Biochimica et Biophysica Acta 1864 (2016) 1104–1109

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

Prediction of change in protein unfolding rates upon point mutations in two state proteins

Priyashree Chaudhary, Athi N. Naganathan, M. Michael Gromiha *

Department of Biotechnology, Bhupat & Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600 036, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 26 March 2016
Received in revised form 5 May 2016
Accepted 1 June 2016
Available online 2 June 2016

Keywords:
Two-state proteins
Unfolding rates
Point mutations
Multiple regression technique
Machine learning
Amino acid properties

ABSTRACT

Studies on protein unfolding rates are limited and challenging due to the complexity of unfolding mechanism and the larger dynamic range of the experimental data. Though attempts have been made to predict unfolding rates using protein sequence-structure information there is no available method for predicting the unfolding rates of proteins upon specific point mutations. In this work, we have systematically analyzed a set of 790 single mutants and developed a robust method for predicting protein unfolding rates upon mutations ($\Delta \ln k_u$) in two-state proteins by combining amino acid properties and knowledge-based classification of mutants with multiple linear regression technique. We obtain a mean absolute error (MAE) of 0.79/s and a Pearson correlation coefficient (PCC) of 0.71 between predicted unfolding rates and experimental observations using jack-knife test. We have developed a web server for predicting protein unfolding rates upon mutation and it is freely available at https://www.iitm.ac.in/bioinfo/proteinunfolding/unfoldingrace.html. Prominent features that determine unfolding kinetics as well as plausible reasons for the observed outliers are also discussed.

1. Introduction

One of the challenges in the field of protein folding is to identify basic protein features that determine the rates of folding as well as unfolding, and hence the conformational stability. Several experimental studies have been carried out to measure the folding rates of proteins [1]. Consequently, several structural parameters such as contact order (CO), long-range order (LRO), total contact distance, cliquishness, and multiple contact index have been proposed to understand and predict protein folding rates from three-dimensional structures of proteins [2–7]. Further, sequence based methods have also been developed for predicting the folding rates of two and three-state proteins [8–10].

On the other hand, the available experimental data on unfolding rate constants and the mechanisms underlying protein unfolding rates show different behaviors to those of folding rates [11]. Importantly, while the folding rates of proteins vary by just over 5 orders of magnitude [12] the unfolding rates exhibit a much larger dynamic range and span ~8–10 orders of magnitude for single domain proteins. The available experimental data on protein unfolding rates have been related with various structure based parameters. For example, Jung et al. [13] associated them with topological parameters such as contact order, clustering coefficient and average path length. They have also utilized a network of contacts in native protein structures using graph theory for relating protein unfolding rates [14]. Protein unfolding rates ($\ln k_u$) have also been studied with LRO [15], free energy surface model combining the inter-atomic contacts with protein stability [16] and a physical one-dimensional free energy surface model [17]. In addition to the structure-based methods, sequence based models have been proposed by using physicochemical, conformational and energetic properties of amino acid residues for predicting protein unfolding rates [18].

Point mutations in a protein alter its structure, folding, stability and function. The factors influencing the stability of proteins upon mutations have been well documented and several methods have been proposed for predicting the stability of proteins upon amino acid substitutions using protein three-dimensional structural information and/or just from amino acid sequence [8]. The effects of point mutations on various diseases, specifically on cancer have been studied in detail [19,20]. In our earlier works, we have developed a method for predicting the folding rates of two-state proteins upon point mutations, which showed a good performance when compared against the experimental data [10]. However, there is no method available for predicting protein unfolding rates upon mutations.

In this work, we have developed a novel method for the prediction of changes in unfolding rates upon point mutation using amino acid properties, secondary structural information and the precise location of the mutation. Our method showed a correlation and MAE of 0.71 and 0.79/s, respectively with experimental data on leave-one-out cross-validation. We have developed a web server for the same and it is freely available at https://www.iitm.ac.in/bioinfo/proteinunfolding/unfoldingrace.html.

