2) are connected if the distance between the two is below or equal to a certain threshold r_c. In general, we have

$$w_{i,j} = \begin{cases} 1, & \text{if } d_{i,j} \leq r_c \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where r_c is the cutoff distance, and $d_{i,j} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}$. The cutoff r_c is carefully chosen so as to consider the required forces of attractions which indeed keep the protein structure stable. A suitable and meaningful cutoff r_c is usually chosen in the range 5-9 Å [21].

2.1.2. Weighted RIG model

While the RIG model captures the overall three-dimensional structure of the protein to some extent, it fails to highlight the interactions among residues lying distant apart along the backbone of the protein sequence. Such long-range interactions are the ones majorly responsible for holding the structure

intact and aids protein folding [9]. Long-range Interaction network have been studied previously in the context of protein folding kinetics [2]. Here, we use a modified model which includes both short and long-range interactions. It takes all edges of the RIG model into consideration and weights the edges such that long-distance edges have higher edge weights. The weights are proportional to the distance between the residues along the backbone of the protein. Thus, long-range interactions are given more importance. This model is referred to as the Weighted RIG ($wRIG$) model. The edge weights are calculated as -

$$wRIG_{i,j} = \begin{cases} |j-i|, & \text{if } d_{i,j} \leq r_c \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

This model is shown to have produced better results as compared to other models in the later sections.

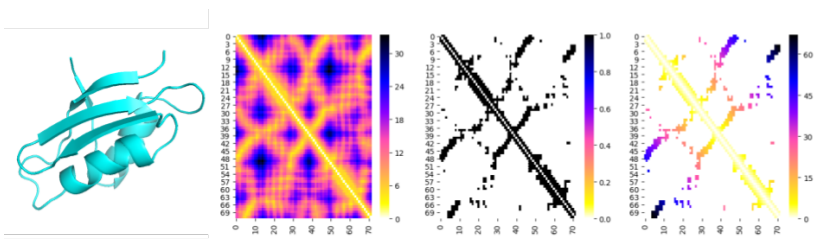

Figure 1: Construction of a weighted Residue Interaction Network for a protein. Starting form a stale tertiary structure

2.2. Spectral Graph Theory

Spectral graph theory is the study of the eigenvalues and eigenvectors of the matrices associated with the graph[6]. The (normalized) Laplacian matrix is often used for the purpose of graph signal processing, which is defined as

$$\mathcal{L}(u,v) = \begin{cases} 1 & \text{if } u = v \text{ and } d_v \neq 0, \\ -\frac{1}{\sqrt{d_u d_v}} & \text{if } u \text{ and } v \text{ are adjacent}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

where u, v are vertices in the network, and d_i is the total degree of the ith vertex. The Laplacian is also a square matrix like the adjacency matrix. The Laplacian matrix is a positive semi-definite matrix, hence it has a complete set of real and orthonormal eigenvectors. Moreover, since the matrix is symmetric with real values, it is diagonalizable. The eigendecomposition of $\mathcal{L}$ is given as $\mathcal{L} = U\Lambda U^{-1}$, where U and Λ are the eigen basis matrix and a diagonal matrix of eigenvalues of $\mathcal{L}$ respectively. Now, since U is the matrix of orthonormal eigenvectors, we have, $\mathcal{L} = U\Lambda U^T$. The corresponding eigenvalues are non-negative. The eigenvectors are denoted by $\{\mathbf{u}_l\}_{l=0,1,\ldots N-1}$. Zero occurs as an eigenvalue in multiplicity of the number of connected components of the graph. The eigenvectors of $\mathcal{L}$ are denoted by $\{\tilde{\mathbf{u}}_l\}_{l=0,1,\ldots N-1}$. The eigenvalues$\{\tilde{\lambda}_l\}_{l=0,1,\ldots N-1}$ of the normalized graph Laplacian matrix satisfy the condition $0 = \tilde{\lambda}_0 < \tilde{\lambda}_1 \leq \cdots \leq \tilde{\lambda}_{N-1} \leq 2$.

