Bioinformatics, 31(13), 2015, 2091–2097
doi: 10.1093/bioinformatics/btv091
Advance Access Publication Date: 16 February 2015
Original Paper

Structural bioinformatics

Folding RaCe: a robust method for predicting changes in protein folding rates upon point mutations

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

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

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on August 11, 2014; revised on January 30, 2015; accepted on February 10, 2015

Abstract

Motivation: Protein engineering methods are commonly employed to decipher the folding mechanism of proteins and enzymes. However, such experiments are exceedingly time and resource intensive. It would therefore be advantageous to develop a simple computational tool to predict changes in folding rates upon mutations. Such a method should be able to rapidly provide the sequence position and chemical nature to modulate through mutation, to effect a particular change in rate. This can be of importance in protein folding, function or mechanistic studies.

Results: We have developed a robust knowledge-based methodology to predict the changes in folding rates upon mutations formulated from amino and acid properties using multiple linear regression approach. We benchmarked this method against an experimental database of 790 point mutations from 26 two-state proteins. Mutants were first classified according to secondary structure, accessible surface area and position along the primary sequence. Three prime amino acid features eliciting the best relationship with folding rates change were then shortlisted for each class along with an optimized window length. We obtained a self-consistent mean absolute error of 0.36 s^{-1} and a mean Pearson correlation coefficient (PCC) of 0.81. Jack-knife test resulted in a MAE of 0.42 s^{-1} and a PCC of 0.73. Moreover, our method highlights the importance of outlier(s) detection and studying their implications in the folding mechanism.

Availability and implementation: A web server ‘Folding RaCe’ has been developed and is available at http://www.iitm.ac.in/bioinfo/proteinfolding/foldingrace.html.

Contact: gromiha@iitm.ac.in

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Understanding protein folding mechanisms involves identifying the series of steps taken by a polypeptide chain at a high structural and temporal resolution during the folding process. Though the temporal aspects are challenging to decipher, the sequential and structural features have been relatively easier to address through protein engineering methodologies (Fersht *et al.*, 1992). This involves the measurement of both the folding rates and stability for a series of carefully designed point mutations along the entire protein sequence. The effects of point mutations on rates and stability are then compared to the wild type to obtain a dimensionless number, the phi-value (Φ) that is generally considered as a proxy for the degree of structure at the transition state and hence shedding light on the mechanism of folding.

Though site-directed mutagenesis experiments have been successfully applied to more than 40 different proteins, the downside is that measuring the stability and folding rates for the series of mutants is exceedingly time and resource-intensive. An alternative to

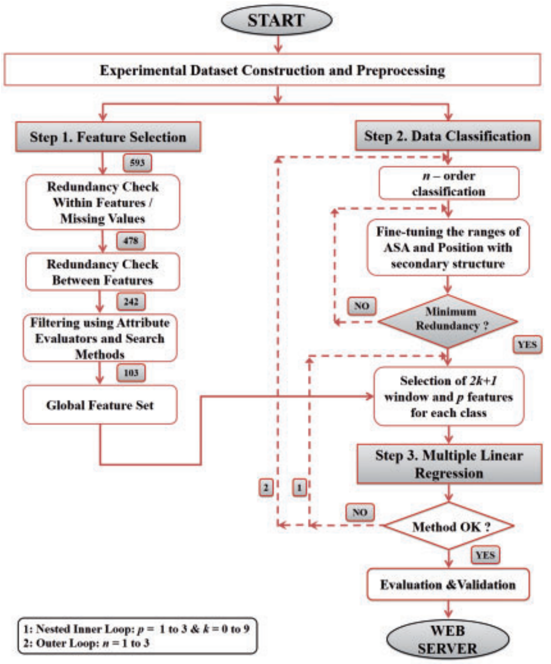
START
Experimental Dataset Construction and Preprocessing
Step 1. Feature Selection
593
Redundancy Check Within Features / Missing Values
478
Redundancy Check Between Features
242
Filtering using Attribute Evaluators and Search Methods
103
Global Feature Set
Step 2. Data Classification
n – order classification
Fine-tuning the ranges of ASA and Position with secondary structure
NO
Minimum Redundancy ?
YES
Selection of 2k+1 window and p features for each class
Step 3. Multiple Linear Regression
2
1
NO
Method OK ?
YES
Evaluation &Validation
WEB SERVER
1: Nested Inner Loop: p = 1 to 3 & k = 0 to 9
2: Outer Loop: n = 1 to 3

