Computers in Biology and Medicine 155 (2023) 106436

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Computational prediction of protein folding rate using structural parameters and network centrality measures

Saraswathy Nithiyanandam [a], Vinoth Kumar Sangaraju [b], Balachandran Manavalan [b,*], Gwang Lee [a,c,**]

[a] Department of Molecular Science and Technology, Ajou University, 206 World Cup-ro, Suwon, 16499, South Korea
[b] Department of Physiology, Ajou University School of Medicine, 206 World Cup-ro, Suwon, 16499, South Korea
[c] Computational Biology and Bioinformatics Laboratory, Department of Integrative Biotechnology, College of Biotechnology and Bioengineering, Sungkyunkwan University, Suwon, 16419, Gyeonggi-do, South Korea

ARTICLE INFO

Keywords:
Two-state
Non-two-state
Folding type
Logarithmic folding rate
Machine learning algorithms
Support vector machine
Network centrality measures
Structural parameters

ABSTRACT

Protein folding is a complex physicochemical process whereby a polymer of amino acids samples numerous conformations in its unfolded state before settling on an essentially unique native three-dimensional (3D) structure. To understand this process, several theoretical studies have used a set of 3D structures, identified different structural parameters, and analyzed their relationships using the natural logarithmic protein folding rate ($\ln(k_f)$). Unfortunately, these structural parameters are specific to a small set of proteins that are not capable of accurately predicting $\ln(k_f)$ for both two-state (TS) and non-two-state (NTS) proteins. To overcome the limitations of the statistical approach, a few machine learning (ML)-based models have been proposed using limited training data. However, none of these methods can explain plausible folding mechanisms. In this study, we evaluated the predictive capabilities of ten different ML algorithms using eight different structural parameters and five different network centrality measures based on newly constructed datasets. In comparison to the other nine regressors, support vector machine was found to be the most appropriate for predicting $\ln(k_f)$ with mean absolute differences of 1.856, 1.55, and 1.745 for the TS, NTS, and combined datasets, respectively. Furthermore, combining structural parameters and network centrality measures improves the prediction performance compared to individual parameters, indicating that multiple factors are involved in the folding process.

1. Introduction

The "functional native structure" of a protein is crucial for understanding the biological functions of proteins, as well as for forming complexes with other molecules or proteins for structural and regulatory processes. The mechanisms of protein folding remain one of the most challenging problems to solve; few of these mechanisms actually untangle the second half of the genetic code. Protein folding occurs in a hierarchical order to produce a stable native structure. Practically, a protein domain attempting to find the native state by randomly traversing all possible interactions would require more time than the entire universe to complete. However, in reality, most protein domains spontaneously fold into their native state in the order of 10^{-6}–10^{-1} s. Protein folding refers to how unfolded proteins are folded into their native form, and this process is studied via thermodynamics and kinetics [1,2].

There are two main characteristics of protein folding: the first is the kinetic order, which determines whether the protein reaches its native structure through one intermediate or multiple intermediates [5]; the second is the rate constant, which varies from nanoseconds to hours, depending on the protein length [6]. In general, smaller proteins tend to fold faster than larger ones [3,4]. Although smaller proteins with 100 or fewer amino acids have simple two-state (TS) kinetics [5,6], larger proteins with over 100 amino acids have non-two-state (NTS) kinetics with stable intermediates [7]. The rule generally holds true, but changes in the experimental conditions can cause it to work in a different way. Furthermore, failure to fold into the native state results in misfolded or aggregated proteins, which can lead to several neurodegenerative disorders including Alzheimer’s disease, Parkinson’s disease, Huntington’s disease, and cystic fibrosis [9].

* Corresponding author.
** Corresponding author. Department of Molecular Science and Technology, Ajou University, 206 World Cup-ro, Suwon, 16499, South Korea.
E-mail addresses: bala2022@skku.edu (B. Manavalan), glee@ajou.ac.kr (G. Lee).

https://doi.org/10.1016/j.compbiomed.2022.106436
Received 17 August 2022; Received in revised form 28 November 2022; Accepted 13 December 2022
Available online 15 February 2023

Protein folding refers to the process of bringing an unfolded protein domain into its native, compact three-dimensional (3D) structure [11, 12]. In the last two decades, predicting the protein-folding rate from its linear chain sequence has become a popular field of research. Furthermore, a strong correlation was established between the folding rate and native structure. Accordingly, topology packing has proven to be the most reliable component for understanding protein folding and structure. Using a simple optimistic method, extensive research has found that protein folding rate is related to structural factors [13–15]. Owing to their unique perspective, protein folding correlations between protein folding rate and its fundamental properties have been studied for more than two decades [5,13,16,17]. Nevertheless, various studies based on protein 3D structures have also been performed to help understand protein folding rates.

Identifying novel structural parameters using 3D structures and studying their linear relationship with the logarithmic folding rate (ln (k_f)) gained popularity in the late 2000s. Plaxco and colleagues [18] proposed the first structure-based parameter, namely the relative contact order (RCO) of TS proteins and their relationship with ln(k_f). This study inspired several researchers to develop various structural parameters based on TS, namely the absolute contact order (ACO) [19], chain length (N_{res}) [20], size-modified contact order (SMCO) [19], native state geometry (cliquishness) [21], amino acid properties (P_{av}) [17], effective chain length (L_{eff}) [22], long-range order (LRO) [14], effective contact order (ECO) [23], combination of contact order and stability [27], topological properties (L_{pre}, L_{post}) [24], fraction of local contacts (FLCO), number of sequence-distant native pairs (Q_d) [6], chain topology parameter (CTP) [25], total contact distance (TCD) [15], and long-range contact order (LRCO) [26].

Kamagata et al. [27] proposed the first structural parameter based on NTS proteins, called non-local contact clusters, which showed a strong correlation with ln(k_f). Generally, the above-mentioned structural parameters are derived from a small set of TS and NTS proteins. Notably, none of these parameters are suitable for accurately identifying ln(k_f) for both types of proteins. Consequently, machine learning (ML)-based methods have been developed using a minimal amount of data [28], but they have not been able to explain possible folding mechanisms.