2.3. Graph Signal Processing

For a graph $G(V, E)$ with V vertices and E edges, a graph signal is represented as a vector $f \in \mathbb{R}^V$, such that the ith component of the vector represents the value of the signal at vertex v_i. A graph Fourier basis is obtained through the spectral decomposition of its Laplacian Matrix [17]. For a signal f defined on the vertices of a graph, its Graph Fourier Transform (GFT) is defined as

$$\hat{f} = U^{-1} f \tag{4}$$

Where U^{-1} is the *Graph Fourier transform matrix*. Since U is the matrix of orthonormal eigenvectors, $U^{-1} = U^T$. The values of $\hat{f}_n$of the signal's graph Fourier transform characterizes the frequency content of the of the signal as a projection on the eigen vector $\mathbf{u}_n$ [18].

The inverse graph Fourier transform (I-GFT) is given by

$$f = U\hat{f} \tag{5}$$

It constructs the original back from its frequency components. According to GSP theory, the transformed signal $\hat{f}$is analogous to the frequency domain transformation [18]. The components towards the smaller eigenvalues are considered informative while the high eigenvalue components capture noise.

In this study, various biophysical attributes of the residues, like their molecular weight, hydrophobicity etc are perceived as signals.

2.4. Implementation Details

Previous studies [2] have shown correlations between network parameters of various protein contact networks with the rate of folding of the proteins. GSP was used to validate these claims and check if residue signals(when visualized in frequency domain) have informative sections, which could be correlated to the rate of folding of the protein into consideration. In order to test our hypothesis, 52 single domain two-state folding proteins data was taken [11]. A complete list of these proteins, with their rate of folding information and protein class is given in Supplementary Table 1. With lower frequencies considered informative and higher frequencies as noise, absolute signal intensities in the lower eigenvalues were considered informative [19]. This informative fraction of signal is referred as Low Frequency Component (LFC)

$$LFC = \Sigma_{\hat{\lambda}_i=0}^{\hat{\lambda}_i \leq \tilde{\lambda}_{cutoff}} |\hat{f}_i| \tag{6}$$

The cutoff frequency $\hat{\lambda}_{cutoff}$ was varied from 0.01 to $\tilde{\lambda}_{max}$ in steps of 0.01 to search for the optimum cutoff $\tilde{\lambda}_{cutoff}$ which maximizes the correlation between LFC and $ln(k_f)$.

3. Results

3.1. Single feature correlation model

Residue hydrophobicity is well known for its role in protein folding [7]. With hydrophobicity signal fixed, we iterated with many variations of PCNs to maximize the correlations. Kyte and Doolittle hydrophobicity scale was taken to consider signal intensities at nodes [13]. The $\tilde{\lambda}_{cutoff}$ was optimised for maximum correlation between LFC and $\ln k_f$. The results obtained are summarized in Table 1. The weighted RIG model was seen to outperform other PCN models, with a correlation value of 0.62.

Table 1: Performance Summary of Various Protein Contact Network Models (Signal: Residue hydrophobicity)

S. No.	PCN Model	l_{cutoff}	Correlation observed	
1	Simple RIG	0.48	0.39	
2	RIG with distance as weights	0.27	0.34	
3	Weighted RIG	0.18	0.62	

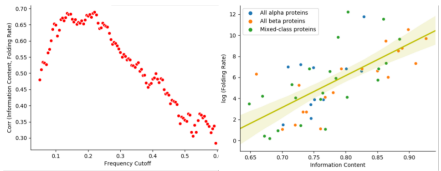

Figure 2: Weighted RIG model with residue hydrophobicity values as signals. (a) Scatter plot of correlation between rate of folding and low frequency component among proteins vs the frequency cutoff. The correlation is seen to hit a maximum value at frequency cutoff value 0.18. (b) Rate of folding vs low frequency component plot for 52 proteins, calculated at frequency cutoff = 0.18

3.2. Validation against random controls

The above-mentioned results were validated against random control networks. Different network models were considered which mimic completely arbitrary proteins in order to verify the results' sanctity. Multiple strategies or random network models were used for each protein's weighted RIG model, a random control was designed such that the number of nodes (or residues) remains the same while the connections are changed. The edge weights were also adjusted as being done for the weighted RIG model. Then, assuming the same hydrophobicity signal values on this weighted RC along with the same frequency cutoff, the correlations were computed. Each type of random control was made as close as possible to the original model as possible, in terms of the number of nodes and edges.

A detailed discussion on the methodology and implementation details per-

Random Control Model	Correlation
Complete Graph	0.07
Path Graph	-0.13
K-Regular Graph	0.22
Modified K-Regular Graph	-0.07

taining to random control based validation is present in Appendix 1.