Fig. 1. Steps involved during model development

this approach would involve the computational prediction of changes in rates and stability upon point mutations. In this regard, predicting the changes in thermodynamic stability upon point mutations are already well established through sequence-based features (Huang *et al.*, 2007), structure-based parameters (Dehouck *et al.*, 2011; Parthiban *et al.*, 2006), force-field based approaches (Schymkowitz *et al.*, 2005; Yin *et al.*, 2007) and statistical-mechanical methods (Naganathan, 2013).

Remarkably, despite the structural complexity of proteins, the absolute folding rates are nowadays easy to predict either from topological/structural considerations (Gromiha, 2009; Gromiha and Selvaraj, 2001; Makarov *et al.*, 2002; Micheletti, 2003; Plaxco *et al.*, 1998; Zhou and Zhou, 2002), protein length (De Sancho *et al.*, 2009; Naganathan and Muñoz, 2004; Thirumalai, 1995), protein length combined with the secondary structure content (Ivankov and Finkelstein, 2004), statistical mechanical approaches (Henry and Eaton, 2004; Muñoz and Eaton, 1999) and simply even from the primary sequence information combining statistical and machine-learning approaches (Capriotti and Casadio, 2007; Cheng *et al.*, 2013; Gromiha *et al.*, 2006; Huang and Gromiha, 2008; Huang and Gromiha, 2010; Lin *et al.*, 2010; Ouyang and Liang, 2008; Punta and Rost, 2005).

However, the development of models for predicting the effects of point mutations on changes in folding rates is still at an infant stage. This is important as the availability of a computational tool could significantly reduce the time taken in experimental approaches. More relevantly, it could serve as a first exploratory step to design mutations of interest that could then be experimentally tested. Moreover, it might aid in synthetic designing of protein mutants with known effect on folding rate, for even therapeutic purposes. Till date, Prediction of protein FOlding RAte change upon point mutation (FORA) is the only available web server for real value prediction of folding rates upon point mutations based on a quadratic regression model (Huang and Gromiha, 2012). The major pitfalls in this approach include the small training dataset of 467 mutants, moderate performance, encompassing both two-state and multistate proteins, less diversity in nature of mutants as well as amino acid feature dataset.

In this work, we have developed a rigorous, knowledge-based model for the prediction of folding rates employing a large database of 790 single-site point mutants specifically from 26 two-state proteins. A triple-order classification scheme together with a multiple linear regression model incorporating three descriptors was able to predict the changes in folding rates with a mean Pearson correlation coefficient (PCC) of 0.73, mean absolute error (MAE) of $0.42\,s^{-1}$ and prediction accuracy of 81.20% after jack-knife validation. The structural implications of the outlying mutants have been exclusively discussed. Web server named '*Folding RaCe*' has been developed for prediction purpose.

2 Materials and Methods

2.1 Dataset

We have constructed a set of 790 mutants from two-state proteins using the data available in Single-Point Mutants Protein Folding Database (Naganathan and Muñoz, 2010) (Supplementary Table S1a). They encompass 26 proteins from all structural classes and sequence lengths varying from 37 to 107 amino acid residues. The distribution of mutants based on experimental $\Delta\ln k_f$ is shown in Supplementary Figure S1a. The $\Delta\ln k_f$ values are in the range of $-5.23\,s^{-1}$ to $2.60\,s^{-1}$. Among the 790 mutants, 180 are accelerating with an average $\Delta\ln k_f$ of $0.40\,s^{-1}$ and the rest are decelerating with an average $\Delta\ln k_f$ of $-0.90\,s^{-1}$.

Further, we have used a non-redundant test set of 59 mutants from FORA (Huang and Gromiha, 2012) (Supplementary Table S1b). The $\Delta\ln k_f$ values in the test set varies from -3.73 to $1.13\,s^{-1}$ and the average values of 9 accelerating and 50 decelerating mutants are 0.42 and $-0.92\,s^{-1}$, respectively (Supplementary Fig. S1b).