In this study, we constructed a TS and NTS dataset based on a recent PFDB database. Firstly, the relationship of eight structural parameters (CTP, TCD, ACO, RCO, FLCO, LRO, LRCO, and N_{res}) with TS, NTS, and combined (TS + NTS) were analyzed to determine whether the reported performance of each parameter is still relevant when applied to a larger dataset. Second, large-scale machine learning regressors (10 different algorithms) were employed to predict ln(k_f) using structural parameters and network-based parameters, including betweenness centrality (C_B), betweenness centrality of an edge (C_E), eigen centrality (E_C), closeness centrality (C_C), and degree centrality (C_D). The results indicate that combining both structural and network parameters with support vector machine (SVM) achieves the best performance regardless of the dataset, indicating that several factors contribute to the prediction of ln(k_f), in contrast to the specific structural parameters mentioned in previous studies.

2. Methods

2.1. Datasets

Recently, Manavalan et al. [29] reported a larger collection of TS and NTS. This study utilized the PFDB database, which contains 141 globular proteins, and their ln(k_f) values were reported at 25 °C. Of the 141 proteins, 89 belonged to the TS and the remaining 52 belonged to the NTS. In terms of structural classes, all- α: 24; all-β: 42; $\alpha/\ \beta$ or $\alpha+\ \beta$: 23 for TS proteins; and all-α: 10; all-β:13; α/β or $\alpha+\beta$: 29 for NTS proteins. The three-dimensional (3D) structure of each protein was downloaded from the Protein Data Bank (PDB) [30]. These data can be downloaded from http://lee.kias.re.kr/~bala/PFDB.

2.2. Structural parameters

We employed seven topological parameters, extracted this information from the 3D structure, and studied their relationship with ln(k_f). The seven structural parameters were as follows.

2.2.1. Relative contact order

RCO is defined as the average sequence distance between all pairs of residues in contact, normalized by the length of the entire sequence [13, 14]. This reflects the relative importance of the local and nonlocal contacts of a protein in a 3D structure.

$$RCO=\frac{1}{K.Nc}\sum_{\substack{s=1\\ |p-q|>2}}^{n}\Delta Apq \tag{1}$$

Unless otherwise noted, ΔApq is the heavy atoms distance (<7 Å), K is the number of residues, p and q are the sequence distances of interacting residues, and Nc is the total number of contacts.

2.2.2. Absolute contact order

ACO is the normalization of relative contact order by sequence length [19,31]. This calculates the average distance of both the local and nonlocal contacts.

$$ACO=\frac{1}{Nc}\sum_{\substack{s=1\\ |p-q|>2}}^{n}\Delta Apq \tag{2}$$

2.2.3. Total contact distance

TCD is the integration of CO and LRO [15] and is described as follows:

$$TCD=\frac{1}{K^2}\sum_{\substack{s=1\\ |p-q|>0}}^{n}\Delta Apq \tag{3}$$

where ΔApq is the distance between the heavy atoms of the contact pair of residues (<8 Å). Notably, the length of the residues is squared compared to the long-range order.

2.2.4. Chain topology parameter

CTP considers both local and nonlocal contacts [25], which are calculated as follows:

$$CTP=\frac{1}{K.Nc}\sum_{\substack{s=1\\ |p-q|>2}}^{n}\Delta Apq^2 \tag{4}$$

where ΔApq is a heavy atoms distance of <7 Å, and the sequence separation is squared compared with CO.

2.2.5. Fraction of local contact

FLCO is based on short-range contacts [32] and is described as follows:

$$FLCO=\frac{\sum_{|p-q|\le 4}\Delta(p,q)}{\sum_{|p-q|>0}\Delta(p,q)} \tag{5}$$

where $\Delta(p,\ q)=1$ if the heavy-atoms distance of the pair of residues is < 5 Å. If not, then $\Delta\ (p,\ q)=0$.

2.2.6. Long-range order

LRO is defined as a pair of residues $C\alpha$ atoms that are far apart in the primary sequence but close in 3D-space [14], which is described as:

$$LRO = \sum \frac{A_{pq}}{K}, \begin{cases} A_{pq} = 1 \ if \ |p-q| > 12 \\ A_{pq} = 0 \ otherwise \end{cases} \tag{6}$$

where p and q are two interacting residues whose $C\alpha$-$C\alpha$ atoms distance ≤ 8 Å.

2.2.7. Long-range contact order

LRCO is the calculated average sequence distance of Cα atoms [26], which is described as:

$$LRCO = \frac{1}{K.Nc} \sum_{\substack{s=1 \\ \\ |p-q|>0}}^{n} \Delta A_{pq} \tag{7}$$

where p and q are two interacting residues whose $C\alpha$-$C\alpha$ atom distance is ≤ 8 Å.

2.3. Network centrality measures

A residue-residue contact map was constructed using the same procedure as that used for RCO. Subsequently, we computed C_D, C_B, C_E, E_C and C_c.

Betweenness centrality (C_B) of node θ is the sum of the fraction of all pairs of shortest paths that pass through.

$$C_B(\theta) = \sum_{p,q \in R} \frac{\sigma(p,q|\theta)}{\sigma(p,q)} \tag{8}$$

where R is the set of nodes, $\sigma(p,q)$ is the number of shortest (p, q)-paths, and $\sigma(p,q|\theta)$ is the number of paths passing through node θ other than (p, q). If $p = q$, $\sigma(p,q) = 1$, and if $\theta \in p,q$, $\sigma(p,q|\theta) = 0$.

Edge betweenness centrality (C_E) is the sum of the fraction of all pairs of shortest paths that pass through the edge (e).

$$C_E = \sum_{p,q \in R} \frac{\sigma(p,q|e)}{\sigma(p,q)} \tag{9}$$

where R is the set of nodes, $\sigma(p,q)$ is the number of shortest (p, q) paths, and $\sigma(p,q|e)$ is the number of paths passing through edge e.

Eigen centrality (E_C) is calculated based on the neighbors' centrality. The eigenvector for node a is the ath element of vector y defined by the equation $Ty = \delta y$, where T is the adjacency matrix of graph G with the eigenvalue. According to the Perron-Frobenius theorem, there is a unique and positive solution y if δ is the largest eigenvalue of adjacency matrix T.

Closeness centrality (C_C) of a node is the reciprocal of the average shortest-path distance to σ over all n-1 reachable nodes

$$C_c(u) = \frac{k-1}{\sum_{u=1}^{k-1} d(a,b)} \tag{10}$$

where d (a, b) is the shortest distance between a and b, k-1 is the number of reachable nodes from u. A higher closeness value indicates higher centrality.

Degree centrality (C_D) is defined as a measure that counts the number of neighbors in the network.

