Physica A 615 (2023) 128603

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

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

Divyanshu Srivastava, Ganesh Bagler, Vibhor Kumar *

Department for Computational Biology, Indraprastha Institute of Information Technology, Delhi, India

ARTICLE INFO

Article history:
Received 2 January 2021
Received in revised form 15 February 2022
Available online 24 February 2023

Dataset link: https://github.com/divyanshusrivastava/Protein-GSP

Keywords:
Graph signal processing
Residue interaction graph
Graph Fourier transform

ABSTRACT

Understanding the physical and chemical properties of proteins is vital, and many efforts have been made to study the emergent properties of the macro-molecules 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 nodes 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.

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 [1]. 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 [2–4]. 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 [3]. 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 [5,6]. Similarly, the folding rate of protein has also been modelled using network properties based on only graph-theoretic approach [7]. 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.

* Corresponding author.
E-mail address: vibhor@iiitd.ac.in (V. Kumar).

https://doi.org/10.1016/j.physa.2023.128603
0378-4371/

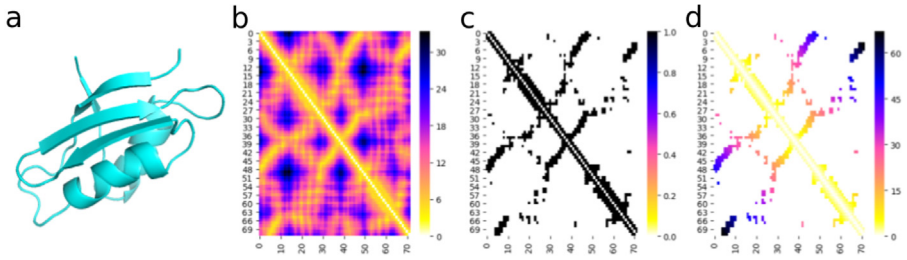

Fig. 1. Construction of a weighted Residue Interaction Network for a protein (PDB ID : 2HQI). (a) A stable tertiary structure of a protein (b) the inter-residue Euclidean distances and all-vs-all contact map (c) The contact map is a heat map of the inter-residue distances of the protein, and is analogous to the network's adjacency matrix. The distance cutoff (r_c) is chosen as 8 Å in this example, and an unweighted RIG_{binary} model is created (d) Sequence based inter-residue distance as weights are added to the unweighted RIG_{binary} model according to distance in sequence of protein, to obtain RIG_{seq} model's adjacency matrix.

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 simple RIG, or $RIG_{spatial}$ 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 spatial proximity in their native-state structures. Given the three-dimensional coordinates of each atom in a protein (available in the Protein Data Bank [8]), inter-residue Euclidean distances are calculated (Fig. 1b). For consistency, the centre of a residue is considered to be its alpha carbon (C_α) atom [9]. 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

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

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 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 [4].

A slight variant of the $RIG_{spatial}$ model is the RIG_{binary} model, which is a binarized version of the $RIG_{spatial}$ model. The model's weights are defined as

$$w_{i,j} = \begin{cases} 1, & \text{if } d_{i,j} \leq r_c \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Fig. 1c shows the adjacency matrix for RIG_{binary} model of a human protein (PDB ID: 2HQI), where $r_c = 8$ Å.

2.1.2. Sequence distance based 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 [10]. 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 (Fig. 1d). 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 RIG_{seq} model. The edge weights are calculated as -

$$(RIG_{seq})w_{i,j} = \begin{cases} |j-i|, & \text{if } d_{i,j} \leq r_c \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

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

2.2. Spectral graph theory

Spectral graph theory is the study of the eigenvalues and eigenvectors of the matrices associated with the graph [11]. The (normalized) Laplacian matrix is often used for the purpose of graph signal processing, which is defined as

$$\mathcal{L}(u,v) = \begin{cases} 1 & \text{if } u = v \text{ and } D_v \neq 0, \\ -\frac{1}{\sqrt{D_u D_v}} & \text{if } u \text{ and } v \text{ are adjacent,} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

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 [12]. For a signal f defined on the vertices of a graph, its Graph Fourier Transform (GFT) is defined as

$$\hat{f} = U^{-1}f \tag{5}$$

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 signal as a projection on the eigenvector $\mathbf{u}_n$ [13].