3.3. Prediction using multiple GSP derived features

So far, the hypothesis was only based on residue's hydrophobicity value as a parameter for the protein folding-rate. Results in the previous sections also validate this hypothesis to a greater extent. Different signals were further used as features, and their overall effect on protein dynamics was studied. The signals chosen were both residue's properties, like their hydrophobicity, molecular weight etc, along with the network properties like node degree, node clustering coefficient etc. To analyze the effect of different signals on the rate of folding of proteins, a linear multivariable regression model [16] was defined. Let y be the dependent variable, and $x_1, x_2, \ldots, x_n$ are the n explanatory variables, a regression model tries to learn the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ such that $\hat{y}$ is approximated as a linear combination of the exploratory variables.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{7}$$

The independent predictor variables considered were critical in the regression analysis. In totality, 6 features were considered. For each of the signal, the frequency cutoff $\tilde{\lambda}_{cutoff}$ was taken to be the one which maximized the correlation between LFC and $\ln k_f$. The features, their corresponding $\tilde{\lambda}_{cutoff}$ and correlation values (ρ) are listed in Table 2.

Once the cutoff yielding the maximum correlation was observed, the corresponding fraction of information in the cutoff section of frequency (LFC) was calculated across proteins. This vector of an informative fraction of the signal in the frequency domain was taken as the predictors x_i. This exercise was carried

Table 2: Features considered for regression model, along with their respective frequency cutoff values and correlations observed

Signal	$\tilde{\lambda}_{cutoff}$	**Correlation**	
Hydrophobicity	0.27	0.63199	
Molecular Weight	0.2	0.73299	
Degree	0.17	0.6898	
Weighted Degree	0.17	0.55679	
Residue Frequency	0.22	0.59294	
Clustering Coefficient	0.26	0.68722	

out for each of the signal listed in Table 2. In this way, a feature vector for each protein was constructed. Throughout the regression analysis, the weighted RIG model was considered. The constructed feature matrix, along with folding rates, is presented in Supplementary Table 2. The proteins used in the regression model are compiled from multiple sources [11, 2]

The learned linear regression model yielded a multiple R-squared value of 0.55 (R=0.74), with an F-Statistic of 9.35 with a p-value of 1.18e-06. Even though the data points were less we performed cross-validation by randomly split 20 times in 60:40 ratio of test-train samples. For each split, LASSO (east absolute shrinkage and selection operator) based regression was used, which resulted in a 10-fold cross-validation R-square value of 0.55 (R=0.74).

An interesting thing to observe here was the contributions of variables independently on the rate of folding[14]. Molecular weight alone contributes to roughly 30 percent for the rate of folding, as shown in Supplementary Figure 1. Other important signals are the network properties of node clustering coefficient and node degree. Residue hydrophobicity, which according to our assumptions, is seen to contribute only about 15 percent. Molecular weight is directly related to the size of amino acids, the number of atoms it has and thereby governing the number of inter-atomic attractions it can handle. [8]

3.4. Alpha vs Beta proteins

Tertiary structural alterations in a protein changes its resultant network model significantly. This leads to a changed eigenbasis for signal transformation. In order to examine this, we processed separately processed alpha-helices and beta-sheet proteins. Proteins which exhibited both alpha-helix and beta sheets in their tertiary structure were kept separately in a third bucket. With common signal on all the set of models, the information quotient in low frequencies were calculated. In particular, node's clustering coefficient when used as signal resulted in a striking different LFC distribution along with groups as seen in Supplementary Figure 2. The mixed class features were seen to lie between the alpha and beta classes

Deeper analysis of the alpha and beta proteins validated the efficacy of using GSP derived features in protein class identification. All the features used in the regression model earlier were calculated for these proteins. To reduce the complexity, $\tilde{\lambda}_{cutoff}$ was chosen as an optimized value which was the same for each signal. The computed feature vectors (with an optimised $\tilde{\lambda}_{cutoff} = 0.42$) are provided in Supplementary Table 3. Alpha, mixed and beta class proteins were labelled 1, 0 and -1 respectively, and a regression model was learned on the data. A 10-fold cross-validation based LASSO regression model resulted in an R-squared value of 0.394 (R=0.627). Further, the mixed class proteins were omitted and the remaining proteins were used to fit another LASSO regression model. This yielded in a 10-fold cross-validated R-square value of 0.587 (R=0.766). The pairwise feature distribution (Figure 3) clearly shows a divide in GSP derived LFC values for different features in the two classes. The most relevant features contributing to this difference were seen to be clustering coefficient and residue molecular weight.