2.4. Machine learning algorithms

We employed ten different shallow machine learning regressors: random forest (RF), extreme gradient boosting (XGB), decision tree (DT), adaboost (AB), artificial neural network (ANN), light gradient boosting (LGB), SVM, gradient boosting (GB), catboost (CB), and extremely randomized tree (ERT). Because the dataset was small, we did not employ deep learning-based algorithms [33]. In general, training datasets and external validation datasets can be used to develop prediction models and evaluate their transferability, respectively [34]. Although the number of proteins in the current study was greater than that reported in previous studies, it was not sufficient to divide them into two datasets. Previous studies have demonstrated that when a dataset is small, the entire dataset can be used for model training [35–37]. Based on these studies, we used all proteins as a training dataset for the development of prediction models. The search ranges for the hyperparameters for each algorithm were based on previous studies and optimized by leave-one-out cross-validation (LOOCV). These algorithms have been widely used for various function prediction problems based on their sequence, structure, images, and expression profiles.

2.5. Evaluation criteria

Three commonly used metrics were employed to evaluate performance: mean absolute difference (MAD), Pearson correlation coefficient (CC), and root mean square error (RMSE).

$$\begin{cases} CC = \dfrac{\sum_{i=1}^{K} (m_i - \overline{m})(n_i - \overline{n})}{\sqrt{\sum_{i=1}^{K} (m_i - \overline{m})^2} \sqrt{\sum_{i=1}^{K} (n_i - \overline{n})^2}} \\ MAD = \dfrac{1}{K} \sum_{i=1}^{K} |m_i - n_i| \\ RMSE = \sqrt{\dfrac{1}{K} \sum_{i=1}^{K} (m_i - n_i)^2} \end{cases} \tag{11}$$

where m_{i} and n_{i} are the predicted and observed folding rates of the ith protein, respectively; and K is the total number of proteins.

3. Results and discussion

3.1. Structural parameters and their relationship with ln(k_f)

We constructed a new dataset containing 89 TS and 52 NTS, which are relatively larger than the number of proteins used in previous studies. Using these datasets, we computed seven structural parameters (CTP, TCD, ACO, RCO, FLCO, LRO, and LRCO) and sequence length (N_{res}), and examined their relationship with ln(k_f). Notably, all structural parameters (except N_{res}) were derived based on the TS. Fig. 1 shows the linear relationship between ln(k_f) and different structural parameters for TS, where LRO achieved superior performance with a CC of −0.758. Interestingly, Gromiha et al. [38] proposed an LRO based on a smaller dataset and demonstrated a high correlation with ln(k_f). Surprisingly, LRO maintained similar levels of performance during our evaluation, particularly when we used a larger dataset. Conversely, the other six parameters (CTP, TCD, ACO, RCO, FLCO, and LRCO) showed significantly reduced correlations when compared with their corresponding values reported in the original paper, indicating that these parameters have limitations when applied to larger datasets.

For NTS proteins, we examined whether TS structural parameters along N_{res} have a linear relationship between NTS proteins and ln(k_f). The results show that N_{res} achieved the highest CC of −0.841, which is in line with the results of previous studies [39]. It is noteworthy that ACO achieved a CC of −0.772, which was significantly better than any other structural parameter. Apart from TCD and RCO, the other four structural parameters (CTP, FLCO, LRO, and LRCO) performed reasonably well on the NTS. As a result of this analysis, some of the structural parameters of TS proteins can also be applied to NTS proteins. In addition, we evaluated these structural parameters using a combined dataset (a combination of NTS and TS proteins). Figure S1shows that ACO achieved the

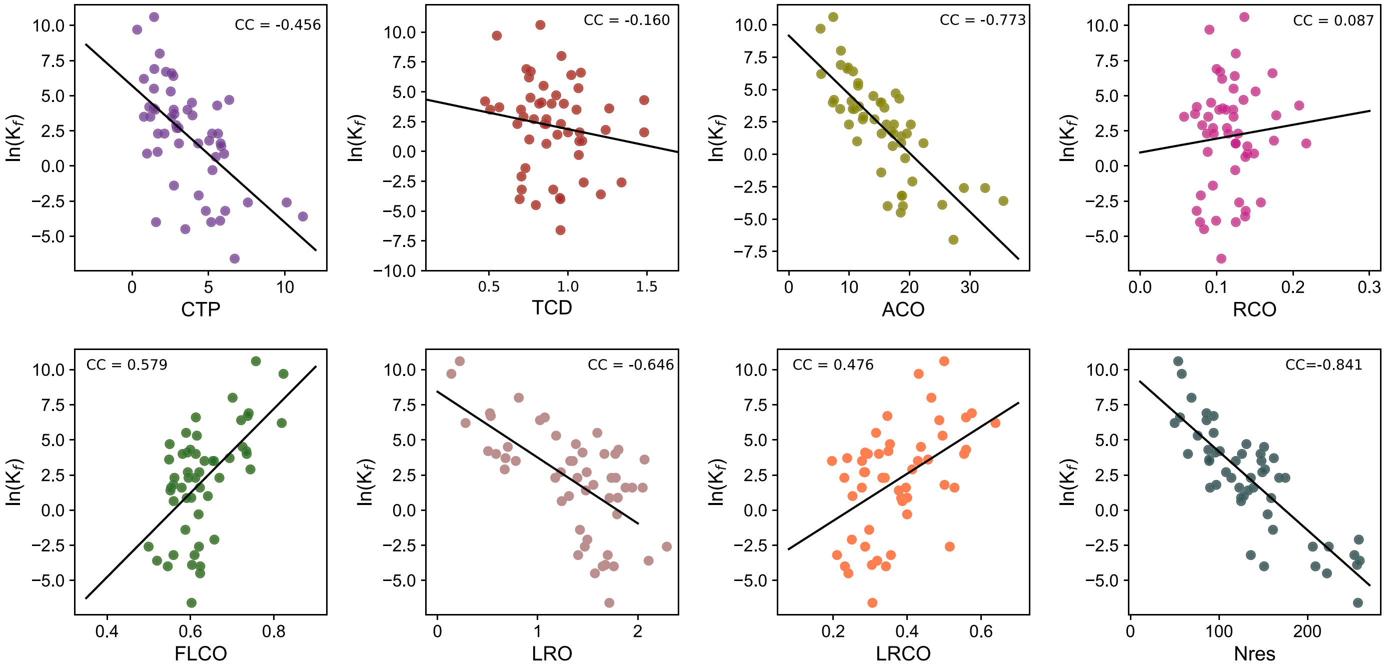

Fig. 1. Relationship between different structural parameters and folding rates of two-state proteins.

highest CC of −0.734, followed by LRO with a CC of −0.704, which was significantly better than the other structural parameters. The results of this study indicate that the structural parameters derived from TS proteins are also applicable when combined with NTS proteins. In general, LRO, N_{res}, and ACO were the most effective parameters for predicting ln (k_f) for the TS, NTS, and combined datasets, respectively. Because LRO and N_{res} have low CC in the combined dataset, they may not be suitable for identifying ln(k_f) in the absence of known protein-folding classes (TS/NTS).