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

$$f = U\hat{f} \tag{6}$$

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 [13]. Since we use Laplacian of graph, the components towards the smaller eigenvalues represent low frequency while the high eigenvalue components represent higher frequency [13]. The higher frequency components are often considered to be dominated by 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 (PCN) 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. With lower frequencies considered informative and higher frequencies as noise, absolute signal intensities in the lower eigenvalues were considered informative [14]. This informative fraction of signal is referred as Low Frequency Component (*LFC*)

$$LFC = \Sigma_{\hat{\lambda}_i=0}^{\hat{\lambda}_i \leq \tilde{\lambda}_{cutoff}} |\hat{f}_i| \tag{7}$$

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 Pearson's correlation between *LFC* and $ln(k_F)$, where k_F is the rate of folding of the protein. The threshold $(\tilde{\lambda}_{cutoff})$ to determine the low-frequency component represents the width of the neighbourhood of co-occurrence of a property of connected residues in protein residue-network, which is correlated with folding rate. The RIG_{seq} distance cutoff (r_c) was optimized to maximize the correlation, and 7.3 Å was fixed for further processing (Fig. 2). While using regression

Table 1
Performance Summary of different Residue Interaction Graph (RIG) models.

S. No.	PCN model	$\tilde{\lambda}_{cutoff}$	Correlation observed
1	RIG_{binary}	0.27	0.34
2	$RIG_{spatial}$	0.48	0.39
3	RIG_{seq}	0.27	0.63

Signal: Residue hydrophobicity.

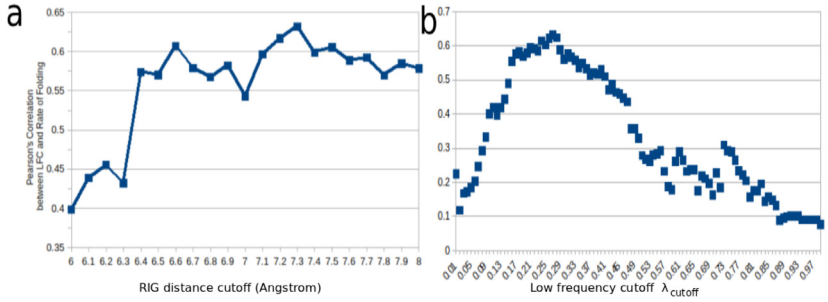

Fig. 2. Optimizing the parameters for the RIG_{seq} model with residue hydrophobicity values as signals. (a) The Euclidean distance cutoff r_c was varied from 6 to 8 Å. The corresponding RIG_{seq} models were constructed, and for each model, the LFC cutoff was varied to obtain maximum correlation between low frequency component and folding rate. An optima were obtained at $r_c = 7.3$ Å. (b) Scatter plot of correlation vs low frequency component cutoff for best performing RIG_{seq} model ($r_c = 7.3$ Å). The correlation is seen to hit a maximum value of 0.63 at $\tilde{\lambda}_{cutoff} = 0.27$.

model for a protein property we used same ($\tilde{\lambda}_{cutoff}$) for amino acid features, in order to reduce number of parameters to be optimized. After finding the optimal ($\tilde{\lambda}_{cutoff}$) using all the protein structures meant to model a property, we did not change it during cross-validation based test using regression model. However, for different properties of protein we had different values for optimal ($\tilde{\lambda}_{cutoff}$). Hence the value of ($\tilde{\lambda}_{cutoff}$) for modelling transmembrane-globular property was different from folding rate prediction.

2.5. Data sources

In order to test our hypothesis, 52 single domain two-state folding proteins data was taken. In order to reduce model complexity, the proteins were specifically chosen to be single domain simple structures, with credible folding rate information available. Moreover, only those proteins were considered whose folding mechanism is in two-states only, without undergoing any intermediate folded structure. Such proteins were gathered from multiple sources [2,15]. A complete list of these proteins, with their rate of folding information and protein family (α, β and $\alpha\beta$ proteins), is given in Supplementary Table 1. Same list of 52 proteins were used for modelling α/β property.

3. Results

3.1. Single feature correlation model

Residue hydrophobicity is well known for its role in protein folding [16]. With hydrophobicity signal fixed, we iterated with many variations of RIG models to maximize the correlations. Kyte and Doolittle hydrophobicity scale for amino-acids was used as signal intensities at nodes [17]. The $\tilde{\lambda}_{cutoff}$ was optimized for maximum correlation between *LFC* and $\ln k_F$. As shown in Fig. 2b, the accumulation of the low-frequency components till certain $\tilde{\lambda}_{cutoff}$ provides maximum correlation using hydrophobicity value (see Supplementary Figure 1). It indicates that when hydrophobic residues interact with each other and form a large connected group, the folding rate is high. However, when they make smaller (below cutoff) connected groups, it has a negative impact on the folding rate. The results obtained are summarized in Table 1. The RIG_{seq} model was seen to outperform other RIG models, with a correlation value of 0.63.