2.2 Amino acid properties

We have utilized a comprehensive collection of 593 diverse amino acid features that includes physicochemical, conformational, thermodynamic and evolutionary properties in this study. It contains 544 descriptors from Amino Acid Index Database (Kawashima and Kanehisa, 2000) and 49 properties from the literature (Gromiha, 2005). These properties were normalized between 0 and 1 to employ a common scale. Further, we have eliminated the redundant and nonrelevant descriptors, and reduced them to 103 using the following procedure (Step 1 in Fig. 1): (i) removed 14 properties in which the values are missing for any of the amino acid residues, (ii) removed 101 properties in which more than two amino acid residues have the same value, which have no effect on mutation, (iii) for avoiding redundancy, we have eliminated 236 properties in such a way that no two properties have the absolute correlation coefficient of more than 0.85 and (iv) utilized an ensemble of attribute selection methods (six attribute evaluators along with four search methods) available in Waikato Environment for Knowledge Analysis (WEKA) (Hall *et al.*, 2009) for reducing the features (Supplementary Table S2). For each method, we detected the important features using all the 790 mutants as well as the ones selected in more than 50% of the cross-validated datasets. The selection criteria for each method have been explained in Witten and Frank (2005). Then, we chose the features, which are identified to be important in at least two of the methods. This procedure along with the inclusion of relevant properties for protein folding such as unfolding entropy change of hydration, surrounding hydrophobicity etc. by manual inspection reduced the number of features to 103 (Supplementary Table S3).

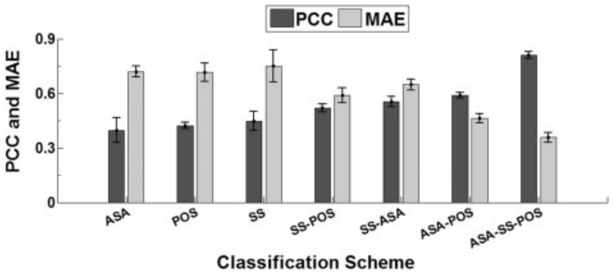

Fig. 2. Influence of increase in the order of classification on model performance along with error bars. SS, Secondary Structure (Strand, Helix, Others); ASA, accessible surface area (Buried, Partially Buried, Exposed); POS, sequence position (N-Terminal, Middle, C-Terminal)

2.3 Computational procedures

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

$$\Delta\ln\ k_f =\ \ln\ k_f^{\text{mutant}} - \ln\ k_f^{\text{wild}} \qquad (1)$$

where ln k_f^{mutant} and ln k_f^{wild} are natural logarithms of folding rates for mutant and wild-type amino acid residues, respectively. The change in folding rates has been related with the change in properties using correlation coefficient.

2.4 Dataset classification

We observed that none of the shortlisted 103 features exhibited an absolute PCC greater than 0.36 with $\Delta\ln\ k_f$ for the entire set of 790 mutants taken together (Supplementary Fig. S2). Hence, we classified the mutants based on secondary structure (helix, strand, coil), normalized Accessible Surface Area (ASA) (buried, ASA<12%, partially buried, $12 < \text{ASA} \leq 36\%$ and exposed, $\text{ASA} > 36\%$) 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 (Step 2 in Fig. 1). ASA and secondary structure of mutants were assigned using Dictionary of Protein Secondary Structure (DSSP) (Kabsch and Sander, 1983).

2.5 Multiple linear regression and selection of prime three features

For each class prime three features were selected (Step 3 in Fig. 1), which elicit the best relationship with the folding rate change using multiple linear regression technique (Grewal, 1987). The model with the same features has been 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; see Section 2.6.2). The method was also tested with 10-fold cross-validation. Further, the same model has been used for evaluating its performance on a blind test set. We noticed that some features have been selected in different models and hence we have used a total of 51 features in all the models.

The contribution of neighboring residues (ΔP_{seq}) has been included in the method using the equation:

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

where k varies from 0 to 9 residues on both directions.