3.2. *Network centrality measures and their relationship with ln(k_f) of TS and NTS*

We computed five network-based parameters (C_D, C_B, C_E, E_C and C_c.) and examine their relationship with ln(k_f). As shown in Figure S2, three parameters (C_D, C_C, and C_E) showed a reasonable performance (CC of ∼ −0.50) and were significantly better than the other two parameters for TS. However, the performance of the best network-based parameter (C_D) (CC = −0.508) was significantly lower than that of the best structural parameter (LRO) (CC of −0.758). In the case of NTS, (C_D, C_C, and C_E) achieved similar performance and were better than their respective performances on TS. Interestingly, C_C (CC of −0.842) achieves the same level of performance as N_{res} (CC of −0.841). When using the combined dataset, C_C (CC of −0.700) achieved a similar performance with the best structural parameter ACO (CC of −0.734). These results demonstrate that network centrality measures perform equally well with existing structural parameters on both the NTS and combined datasets, suggesting that residue communication is also important in the folding of proteins.

3.3. *Large scale machine learning regression models*

We employed ten different popular ML algorithms, including RF, XGB, DT, AB, ANN, LGB, SVM, GB, CB, and ERT. In this study, we evaluated all ten regressors because they have their own advantages and disadvantages with respect to predicting ln(k_f). To develop these ML regressors, three different types of information were considered as input features: a linear combination of structural parameters, a linear combination of network parameters, a linear combination of both structural and network parameters, and N_{res}. The performance was evaluated using three metrics (CC, MAD, and RMSE). Among these, MAD was recommended in previous studies [40,41] as being the most effective for real-value prediction, and we used the same for ranking the models.

The performance of different ML regression models based on TS proteins with three different feature sets is shown in Table 1. The performance of each ML model based on network properties was significantly lower than that of its counterpart based on structural parameters. Surprisingly, when these two types of information are incorporated into the respective ML model, the performance is similar in terms of CC; however, the MAD is slightly reduced, which indicates that multiple pieces of information are required to accurately predict ln(k_f) for TS. Furthermore, we observed that SVM is consistently better than other ML regressors regardless of the features, suggesting that it is better suited for predicting ln(k_f), possibly because of the small dataset size.

In the case of NTS, five regressors (ERT, AB, ANN, CB, and SVM) achieved similar performances in terms of the correlation coefficient between the network properties and structural parameters (Table 2). In terms of MAD, the corresponding regressor based on the structural parameters is superior. In the same manner as we observed in TS, the combination of both of these variables improves the correlation and reduces MAD. This indicates that both structural and network parameters are crucial for predicting ln(k_f) of NTS proteins. In general, the SVM regressor performs better on different sets of features than all the other regressors.

Based on the network properties in the combined dataset, each ML model performed significantly worse than its counterpart computed on the basis of structural parameters (Table 3). Interestingly, when these two types of information were incorporated into the respective ML models, both the CC and MAD improved. Furthermore, the SVM-based model performance is closer to the experimental values (Fig. 3) than the other regressors and ensemble models (described below). Table S1 lists the optimal hyperparameters of the final SVM model. As there are no publicly available structure-based prediction methods, we are unable to compare the current performance with existing predictors.

3.4. *Comparison of SVM-based single model with the ensemble models*

In general, ensemble models perform better than single models do

Table 1
Performance of different ML regressors on predicting ln(k_f) based on TS proteins.

	Network properties (NP)			Structural parameters (SP)			NP + SP		
ML regressors	CC	MAD	RMSE	CC	MAD	RMSE	CC	MAD	RMSE
RF	0.597	2.483	3.077	0.776	1.969	2.410	0.791	1.867	2.342
ERT	0.608	2.416	3.030	0.785	1.939	2.365	0.785	1.869	2.366
GB	0.573	2.546	3.129	0.769	1.994	2.440	0.787	1.885	2.355
AB	0.577	2.526	3.119	0.766	1.937	2.460	0.772	1.891	2.430
ANN	0.612	2.491	3.081	0.771	1.927	2.435	0.746	2.048	2.552
CB	0.610	2.409	3.033	0.794	1.902	2.331	0.788	1.884	2.359
DT	0.534	2.762	3.653	0.711	2.262	2.868	0.687	2.242	2.981
LGB	0.567	2.504	3.159	0.770	1.968	2.442	0.773	1.954	2.424
SVM	0.616	2.403	3.033	0.790	1.896	2.353	0.791	1.856	2.344
XGB	0.571	2.542	3.147	0.731	2.122	2.614	0.755	2.016	2.530

Table 2
Performance of different ML regressors on predicting ln(k_f) based on NTS proteins.

	Network properties (NP)			Structural parameters (SP)			NP + SP		
ML regressors	CC	MAD	RMSE	CC	MAD	RMSE	CC	MAD	RMSE
RF	0.780	1.948	2.378	0.817	1.757	2.178	0.832	1.656	2.111
ERT	0.800	1.876	2.268	0.827	1.726	2.132	0.843	1.649	2.052
GB	0.726	2.037	2.603	0.802	1.854	2.264	0.838	1.609	2.098
AB	0.820	1.749	2.173	0.836	1.682	2.088	0.845	1.642	2.066
ANN	0.820	1.803	2.164	0.847	1.654	2.014	0.841	1.674	2.051
CB	0.800	1.843	2.266	0.810	1.775	2.220	0.827	1.708	2.141
DT	0.733	2.128	2.692	0.775	2.019	2.486	0.764	2.019	2.525
LGB	0.638	2.502	2.913	0.740	2.054	2.542	0.741	1.990	2.562
SVM	0.837	1.677	2.094	0.862	1.603	1.915	0.866	1.550	1.898
XGB	0.784	1.907	2.365	0.813	1.762	2.239	0.822	1.726	2.220

Table 3
Performance of different ML regressors on predicting ln(k_f) based on (TS + NTS) proteins.