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 RIG_{seq} 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

Table 2
Performance Summary for modelling folding rate when other graph is used for protein.

Random control model	Correlation
Complete graph	0.07
Path graph	−0.13
K-Regular graph	0.22
Modified K-Regular graph	−0.07

Signal: Residue hydrophobicity.

Table 3
Features considered for regression model, along with their respective frequency cutoff values and correlations observed.

Signal	$\tilde{\lambda}_{cutoff}$	Correlation
Hydrophobicity	0.27	0.63
Molecular weight	0.2	0.73
Degree	0.17	0.69
Weighted degree	0.17	0.56
Residue frequency	0.22	0.59
Clustering coefficient	0.26	0.69
Betweenness centrality	0.17	0.69
Page rank	0.17	0.65

the RIG_{seq} model. Then, assuming the same hydrophobicity signal values on this weighted RIG (RIG_{seq}) along with the same frequency cutoff, the correlations were computed. The correlation values for each random control are listed in Table 2. 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 pertaining to random control based validation is present in Appendix 1 (supplementary Material).

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, centrality, page rank etc. To analyse the effect of different signals on the rate of folding of proteins, a linear multivariable regression model [18] 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.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{8}$$

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 3.

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 out for each of the signal listed in Table 3. In this way, a feature vector for each protein was constructed. Throughout the regression analysis, the RIG_{seq} 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 [2,15]

The learned linear regression model yielded a multiple R-squared value of 0.6 (R=0.78). We performed cross-validation by randomly split of the data in 5-fold test-train samples. For each split, LASSO (Least Absolute Shrinkage and Selection Operator) based regression was used, which resulted in a 5-fold cross-validation R-square value of 0.56 (R=0.75). Our GFT based method capturing node-level properties outperforms previously reported method using holistic network properties of similarly constructed protein network models for folding rate prediction [2]. The multi-feature regression model predicted the folding rate with root-mean-square error of 1.92 and mean-absolute-error of 1.6. A comparison of other protein folding rate prediction tools done by another group is used to compare different methods [19]. As per the error reported by other methods, our approach is seen to perform better than other models [20–22] (Supplementary Table 3).

An interesting thing to observe here were the contributions of variables independently on the rate of folding [23]. Molecular weight alone contributes to roughly 20 percent of variance explained (R-square) for folding-rate prediction,

as shown in Supplementary Figure 2. 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 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 [24].

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 3. 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. Alpha, mixed and beta class proteins were labelled 1, 0 and -1 respectively, and a regression model was learned on the data. A 5-fold cross-validation based LASSO regression model resulted in an R-squared value of 0.39 ($R = 0.62$). Further, the mixed class proteins were omitted, and the remaining proteins were used to fit another LASSO regression model. This yielded in a 5-fold cross-validated R-square value of 0.63 ($R = 0.79$). The pairwise feature distribution (Fig. 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 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 [1]. This difference in the structural placement of hydrophobic and hydrophilic residues was assessed using the proposed frameworks. We collected a set of 237 transmembrane and 59 globular proteins and constructed their RIG_{seq} models. All the 8 signals were used on the models and the feature matrix was curated. This feature matrix was used to train a Logistic Regression based classification model.

The feature distribution shown in supplementary Figure 4 for both the classes shows some variability in the two sets of proteins. Pairwise feature distribution indicated a greater overlap of the two classes, as seen in supplementary Figure 4. However we utilized the power of multiple features for classification. A classification model similar to the Alpha-Beta protein case study was defined for transmembrane and globular proteins. Imbalance in classes was handled by near-miss under-sampling of the over-represented group, i.e. transmembrane proteins [25]. This resulted into a 10-fold cross-validated accuracy value of 0.813, AUC score = 0.92 and F1 score of 0.822 at $\tilde{\lambda}_{cutoff} = 0.59$, as depicted in Fig. 4. Supplementary Figure 5 shows cross-validation based results for modelling the properties of protein.

4. Discussion

Protein residue network can provide information to model many biophysical properties. However, most often using only graph-theoretical approach for analysing 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. 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. 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. For signal of hydrophobicity amino-acids, we have used Kyte and Doolittle scale, which is based on both experimentally measured water-vapour transfer free energies for amino acids and observed interior–exterior distribution of their side-chains [17]. Kyte and Doolittle scale of hydrophobicity has also been widely used for prediction of membrane protein topology and for folding prediction [26]. However, other hydrophobicity scales for amino acids [27] could also provide similar correlation with certain property.