* Corresponding author.
E-mail address: gromiha@iitm.ac.in (M.M. Gromiha).

http://dx.doi.org/10.1016/j.bbapap.2016.06.001
1570-9639/

2. Materials and methods

2.1. Dataset

We have constructed a set of 790 mutants from 26 two-state-like proteins, which contain 14 to 68 mutants in each protein using the data available in the literature [21]. Due to the unavailability of extensive mutational data for three-state and large proteins we considered only two-state small proteins in this work. The unfolding rates of proteins span more than eight orders of magnitude and it is difficult to treat the difference of unfolding rates upon point mutations. Moreover, the difference in logarithm of unfolding rates for the wild type and mutation is proportional to the difference in the unfolding activation free energy that is widely used in the literature. Hence, we used $\Delta\ln k_u$ as change in protein unfolding rates upon mutation. The $\Delta\ln k_u$ values are in the range between −3.46/s and 8.05/s. Further, we used a test set of 20 non-redundant mutants, which have the $\Delta\ln k_u$ values between −0.27/s and 4.20/s. Interestingly, the $\Delta\ln k_u$ values have a wider range and twice the standard deviation compared to the $\Delta\ln k_f$ for the same point mutations [10]. The complete dataset of 790 mutants along with mutation details and unfolding rates are available at https://www.iitm.ac.in/bioinfo/proteinunfolding/unfoldingrace.html.

2.2. Amino acid properties

We start with a comprehensive set of 242 diverse amino acid features obtained from Amino Acid Index Database [22] and literature [23] that includes physicochemical, conformational, thermodynamic and evolutionary properties, which are relevant for the present study. We then utilized an ensemble of attribute selection methods available in WEKA [24] for reducing the features as explained earlier [10] and **Supplementary Table S1** shows the reduced list of features.

2.3. Computational procedures

We have followed the procedure described in Chaudhary et al. [10] for developing the method for predicting the changes in unfolding rates of two-state proteins upon point mutations.

The change in unfolding rate upon point mutations $\Delta\ln k_u$, is calculated as:

$$\Delta\ln k_u = \ln k_u^{mut} - \ln k_u^{WT} \quad (1)$$

where $\ln k_u^{mut}$ and $\ln k_u^{WT}$ are natural logarithms of folding rates for mutant and wild-type amino acid residues, respectively.

We observe that none of the shortlisted features exhibited a PCC of more than ±0.47 with $\Delta\ln k_u$ for the entire set of 790 mutants taken together. Therefore, we classified the mutants based on secondary structure, SS (helix, strand, coil), normalized accessible surface area, ASA (buried, ASA ≤ 12%, partially buried, 12 < ASA ≤ 36% and exposed, ASA > 36%) and sequence position (*N*-terminal, ≤33%, Middle, 33–67% and *C*-terminal, ≥67%) of the wild-type residues so that each class contains uniform distribution of data and minimum redundancy. The single state classification has nine classes (3 classes each), double and triple state classifications have 27 classes. The grouping of 27 classes using the combinations of SS, ASA and sequence position are presented in Table 1. ASA and secondary structure of mutants were assigned using DSSP [25].

For each class, we utilized multiple linear regression technique to identify the best combination of three features to relate protein unfolding rates, $\Delta\ln k_u$ [26]. We have set up a total of 27 models for each class and the regression equations are shown in Supplementary Table S2. These models are subjected to leave-one-out cross-validation (jack-knife test) with n iterations, where n is the total number of data (trained with n-1 data and tested the omitted one) to evaluate the performance of the method. The same model is also used on a blind test set for validating the method.