3.5. Transmembrane vs globular proteins

Differences in proteins' hydrophilic and hydrophobic orientation are also observed based on the location where they are present. Proteins present in the cytosol tend to be hydrophilic in order to easily remain in a stable state. These

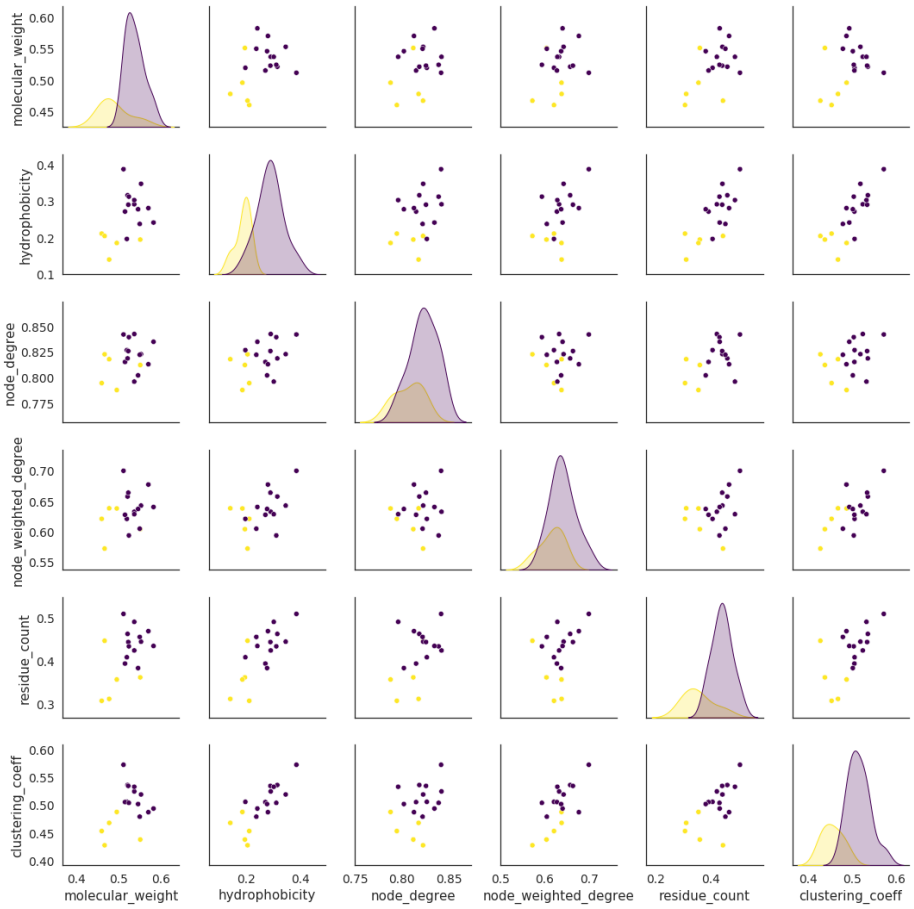

Figure 3: Pairwise GSP based feature distribution for alpha and beta proteins. Alpha and beta proteins are represented by yellow and blue dots respectively.

globular proteins have hydrophilic residues towards the outer side, protecting the hydrophobic ones in the interior side of the structure. On the contrary, transmembrane proteins are a part of the lipid-rich membrane, and are thus oriented in a such a way that the hydrophobic residues are towards the external side of the structure [15]. This difference in the structural placement of hydrophobic and hydrophilic residues was accessed using the proposed frameworks. We collected a set of 237 transmembrane and 59 globular proteins and constructed their weighted rig models. Residue hydrophobicity values were then

used as signals on these graphs, and information in the low frequency component was processed individually for the two groups.