2.6 Model evaluation and validation

2.6.1 Model performance

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 significance of prediction using R^2 statistic, the F statistic, P value and mean squared error. In addition, the method was tested with mean absolute percentage error (MAPE) and symmetric mean absolute percentage error (sMAPE), which are defined as follows:

$$\text{MAPE} = \sum \frac{|\text{Predicted}\ \Delta\ln\ k_f - \text{Experimental}\ \Delta\ln\ k_f|}{|\ \text{Experimental}\ \Delta\ln\ k_f|} \times 100 \qquad (3)$$

$$\text{sMAPE} = \frac{\sum\ |\text{Predicted}\ \Delta\ln\ k_f - \text{Experimental}\ \Delta\ln\ k_f|}{\sum\ |\text{Predicted}\ \Delta\ln\ k_f + \text{Experimental}\ \Delta\ln\ k_f|} \times 100 \qquad (4)$$

2.6.2 Model validation

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 40-fold cross validation tests were performed with 10^7 iterations each for model validation.

iii. *Blind Test:* An independent test dataset of change in folding rates of 59 protein mutants from the FORA dataset.

3 Results and Discussion

3.1 Model training

The first-order classification of mutants based on secondary structure, ASA and sequential position using three features showed a mean PCC of 0.42 (Fig. 2). On the other hand, the mean PCC rose to 0.55 and 0.81, respectively, in second and third-order classification (simultaneously considering secondary structure, ASA and sequence position together for the relationship with $\Delta\ln\ k_f$). Similarly, there was a marked decrease in mean absolute error from 0.73 to $0.36\,\text{s}^{-1}$ from first to third-order classification. Hence, the predictive power of the method drastically enhanced upon incorporating third-order classification.

We have developed specific models pertaining to each of the 27 classes using simple and multiple linear regression approach (maximum of up to three descriptors). The enhancement in PCC values with increase in the number of descriptors is not significant in single (Supplementary Table S4; Supplementary Fig. S3a) and double-order classification (Supplementary Table S5; Supplementary Fig. S3b). However, in the triple-order classification (Supplementary Table S6; Supplementary Fig. S3c) scheme adopted in the current study, we observed that there was a drastic improvement in PCC upon increase in the number of associated features during multiple linear regression analyses.

3.1.1 Class-wise self-consistency performance of model

We observed that the PCC lies in the range of 0.61–0.98 (mean: 0.81) (Supplementary Fig. S4a) in the considered 27 classes of third-order model. The MAE ranges from 0.10 to $0.58\,\text{s}^{-1}$

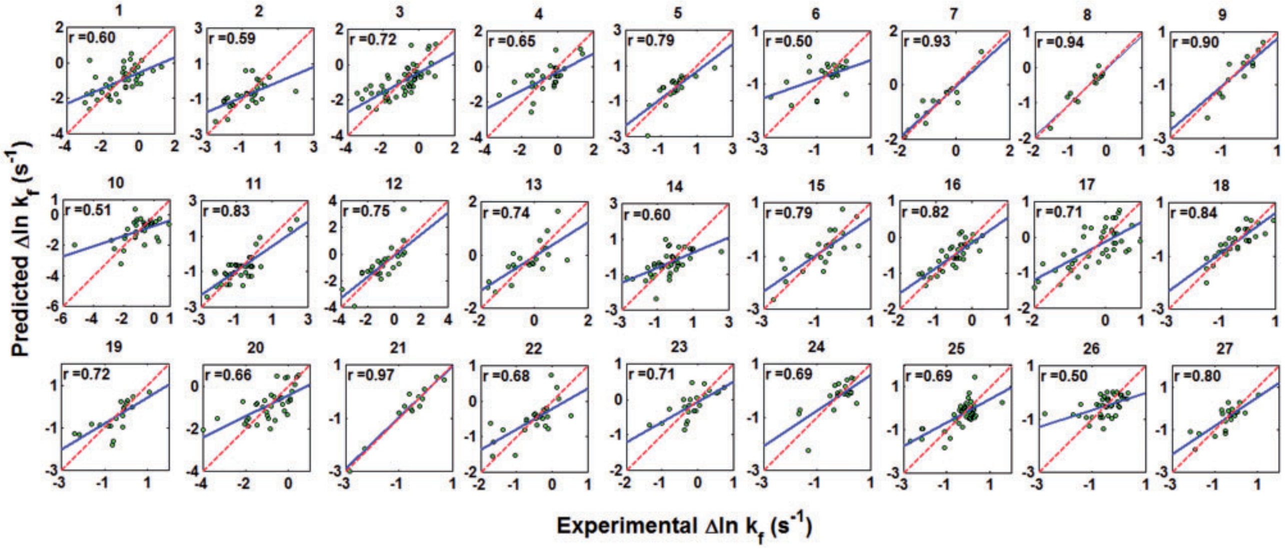

Fig. 3. Class-wise regression plots between the experimental and predicted folding rates change after Jack-Knife cross validation. Dotted line is reference and solid line is regression line. Refer to Supplementary Table S6 for details of each class index