	Network properties (NP)			Structural parameters (SP)			NP + SP		
ML regressors	CC	MAD	RMSE	CC	MAD	RMSE	CC	MAD	RMSE
RF	0.679	2.413	3.022	0.784	1.957	2.541	0.806	1.829	2.426
ERT	0.709	2.269	2.888	0.796	1.935	2.491	0.811	1.822	2.403
GB	0.635	2.562	3.167	0.756	2.061	2.695	0.796	1.897	2.481
AB	0.708	2.291	2.893	0.788	1.912	2.521	0.802	1.845	2.447
ANN	0.699	2.350	2.941	0.798	1.862	2.473	0.803	1.854	2.441
CB	0.709	2.254	2.896	0.784	1.949	2.540	0.789	1.908	2.518
DT	0.584	2.992	3.710	0.696	2.403	3.168	0.741	2.283	2.908
LGB	0.659	2.508	3.082	0.750	2.111	2.709	0.793	1.992	2.496
SVM	0.716	2.242	2.865	0.811	1.833	2.398	0.815	1.745	2.379
XGB	0.680	2.392	3.014	0.775	1.961	2.598	0.777	1.991	2.581

[42]. In brief, ln(k_f) predicted by 10 different regressors was averaged and considered as the final output of the ensemble model. As shown in Fig. 4, ensemble models on two different datasets (TS and combined) achieved MADs of 1.89 and 1.81, respectively; these are better than those resulting from nine different regressors. However, the corresponding metrics (1.856 and 1.745, respectively) for the SVM-based single model were higher. In the case of NTS, the SVM-based single model is better than the ensemble model, indicating that the ensemble model is not suitable for predicting ln(k_f).

3.5. Model interpretation

Using the Shapley additive explanations (SHAP) [43], we estimated the contribution of each feature derived from our model. In SHAP, the contribution of each feature to the model output is allocated based on its marginal contribution to a group sample N (with n features) [44]. Fig. 5 shows the contribution of the top nine essential features and other features on the three different datasets. In the 2S dataset, the LRO and ACO features had high SHAP scores, indicating a significant contribution of these features. FLCO, LRCO, and C_B contributed moderately to 2S proteins. For the N2S dataset, ACO contributed the most significantly, followed by C_E, C_C, LRCO, and N_{res}. In the combined dataset, LRO contributed the most significantly, followed by ACO, C_C, C_B, and C_E. Both the structural parameters and network properties contributed to the top five, demonstrating the importance of combining different properties for accurate ln(k_f) prediction.

3.6. Comparison of SVM-based models with the statistical parameters

The parameters LRO, N_{res}, and ACO achieved CC values of −0.758, −0.841, and −0.734, respectively, for the TS, NTS, and combined datasets (Figs. 1–3). The corresponding CCs calculated based on the SVM were 0.791, 0.866, and 0.815, respectively (Tables 1–3). The performance of the ML-based models was significantly better than that of the best parameter, which is not surprising. Nevertheless, ML-based performance highlighting the importance of multiple factors extracted from the structure, is essential for predicting ln(k_f). In contrast to previous studies on protein folding properties, our results demonstrate that both long-range and short-range interactions, as well as size and residue communication, are necessary.

4. Conclusion

Several structural parameters based on 3D structures have been

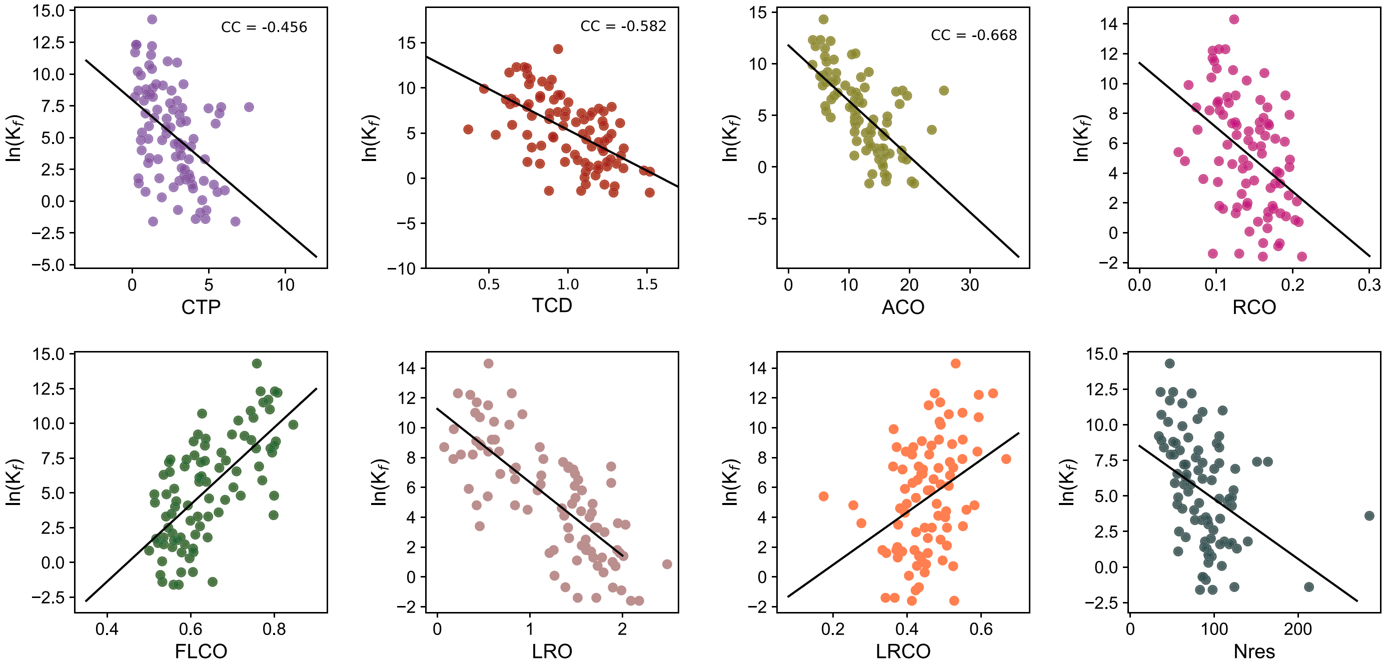

Fig. 2. Relationship between different structural parameters and folding rates of non-two state proteins.