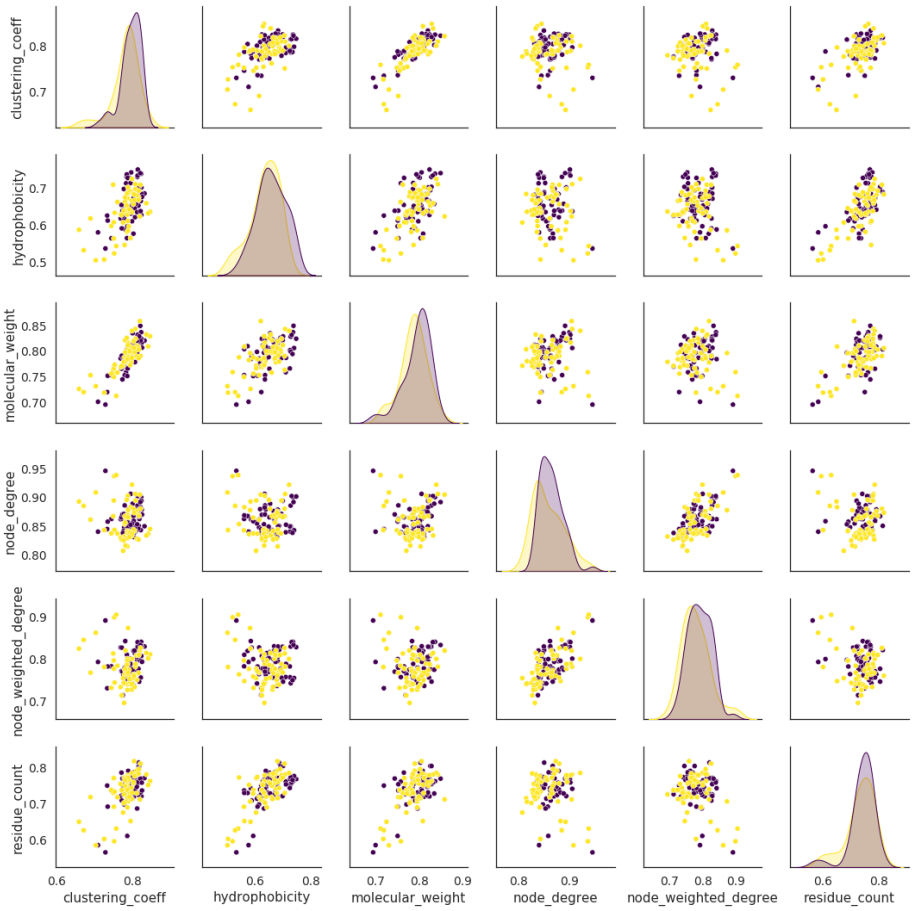

Figure 4: Pairwise GSP based feature distribution for transmembrane vs globular proteins. Yellow and blue points represent transmembrane and globular proteins respectively

Though both globular and transmembrane proteins had a similar median LFC, globular proteins tend to have a much shorter spread, as seen in Supplementary Figure 3.

A LASSO regression model similar to Alpha-Beta protein case study was defined for transmembrane and globular proteins. Imbalance in classes for handled by near-miss under-sampling of the over-represented group i.e. transmembrane

proteins [3]. Pairwise feature distribution indicated a greater overlap of the two classes, as seen in Figure 4. This resulted into a 10-fold cross-validated R-square value of 0.407 (R=0.638) at $\tilde{\lambda}_{cutoff} = 0.59$. Clustering coefficient and molecular weight were seen as the most important features, with -9.8 and 7.8 weights, respectively.

4. Discussion

Protein residue network can provide information to model many biophysical properties. However, most often using only graph-theoretical approach for analyzing protein residue network causes under-utilization of the information about residues' properties. On the other hand, graph signal processing allows the use of information (or signal) of residue properties hand in hand with a graph-theoretic approach. Here, we have shown that properties of nodes based on graph theory, such as centrality, clustering coefficient can also be used together with other features of amino-acids as signal with GSP based approach.

Using a set of 52 single-domain two-state folding proteins we generated various kinds of contact networks using their known structure. The contact was decided based on Euclidean distance cutoffs and residue distance cutoffs, and residue properties are taken as signals on it. Different features of amino acids in graph Fourier domain seem to be informative about rate of folding of the protein. Such as intensity in low-frequency components while using residue molecular-weight as signal is highly informative about rate of folding of protein. It also provides insight into the modularity of protein-residue network hand in hand with residue molecular weight distribution, which affects folding rate. It hints that when heavy residues get connected with each-other in to single hub, the folding rate is high. However if separate modules in protein fold independently of each other or heavy residues are well separated from each other the folding rate is low. We validated our findings with a number of random controls. In this way, it is observed that using GSP, signals on a contact network can be discerned into different segments, which govern the biophysical processes involved

in protein folding. A multi-variable linear regression model is also fitted, to see the contribution of various signals in folding time.