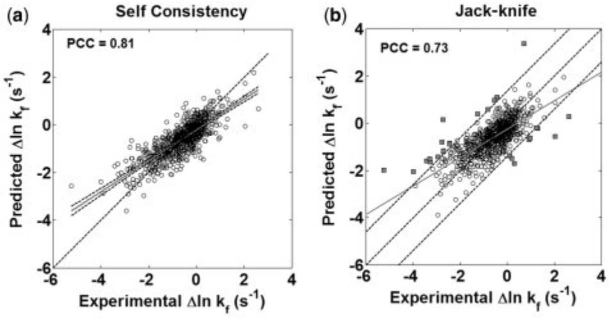

Fig. 4. Unified regression plots for the global dataset of 790 mutants using (**a**) self-consistency and (**b**) Jack-Knife test. The dotted- and solid-lines correspond to the expected 1:1 correlation and the regression lines with 95% confidence interval, respectively. The lines parallel to the regression line are drawn with a deviation of 1.39 s^{-1}. Filled squares outside these lines correspond to the outliers

(Mean: 0.36 s^{-1}). The MAE for accelerating and decelerating mutants are 0.43 and 0.38 s^{-1}, respectively, which have the average Δln k_f of 0.40 and −0.90 s^{-1}. It is important to note that value of PCC is largely dependent on the number and variance of the data in a given dataset. Hence, the class with minimum MAE (0.10 s^{-1}) was identified as the best performing class (i.e. class index 8 in Supplementary Fig. S4b and Supplementary Table S7 in which the mutants belong to 'Strand', 'Exposed' and 'Middle' of the protein sequence). Accuracy (Gromiha and Huang, 2011) of our model (i.e. the model's performance in discriminating between the accelerating and decelerating mutants), ranges from 60 to 100% (Mean: 83.80%). Hence, our model depicts a dual performance of prediction as well as discrimination by adopting a single protocol. For improving the precision of performance, we fine-tuned the prediction using an optimized sequence window length for each class. We noticed that the inclusion of window-lengths improved the mean PCC from 0.75 to 0.81. This indicates the variable influence of neighboring residues in determining the folding rate change of different classes of mutants and accentuates the variable role of short and long range interactions.

We have assessed the statistical significance of the present model and the P value ranges between 10^{-3} and 10^{-11}. Other parameters testifying the significance of prediction are presented in Supplementary Table S8.

3.1.2 Global self-consistency and discriminatory performance

A unified model was developed by combining the prediction results of all the classes for the entire global dataset of 790 mutations. We obtained the PCC of 0.81, which is the same as the mean PCC for class-wise prediction. Further, we have analyzed the performance using iterative elimination of different sets of data with 10^7 iterations and obtained the same average PCC of 0.81 (Supplementary Fig. S5). MAE marginally changed to 0.39 s^{-1} for the equivalent model developed for global dataset as compared to the class wise prediction. This attests to the robustness and flexibility of the cumulative effect of class wise generated models on the entire dataset. The model is also able to discriminate well between accelerating and decelerating mutants with 83% accuracy.

3.2 Model validation

3.2.1 Jack-Knife cross validation

We have evaluated the performance of our method using Jack-Knife test for all the 27 classes. We observed that the PCC is in the range of 0.50–0.97 (mean: 0.73) (Fig. 3), MAE ranges from 0.14 to 0.68 s^{-1} (mean: 0.42 s^{-1}) and accuracy ranges from 60 to 97% (mean: 81.20%) for all 27 classes, which varies slightly from self-consistent mean PCC of 0.81, mean MAE of 0.36 s^{-1} and mean accuracy of 83.80%. The MAE for accelerating and decelerating mutants are 0.51 and 0.44 s^{-1}, respectively, which have the average Δln k_f of 0.40 and −0.90 s^{-1}. The minuscule deviation in the performance before and after Jack-Knife test confirms the robustness of the model as seen in Figure 4a and b, respectively.