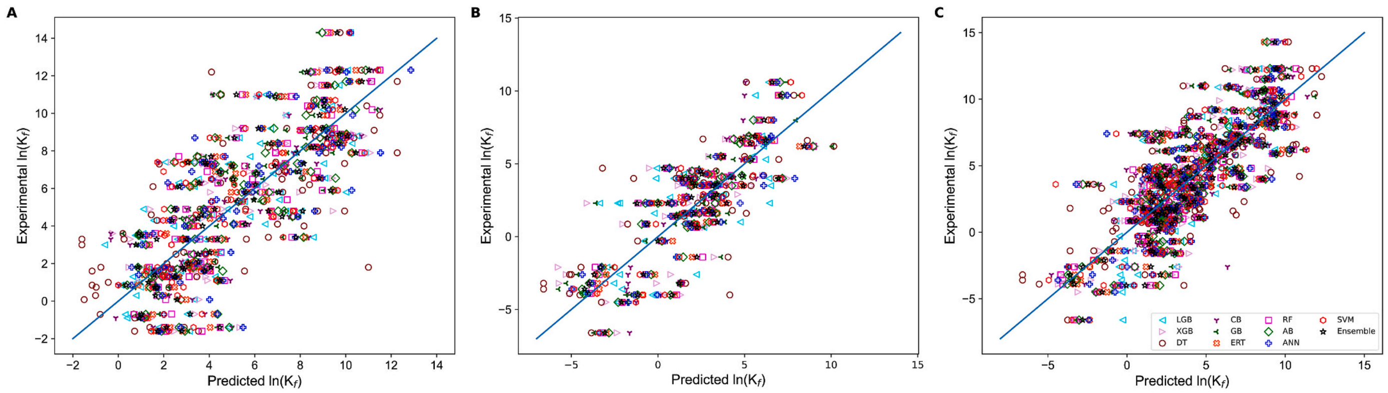

Fig. 3. Pairwise comparison between the experimental ln(k_f) and predicted ln(k_f) value of various methods on the training dataset. (A) TS (two-state proteins), (B) NTS (non-two-state proteins), and (C) combined (TS + NTS).

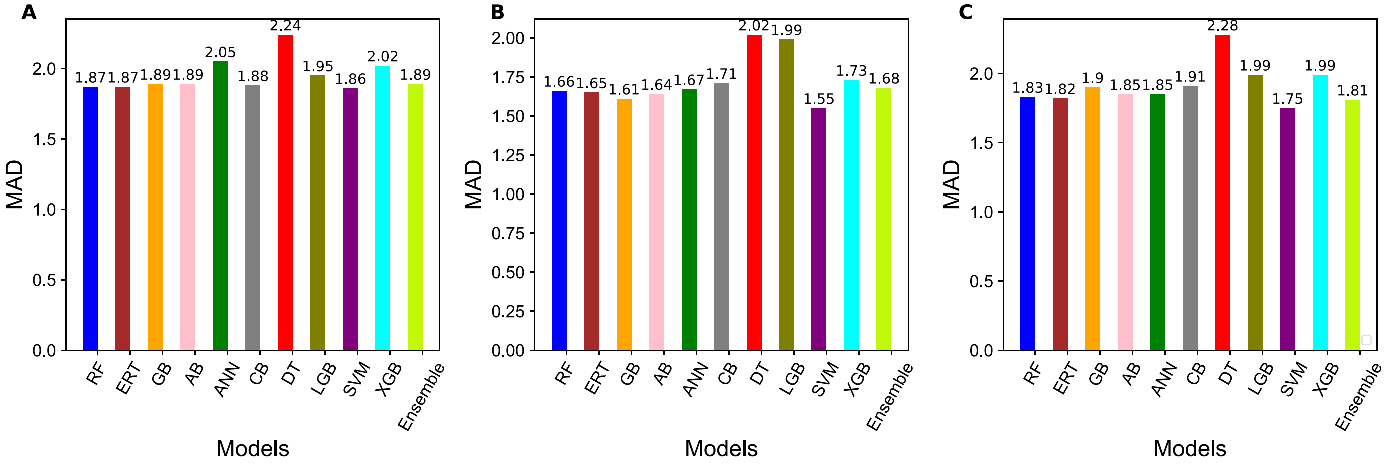

Fig. 4. Performance comparison in terms of MAD between different regressors on the training dataset. (A) TS, (B) NTS, and (C) combined.