Biophysical property of protein actually emerge as a collective result of many factors. Our finding that frequency domain transformations extract informative segments with correlations with biophysical property can provide answers to unsolved puzzles regarding the distribution of amino-acids in the 3D structure of proteins. We also achieved good correlation between predicted and actual alpha/beta property ($R > 0.76$) and globularity ($R > 0.63$) of proteins using our approach of GSP and Lasso based regression model(with cross-validation). Our approach of amalgamating multiple types of signals on residue hand in hand with graph topology of proteins for machine-learning based modelling has rarely been used before. Thus our approach of extracting features using graph-Fourier transforms for down-stream machine learning approach also opens new avenues for feature extraction to predict multiple other properties of proteins. Hence in future, we would extend our approach for modelling other properties of proteins.

Availability

All the codes for graph signal processing-based analysis of protein are available at `https://github.com/divyanshusrivastava/Protein-GSP`.

conflict of Interest

There is no conflict of interest with authors

References

[1] Bagler, G. and Sinha, S. (2005). Network properties of protein structures. *Physica A: Statistical Mechanics and its Applications*, 346(1-2):27–33.

[2] Bagler, G. and Sinha, S. (2007). Assortative mixing in protein contact networks and protein folding kinetics. *Bioinformatics*, 23(14):1760–1767.

[3] Bao, L., Juan, C., Li, J., and Zhang, Y. (2016). Boosted near-miss undersampling on svm ensembles for concept detection in large-scale imbalanced datasets. *Neurocomputing*, 172:198–206.

[4] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242.

[5] Chakrabarty, B. and Parekh, N. (2016). Naps: Network analysis of protein structures. *Nucleic acids research*, 44(W1):W375–W382.

[6] Chung, F. R. and Graham, F. C. (1997). *Spectral graph theory*. Number 92. American Mathematical Soc.

[7] Dyson, H. J., Wright, P. E., and Scheraga, H. A. (2006). The role of hydrophobic interactions in initiation and propagation of protein folding. *Proceedings of the National Academy of Sciences*, 103(35):13057–13061.

[8] Finkelstein, A. V., Bogatyreva, N. S., and Garbuzynskiy, S. O. (2013). Restrictions to protein folding determined by the protein size. *FEBS letters*, 587(13):1884–1890.

[9] Go, N. and Taketomi, H. (1978). Respective roles of short-and long-range interactions in protein folding. *Proceedings of the National Academy of Sciences*, 75(2):559–563.

[10] Greene, L. H. (2012). Protein structure networks. *Briefings in functional genomics*, 11(6):469–478.

[11] Gromiha, M. M., Thangakani, A. M., and Selvaraj, S. (2006). Fold-rate: prediction of protein folding rates from amino acid sequence. *Nucleic acids research*, 34(suppl_2):W70–W74.

[12] Jiang, X., Chen, C., and Xiao, Y. (2010). Improvements of network approach for analysis of the folding free-energy surface of peptides and proteins. *Journal of computational chemistry*, 31(13):2502–2509.

[13] Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132.

[14] Lindeman, R. H. (1980). Introduction to bivariate and multivariate analysis. Technical report.

[15] Nelson, D. L., Lehninger, A. L., and Cox, M. M. (2008). *Lehninger principles of biochemistry*. Macmillan.

[16] Ross, S. M. (2004). *Introduction to probability and statistics for engineers and scientists*. Elsevier.

[17] Sandryhaila, A. and Moura, J. M. (2014). Discrete signal processing on graphs: Frequency analysis. *IEEE Transactions on Signal Processing*, 62(12):3042–3054.

[18] Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., and Vandergheynst, P. (2013). The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98.

[19] Srivastava, D. and Kumar, V. (2018). *Graph signal processing based analysis of biological networks*. PhD thesis, IIIT-D.

[20] Szilágyi, A., Nussinov, R., and Csermely, P. (2013). Allo-network drugs: extension of the allosteric drug concept to protein-protein interaction and signaling networks. *Current Topics in Medicinal Chemistry*, 13(1):64–77.

[21] Yan, W., Zhou, J., Sun, M., Chen, J., Hu, G., and Shen, B. (2014). The construction of an amino acid network for understanding protein structure and function. *Amino acids*, 46(6):1419–1439.