3.2.2 n-fold cross validation

Models developed from the global dataset have been examined with 10, 20, 30 and 40-fold cross validation with 10^7 iterations. We obtained a mean PCC of 0.71, which is similar to that obtained in the Jack-Knife test.

3.2.3 Blind set test

The robustness of this model is also evident from the fact that it predicts the changes in folding rates from the FORA dataset (Huang and Gromiha, 2012) comprising of 59 mutants, with a MAE of 0.78 s^{-1}. We noticed that three of them are outliers and their removal reduced the MAE to 0.62 s^{-1}. The higher MAE compared with cross-validation dataset may be attributed with the nature of the proteins in the FORA dataset including α-lactalbumin that has four disulphide bonds in its native structure and tenascin, which is a fibrous protein with a curved Chevron plot. Removal of mutants in these proteins showed a MAE of 0.46 s^{-1}. The comparison with FORA showed a better performance of our method with respect of PCC and MAE despite of the fact that these mutants were a part of training set of FORA (Supplementary Table S9a).

3.2.4 Comparison of the method with FORA

FORA showed a mean PCC of 0.53, MAE of 0.50 s^{-1} and accuracy of 81.54% after cross-validation in a set of 467 mutants (Supplementary Table S9b). However, it did not perform equally well in a blind assessment carried out on a non-redundant dataset of 568 mutants (PCC: 0.20, MAE: 0.75 s^{-1} and accuracy: 75.22%) (Supplementary Table S9c). It clearly shows that the performance of the present method is better than the only available other method, FORA.

3.2.5 Prediction performance in different proteins

We have evaluated the performance of the present method in all the 26 proteins and the results are presented in Supplementary Table S10. We observed that the MAE is in the range of 0.27 to 0.76 s^{-1}. This result reveals that the present method is not depending on a specific protein instead it can be applied to all the two-state proteins.

3.3 Feature set description

3.3.1 Sequence-based amino acid feature set

We found that the changes in folding rates upon point mutations are primarily affected by the following order of amino acid features: hydrophobicity ≫ secondary structure propensity ≈ physicochemical > thermodynamic > conformational ≫ evolutionary, emphasizing the dominant influence of hydrophobic effect during the folding process. The list of properties, which are identified as important in multiple regression equations are highlighted in italics in Supplementary Table S3. Further, the properties, which have the single correlation of more than 0.6 with experimental change in folding rates within a class, are highlighted in bold. These include unfolding hydration heat capacity change (Oobatake and Ooi, 1993), equilibrium constant with reference to the ionization property of COOH group (Zimmerman *et al.*, 1968), residue volume (Bigelow, 1967), SWEIG index (Cornette *et al.*, 1987), normalized frequency of coil (Nagano, 1973), unfolding entropy change of chain (Oobatake and Ooi, 1993) and long range nonbonded energy per atom (Oobatake and Ooi, 1977), which belong to physical, chemical, energetic and conformational categories, and highlights the necessity of combining properties for predicting the changes in folding rates.

3.3.2 Structure-based amino acid feature set

The structures of all the 790 point mutants were modeled employing homology modeling with the aid of MODELLER v9.7 (Šali and Blundell, 1993), fold recognition with Phyre2 (Kelley and Sternberg, 2009) and *ab initio* with I-Tasser (Zhang, 2008) protocols. However, the mutant structure were minimally different from the wild type (all atom RMSD < 1 Å). All relevant structural descriptors such as contact order, long range order, number of hydrogen bonds, total contact index, multiple contact index and surrounding hydrophobicity for different spatial and sequence separations, exhibited only minor variations between wild type and mutant structures. None of the combinations of structural features exhibited PCC greater than 0.63 in any classification.

3.4 Interpretation of outliers

We find that the classes with a correlation coefficient below 0.75 were profoundly affected by some outlying data points. We opted for Tukey's box plot approach (McGill *et al.*, 1978) to identify outliers as it is less sensitive to extremities of the data. In order to decipher the biological relevance of such mutants, we performed rigorous analysis of outliers at two levels: (i) global dataset and (ii) protein wise.