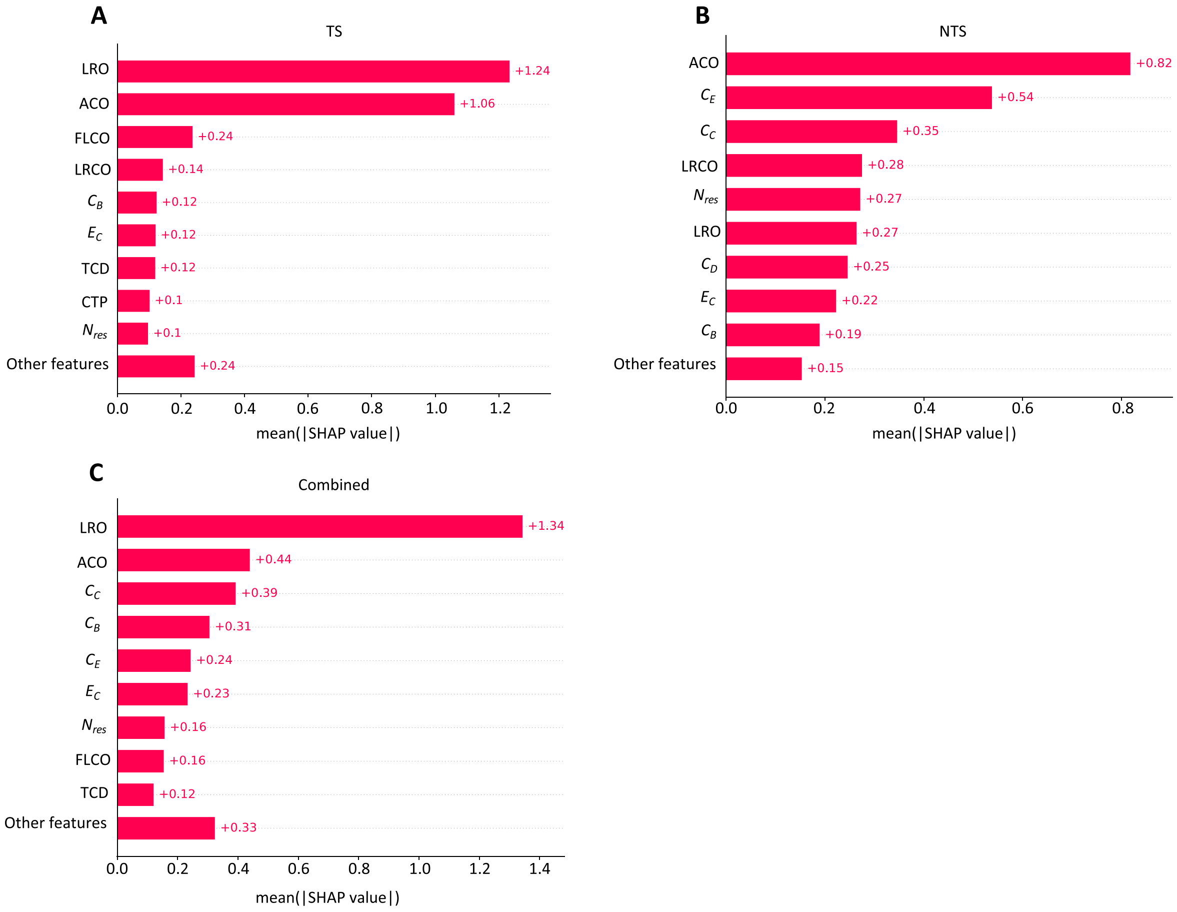

Fig. 5. SHAP interpretation of important features. The summary plot of the distribution of the most important features for (A) TS, (B) NTS, and (C) combined.

reported previously, and their relationships with ln(k_f) have been investigated [45–47]. On the basis of this strong relationship, they proposed a possible mechanism for protein folding. The purpose of this study is to determine the performance of each of these structural parameters using a larger dataset. Our results show that LRO [38] is capable of achieving a similar performance on TS proteins. Interestingly, ACO was constructed using TS proteins, whose parameters can be transferred to the NTS and combined datasets. Unfortunately, none of these parameters provides a reliable prediction of ln(k_f) based on the 3D structure. Thus, we evaluated a large-scale ML algorithm to identify a suitable algorithm for ln(k_f) prediction using seven structural parameters, chain length, and five network-based features. SVM achieved the best results regardless of the datasets, suggesting that local contacts, non-local contacts, chain lengths, and residue communications may cooperate and play a critical role in the folding process. In previous studies, TS and NTS proteins were treated separately to develop ML-based models [41,48]. It is noteworthy that the performance of the SVM models trained on combined datasets is similar to that of SVM models trained on TS and NTS datasets, indicating that folding class information is not necessary when predicting ln(k_f). AlphaFold [49] recently made a breakthrough in protein structure prediction by predicting structures closer to native structures based on sequence information. Moreover, over 200 million predicted structures have been deposited in the databases [50,51]. Our approach may be useful if researchers wish to find ln(k_f) prior to performing folding experiments. Although we conducted large-scale ML-based prediction models, all were single models. Recent studies have highlighted the importance of integrating multiple models to solve various problems [52–57]. In this regard, we plan to test different computational approaches and make the best model publicly available on a web server.

Author contribution

G.L. and B.M. conceived and designed the study; data analysis, experimentation, and manuscript draft preparation were performed by S.N. and V. K. S.; proofreading and manuscript finalization was done by B.M. and G.L. Resources and funding acquisition by G.L. All authors read and approved the final manuscript.

Funding

This work was supported by the National Research Foundation (NRF), funded by the Ministry of Science and ICT (MSIT), Korea [2020R1A4A4079722 and 2020M3E5D9080661].

Declaration of competing interest

The author declares no conflict of interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.106436.

References

[1] D. Lancet, Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.

[2] H. Chial, DNA sequencing technologies key to the Human Genome Project, Nat. Educ. (2008) 1.

[3] H. Gelman, M. Gruebele, Fast protein folding kinetics, Q. Rev. Biophys. 47 (2014) 95–142.

[4] K. Lindorff-Larsen, S. Piana, R.O. Dror, D.E. Shaw, How fast-folding proteins fold, Science 334 (2011) 517–520.

[5] S.E. Jackson, How do small single-domain proteins fold? Folding Des. 3 (1998) R81–R91.

[6] K. Kamagata, M. Arai, K. Kuwajima, Unification of the folding mechanisms of non-two-state and two-state proteins, J. Mol. Biol. 339 (2004) 951–965.

[7] O. Ptitsyn, Molten globule and protein folding, Adv. Protein Chem. 47 (1995) 83–229.

[9] C. Soto, L.D. Estrada, Protein misfolding and neurodegeneration, Arch. Neurol. 65 (2008) 184–189.

[11] M.M. Gromiha, A.M. Thangakani, S. Selvaraj, FOLD-RATE: prediction of protein folding rates from amino acid sequence, Nucleic Acids Res. 34 (2006) W70–W74.

[12] P.C. Whitford, J.N. Onuchic, What protein folding teaches us about biological function and molecular machines,, Curr. Opin. Struct. Biol. 30 (2015) 57–62.

[13] K.W. Plaxco, K.T. Simons, D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins, J. Mol. Biol. 277 (1998) 985–994.

[14] M.M. Gromiha, S.J. Selvaraj, Comparison between Long-Range Interactions and Contact Order in Determining the Folding Rate of Two-State Proteins: Application of Long-Range Order to Folding Rate Prediction, J. Mol. Biol. 310 (2001) 27–32.

[15] H. Zhou, Y. Zhou, Folding rate prediction using total contact distance, Biophys. J. 82 (2002) 458–463.

[16] L. Mirny, E. Shakhnovich, Protein folding theory: from lattice to all-atom models, Annu. Rev. Biophys. Biomol Struct. 30 (2001) 361–396.

[17] M.M.J. Gromiha, Importance of Native-State Topology for Determining the Folding Rate of Two-State Proteins, J. Chem. Inf. Comput. Sci. 43 (2003) 1481–1485.

[18] K.W. Plaxco, K.T. Simons, D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins, J. Mol. Biol. 277 (1998) 985–994.

[19] D.N. Ivankov, S.O. Garbuzynskiy, E. Alm, K.W. Plaxco, D. Baker, A.V. Finkelstein, Contact Order Revisited: Influence of Protein Size on the Folding Rate, Protein Sci. vol. 12 (2003) 2057–2062.

[20] O.V. Galzitskaya, S.O. Garbuzynskiy, D.N. Ivankov, A.V. Finkelstein, Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics, Proteins 51 (2003) 162–166.

[21] C. Micheletti, Prediction of folding rates and transition-state placement from native-state geometry, Proteins 51 (2003) 74–84.

[22] D.N. Ivankov, A.V. Finkelstein, Prediction of Protein Folding Rates from the Amino Acid Sequence-Predicted Secondary Structure, Proc. Natl. Acad. Sci. USA vol. 101 (2004) 8942–8944.

[23] Z. Ouyang, J. Liang, Predicting protein folding rates from geometric contact and amino acid sequence, Protein Sci. 17 (2008) 1256–1263.