For simplicity, the information in graph Fourier domain can be understood in terms of sizes of components(modules) made of connected nodes with same property. When a large number of nodes connected with each other have same feature (such as high hydrophobicity or high molecular weight) then we have higher values for low-frequency coefficients. On the other hand, if the distribution of property of amino acids on the residue interaction graph is grainy (smaller components or modules) the value of high frequency component is high. However, if, it is too grainy then often it

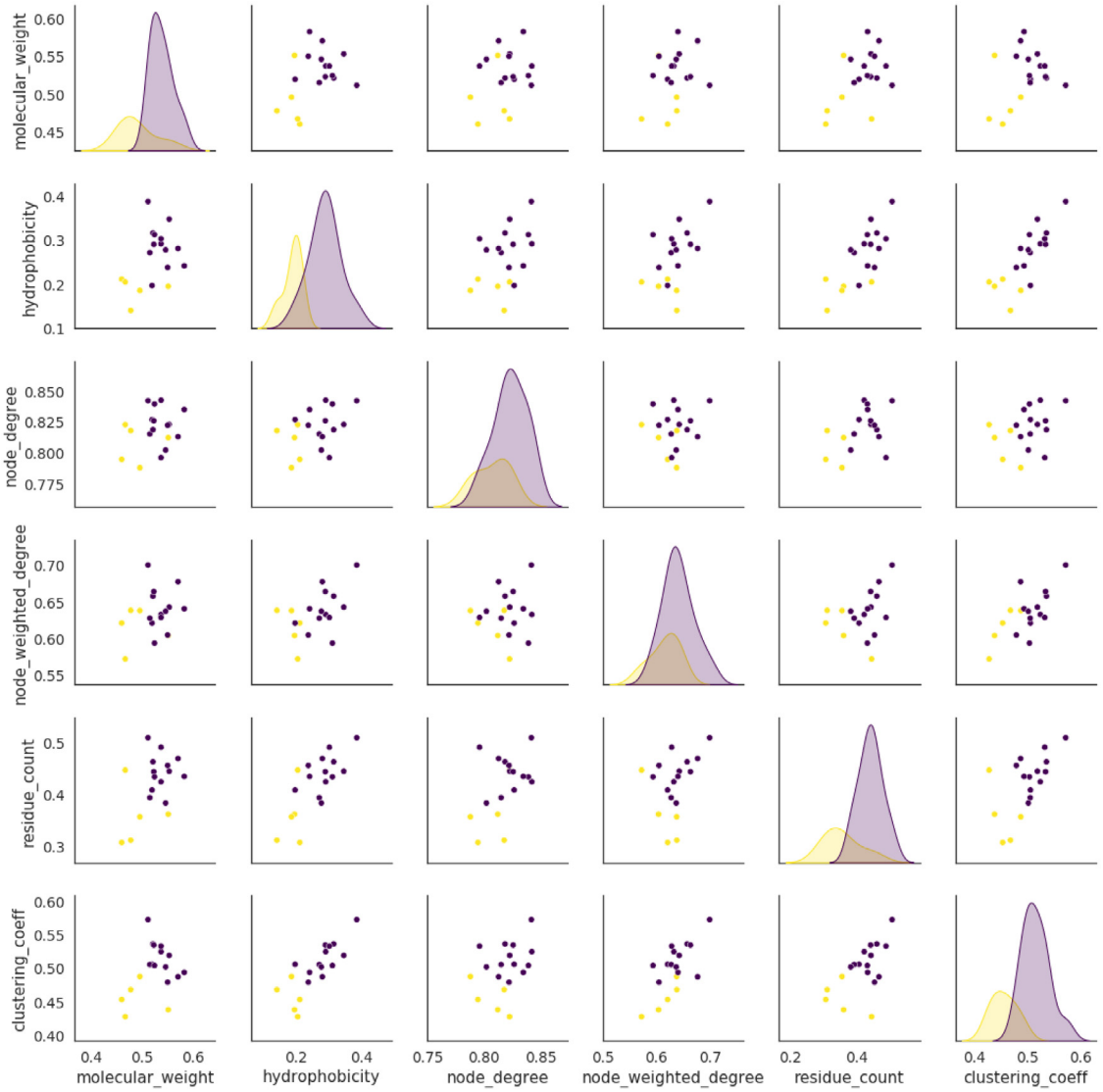

Fig. 3. Pairwise GSP based feature distribution for alpha and beta proteins. Alpha and beta proteins are represented by yellow and blue dots, respectively.