With this approach, however, we were unable to predict the effect of the same mutation type (say, V → A) at different positions, suggesting the importance of neighboring residues close to the mutated site. Hence, we included the information about neighboring residues, which varies with location for each mutation depending on the nearby residues. The contribution of neighboring residues (ΔP_{seq}) has been obtained using average property value for window lengths of 3 to 19 residues. It is computed using the equation [10]:

$$\Delta P_{seq} = P_{mut}(i) - \left[\left(\sum_{j=i-k}^{j=i+k} P_j(i)/(2k+1)\right)\right] \quad (2)$$

where, k varies from 0 to 9 residues on both directions; zero represents only the mutations and without neighboring residue information, 1 uses a window of 3 residues and so on.

2.4. Evaluation and validation of the method

Two measures viz. PCC and MAE (mean of absolute difference between experimental and predicted values of the logarithmic change in folding rates) were employed to evaluate the performance of the present method for each class. We have also examined the quality of prediction using *p*-values.

The performance of the method was validated with three methods:

(i) Jack-knife/*Leave-One-Out cross validation*: Each mutant from the dataset is left out and the prediction is performed by training *n*-1 dataset for the omitted mutant. Likewise, the procedure is iterated for the entire dataset to obtain the mean measure.

(ii) *n-fold cross validation*: 'n' percentage of entire data is eliminated from the training set and is used as a validation set for testing the model, constituted by the rest of the database. 5, 10, 20, 30 and

Table 1
Correlation coefficient and mean absolute error in 27 classes of mutants based on secondary structure, solvent accessibility and sequence position.

S. No	SS	ASA	Position	No. of mutants	Window	Performance		*p*-value
						r	MAE (s^{-1})	
1	Strand	0–12%	0–33%	45	3	0.64	0.96	6.64E−05
2	Strand	0–12%	33–67%	30	3	0.65	0.93	2.39E−03
3	Strand	0–12%	>67%	52	5	0.6	1.12	9.47E−05
4	Strand	12–36%	0–33%	30	9	0.75	0.63	7.09E−05
5	Strand	12–36%	33–67%	21	None	0.86	0.66	2.59E−05
6	Strand	12–36%	>67%	24	None	0.75	0.48	6.82E−04
7	Strand	>36%	0–33%	16	15	0.94	0.1	4.41E−06
8	Strand	>36%	33–67%	14	None	0.98	0.14	1.29E−07
9	Strand	>36%	>67%	13	None	0.99	0.12	3.37E−07
10	Helix	0–12%	0–33%	36	3	0.73	0.98	1.42E−05
11	Helix	0–12%	33–67%	42	9	0.73	1.02	2.40E−06
12	Helix	0–12%	>67%	29	None	0.72	1.1	3.92E−04
13	Helix	12–36%	0–33%	22	9	0.8	0.94	2.57E−04
14	Helix	12–36%	33–67%	43	3	0.74	1.03	8.37E−07
15	Helix	12–36%	>67%	23	None	0.85	0.78	1.37E−05
16	Helix	>36%	0–33%	37	9	0.73	0.56	1.16E−05
17	Helix	>36%	33–67%	40	15	0.69	0.85	2.79E−05
18	Helix	>36%	>67%	32	5	0.77	0.73	1.40E−05
19	Others	0–12%	0–33%	19	5	0.85	0.57	1.75E−04
20	Others	0–12%	33–67%	37	19	0.72	0.85	1.84E−05
21	Others	0–12%	>67%	13	7	0.99	0.26	1.45E−07
22	Others	12–36%	0–33%	24	None	0.86	0.57	6.18E−06
23	Others	12–36%	33–67%	22	5	0.84	0.55	5.79E−05
24	Others	12–36%	>67%	20	7	0.87	0.66	3.84E−05
25	Others	>36%	0–33%	42	7	0.64	0.58	1.56E−04
26	Others	>36%	33–67%	42	7	0.67	0.53	4.10E−05
27	Others	>36%	>67%	22	15	0.89	0.54	1.81E−06
Mean						0.79	0.67	

SS: secondary structure, ASA: normalized accessible surface area.