[24] N.V. Dokholyan, L. Li, F. Ding, E.I. Shakhnovich, Topological determinants of protein folding, Proc. Natl. Acad. Sci. USA 99 (2002) 8637–8641.

[25] B. Nölting, W. Schälike, P. Hampel, F. Grundig, S. Gantert, N. Sips, W. Bandlow, P. X. Qi, Structural determinants of the rate of protein folding, Cell. Mol. Life Sci. 223 (2003) 299–307.

[26] B.-G. Ma, L.-L. Chen, H.-Y. Zhang, What determines protein folding type? An investigation of intrinsic structural properties and its implications for understanding folding mechanisms, J. Mol. Biol. 370 (2007) 439–448.

[27] K. Kamagata, K. Kuwajima, Surprisingly high correlation between early and late stages in non-two-state protein folding, J. Mol. Biol. 357 (2006) 1647–1654.

[28] E. Capriotti, R. Casadio, K-Fold: a tool for the prediction of the protein folding kinetic order and rate, Bioinformatics 23 (2007) 385–386.

[29] B. Manavalan, K. Kuwajima, J. Lee, PFDB: a Standardized Protein Folding Database with Temperature Correction, Sci. Rep. vol. 9 (2019) 1–9.

[30] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I. N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.

[31] Y. Shi, J. Zhou, D. Arndt, D.S. Wishart, G. Lin, Protein contact order prediction from primary sequences, BMC Bioinf. 9 (2008) 1–9.

[32] I.B. Kuznetsov, S. Rackovsky, Class-specific correlations between protein folding rate, structure-derived, and sequence-derived descriptors, Proteins 54 (2004) 333–341.

[33] M.M. Hasan, S. Tsukiyama, J.Y. Cho, H. Kurata, M.A. Alam, X. Liu, B. Manavalan, H.W. Deng, Deepm5C: a deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy, Mol. Ther. 30 (2022) 2856–2867.

[34] S. Basith, B. Manavalan, T. Hwan Shin, G. Lee, Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening, Med. Res. Rev. 40 (2020) 1276–1314.

[35] B. Liu, L. Fang, R. Long, X. Lan, K.C. Chou, iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition, Bioinformatics 32 (2016) 362–369.

[36] W. Su, F. Wang, J.-X. Tan, F.-Y. Dao, H. Yang, H. Ding, The prediction of human DNase I hypersensitive sites based on DNA sequence information, Chemometr. Intell. Lab. Syst. 209 (2021), 104223.

[37] X. Qiang, C. Zhou, X. Ye, P.F. Du, R. Su, L. Wei, CPPred-FL: a Sequence-Based Predictor for Large-Scale Identification of Cell-Penetrating Peptides by Feature Representation Learning, Brief Bioinform, 2018.

[38] M.M. Gromiha, S. Selvaraj, Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction, J. Mol. Biol. 310 (2001) 27–32.

[39] O.V. Galzitskaya, S.O. Garbuzynskiy, D.N. Ivankov, A.V. Finkelstein, Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics, Proteins: Struct., Funct., Bioinf. 51 (2003) 162–166.

[40] C.C. Chang, B.T. Tey, J. Song, R.N. Ramanan, Towards more accurate prediction of protein folding rates: a review of the existing Web-based bioinformatics approaches, Briefings Bioinf. 16 (2015) 314–324.

[41] B. Manavalan, J. Lee, FRTpred: a novel approach for accurate prediction of protein folding rate and type, Comput. Biol. Med. 149 (2022), 105911.

[42] H. Lv, Y. Zhang, J.-S. Wang, S.-S. Yuan, Z.-J. Sun, F.-Y. Dao, Z.-X. Guan, H. Lin, K.-J. Deng, iRice-MS: an integrated XGBoost model for detecting multitype post-translational modification sites in rice, Briefings Bioinf. 23 (2022) bbab486.

[43] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. (2017) 30.

[44] F. Wang, L. Wei, Multi-scale deep learning for the imbalanced multi-label protein subcellular localization prediction based on immunohistochemistry images, Bioinformatics 38 (2022) 2602–2611.

[45] K.W. Plaxco, K.T. Simons, D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins, J. Mol. Biol. 277 (1998) 985–994.

[46] M.M. Gromiha, Importance of native-state topology for determining the folding rate of two-state proteins, J. Chem. Inf. Comput. Sci. 43 (2003) 1481–1485.

[47] D.N. Ivankov, S.O. Garbuzynskiy, E. Alm, K.W. Plaxco, D. Baker, A.V. Finkelstein, Contact order revisited: influence of protein size on the folding rate, Protein Sci. 12 (2003) 2057–2062.

[48] J. Song, K. Takemoto, H. Shen, H. Tan, M.M. Gromiha, T. Akutsu, Prediction of protein folding rates from structural topology and complex network properties, IPSJ Trans. Bioinf. 3 (2010) 40–53.

[49] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S.A. A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature 596 (2021) 583–589.

[50] M. Varadi, S. Velankar, The impact of AlphaFold Protein Structure Database on the fields of life sciences, Proteomics (2022), e2200128.

[51] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Zidek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, Nucleic Acids Res. 50 (2022) D439–D444.

[52] S. Basith, G. Lee, B. Manavalan, STALLION: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction, Briefings Bioinf. 23 (2022).

[53] P. Charoenkwan, N. Schaduangrat, P. Lio, M.A. Moni, B. Manavalan, W. Shoombuatong, NEPTUNE: a novel computational approach for accurate and large-scale identification of tumor homing peptides, Comput. Biol. Med. (2022), 105700.

[54] M.M. Hasan, S. Tsukiyama, J.Y. Cho, H. Kurata, M.A. Alam, X. Liu, B. Manavalan, H.W. Deng, Deepm5C: a deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy, Mol. Ther. 30 (2022) 2856–2867.

[55] Y.J. Jeon, M.M. Hasan, H.W. Park, K.W. Lee, B. Manavalan, TACOS: a novel approach for accurate prediction of cell-specific long noncoding RNAs subcellular localization, Briefings Bioinf. 23 (2022).

[56] W. Shoombuatong, S. Basith, T. Pitti, G. Lee, B. Manavalan, THRONE: a new approach for accurate prediction of human rna N7-methylguanosine sites, J. Mol. Biol. 434 (2022), 167549.

[57] F.Y. Dao, H. Lv, M.J. Fullwood, H. Lin, Accurate identification of DNA replication origin by fusing epigenomics and chromatin interaction information, Research 2022 (2022), 9780293.