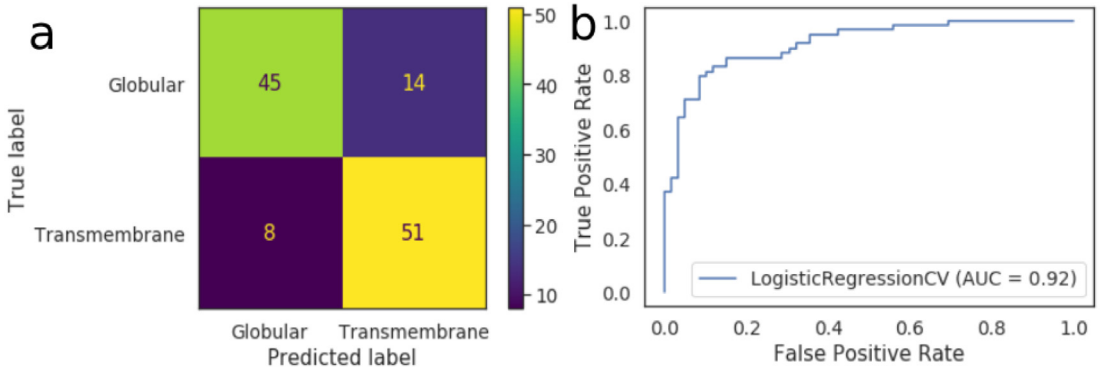

Fig. 4. GSP derived features based Transmembrane vs Globular proteins' classification, using 10-fold cross validation on Logistic regression. (a) Confusion Matrix of the classifier. (b) Receiver Operating Characteristic (ROC) Curve for the classification.

would overlap with noise. Different properties of amino acids in graph-Fourier domain seem to be informative about rate of folding of the protein. Such as sum of intensities of low-frequency graph-Fourier coefficients while using residue molecular-weight as signal is correlated with 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. Higher frequency components in graph signal represent more graininess or more smaller local modules in residue interaction graph. It hints that when heavy residues get connected with each-other in to single hub (or fewer larger modules or connected-components), the folding rate is high. We provide higher weightage for interaction among residues lying far from each other in sequence. Thus, our results hint that interaction among residues with higher molecular weight and lying far from each other, leads to higher folding rate. However, if many small modules in protein fold locally (in sequence) of each other and there is less distal interaction among higher molecular-weight residues, the folding rate is low.

Biophysical property of protein actually emerge as a collective result of many factors. Therefore we used multi-variable linear regression model is also fitted, to see the contribution of various signals in folding rate. Using lasso based regression models (with cross-validation) we also achieved good correlation between predicted and actual alpha/beta property ($R > 0.76$) and globularity ($R > 0.63$). 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. Due to scarcity of flexibility information of proteins used in our model for prediction, we have assumed fixed coordinates provided in PDB structure like previous studies using only graph theory [3,7]. Residue interaction graph could also be adopted to include flexibility information of protein structure if it is available. However in that case we may have to make two version of edge weights (or two kinds of graph): one corresponding to probability of interaction among residue accounting for flexibility and other based on distance in sequence. Combining the information in Fourier spectrum for both kinds of graphs for a protein could help in achieving better predictive model, which could partially include flexibility information. However, it was out of scope of our current strategy due to non-availability of flexibility information. Overall, 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.

CRediT authorship contribution statement

Divyanshu Srivastava: Data curation, Coding, Testing, Writing – original draft. **Ganesh Bagler:** Quality control, Supervision, Revision. **Vibhor Kumar:** Conceptualization, Methodology, Supervision, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

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

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.physa.2023.128603.

References

[1] David L. Nelson, Albert L. Lehninger, Michael M. Cox, Lehninger principles of biochemistry, Macmillan, 2008.
[2] Ganesh Bagler, Somdatta Sinha, Assortative mixing in protein contact networks and protein folding kinetics, Bioinformatics 23 (14) (2007) 1760–1767.
[3] Broto Chakrabarty, Nita Parekh, NAPS: Network analysis of protein structures, Nucleic Acids Res. 44 (W1) (2016) W375–W382.
[4] Wenying Yan, Jianhong Zhou, Maomin Sun, Jiajia Chen, Guang Hu, Bairong Shen, The construction of an Amino acid network for understanding protein structure and function, Amino acids 46 (6) (2014) 1419–1439.
[5] András Szilágyi, Ruth Nussinov, Péter Csermely, Allo-network drugs: Extension of the allosteric drug concept to protein-protein interaction and signaling networks, Curr. Top. Med. Chem. 13 (1) (2013) 64–77.
[6] Xuewei Jiang, Changjun Chen, Yi Xiao, Improvements of network approach for analysis of the folding free-energy surface of peptides and proteins, J. Comput. Chem. 31 (13) (2010) 2502–2509.
[7] Ganesh Bagler, Somdatta Sinha, Network properties of protein structures, Physica A: Stat. Mech. Appl. 346 (1–2) (2005) 27–33.
[8] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, Philip E Bourne, The protein data bank, Nucleic Acids Res. 28 (1) (2000) 235–242.
[9] Lesley H. Greene, Protein structure networks, Brief. Funct. Genom. 11 (6) (2012) 469–478.
[10] Nobuhiro Go, Hiroshi Taketomi, Respective roles of short-and long-range interactions in protein folding, Proc. Natl. Acad. Sci. 75 (2) (1978) 559–563.
[11] Fan R.K. Chung, Fan Chung Graham, Spectral Graph Theory, no. 92, American Mathematical Soc., 1997.

[12] Aliaksei Sandryhaila, Jose M.F. Moura, Discrete signal processing on graphs: Frequency analysis, IEEE Trans. Signal Process. 62 (12) (2014) 3042–3054.
[13] David I Shuman, Sunil K Narang, Pascal Frossard, Antonio Ortega, Pierre Vandergheynst, The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains, IEEE Signal Process. Mag. 30 (3) (2013) 83–98.
[14] Divyanshu Srivastava, Vibhor Kumar, Graph signal processing based analysis of biological networks (Ph.D. thesis), IIIT-D, 2018.
[15] M. Michael Gromiha, A. Mary Thangakani, Samuel Selvaraj, FOLD-RATE: Prediction of protein folding rates from Amino acid sequence, Nucleic Acids Res. 34 (suppl_2) (2006) W70–W74.
[16] H. Jane Dyson, Peter E. Wright, Harold A. Scheraga, The role of hydrophobic interactions in initiation and propagation of protein folding, Proc. Natl. Acad. Sci. 103 (35) (2006) 13057–13061.
[17] Jack Kyte, Russell F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1) (1982) 105–132.
[18] Sheldon M. Ross, Introduction to Probability and Statistics for Engineers and Scientists, Elsevier, 2004.
[19] Catherine Ching Han Chang, Beng Ti Tey, Jiangning Song, Ramakrishnan Nagasundara Ramanan, Towards more accurate prediction of protein folding rates: A review of the existing web-based bioinformatics approaches, Brief. Bioinform. 16 (2) (2015) 314–324.
[20] Hong-Bin Shen, Jiang-Ning Song, Kuo-Chen Chou, et al., Prediction of protein folding rates from primary sequence by fusing multiple sequential features, J. Biomed. Sci. Eng. 2 (03) (2009) 136.
[21] Chou Kuo-Chen, Shen Hong-Bin, FoldRate: A web-server for predicting protein folding rates from primary sequence, Open Bioinform. J. 3 (1) (2009).
[22] Xiang Cheng, Xuan Xiao, Zhi-cheng Wu, Pu Wang, Wei-zhong Lin, Swfoldrate: Predicting protein folding rates from Amino acid sequence with sliding window method, Proteins: Struct. Funct. Bioinform. 81 (1) (2013) 140–148.
[23] Richard Harold Lindeman, Introduction to bivariate and multivariate analysis, Technical report, 1980.
[24] Alexei V Finkelstein, Natalya S Bogatyreva, Sergiy O Garbuzynskiy, Restrictions to protein folding determined by the protein size, FEBS Lett. 587 (13) (2013) 1884–1890.
[25] Lei Bao, Cao Juan, Jintao Li, Yongdong Zhang, Boosted near-miss under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets, Neurocomputing 172 (2016) 198–206.
[26] Mauro Degli Esposti, Massimo Crimi, Giovanni Venturoli, A critical evaluation of the hydropathy profile of membrane proteins, Eur. J. Biochem. 190 (1) (1990) 207–219.
[27] Stefan Simm, Jens Einloft, Oliver Mirus, Enrico Schleiff, 50 Years of Amino acid hydrophobicity scales: Revisiting the capacity for peptide classification, Biol. Res. 49 (1) (2016) 1–19.