3.4.1 Outliers in the global dataset

Twelve of the 27 classes were found to be affected by the outliers in class wise analysis (Supplementary Fig. S6). In class index 7, we observe that PCC is highly affected by the outliers despite exhibiting only minimal changes in the mean absolute deviation. This is another reason for giving preeminence to MAE than PCC in analyzing the model's relative performance. Similar analysis, performed on global dataset led to the detection of 20 statistically significant outliers out of total 27. MAE between their experimental and predicted change in folding rates after Jack-Knife test was greater than 1.39 s^{-1}. The PCC markedly improved to 0.83 from 0.73 and MAE dropped to 0.35 from 0.42 s^{-1} after removal of all the outliers (and after Jack-Knife test) confirming the authenticity of detected outliers. The MAPE value is 194.23 and 166.05% as well as sMAPE value is 31.16 and 24.87%, respectively, with and without outliers. In the test set, the MAPE value is 205.89 and 201.65% as well as sMAPE value is 49.08 and 40.78%, respectively with and without outliers. The high MAPE is attributed with the low value of $\Delta\ln k_f$ in which a small deviation would cause a high MAPE (Equation 3). The outlying mutants belonged to three categories: (i) aromatic to aliphatic mutations (e.g. Phe to Val), i.e. truncation of the aromatic ring, creating a void and severely affecting local packing, (ii) mutations involving glycine (e.g. Gly to Arg): significantly affecting the backbone flexibility, (iii) large side chain truncations in aliphatic residues (e.g. Ile to Ser) (Supplementary Table S11).

3.4.2 Protein-wise identification of outliers

For all of the 26 single domain proteins, the outlying mutations identified using an absolute deviation cut-off of 1.39 s^{-1} were scrutinized at the tertiary structure level to gain a physical insight into the origins of this behavior (Supplementary Table S12; Supplementary Fig. S7). We observed that the structural and energetic factors are dominant for accounting the folding rates of these mutants.

In colicin E9 immunity protein Im9, the mutant F15A is identified as a potential outlier (Fig. 5a). The exclusion of this mutant drastically increased the PCC from 0.78 to 0.87 and decreased MAE from 0.47 to 0.35 s^{-1}.

One of the striking observations with such aromatic outliers was their role as a mediator in forming an aromatic core triad or dyad in their respective proteins. For example in Im9 protein, a phenylalanine triad is observed in the protein core comprising of three phenylalanine residues (F15, F40 and F83) from three different regions, N-terminal, middle and C-terminal, respectively (Fig. 5b);

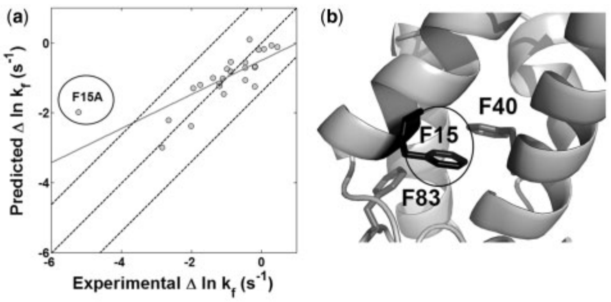

Fig. 5. (**a**) Influence of F15A as an outlier in Im9 (**b**) Core aromatic triad comprising of N-terminal F15 acting as a bridge between the middle F40 and C-terminal F83 (PDB Id: 1IMQ)

the experimental mutation position F15 (outlier), seems to act as a bridge between the F40 and F83, thereby bringing the distant regions of the sequence close in space and hence packing the structure. Other aromatic outliers were also found to form such triads in acyl coenzyme A binding protein (PDB ID: 2ABD), fyn tyrosine kinase SH3 (PDB ID: 1SHF) and acylphosphatase (PDB ID: 1APS) and an aromatic dyad in protein L (PDB ID: 1HZ6) (Supplementary Table S12). These observations suggest the prediction of rates upon mutations involving these regions is complicated by the presence of multiple structural and energetic factors that is challenging to be taken into account by the simple model developed in this work.

3.5 Limitations of this model

Though we have presented a simple procedure to predict the changes in rates upon mutations, it still has some deficiencies due to the knowledge-based approach. First, the training dataset mainly comprises of only specific types of mutations and hence the prediction will be less accurate upon some uncommon mutations. Second, only 22% of the total dataset comprise of accelerating mutants due to which the model is intrinsically less sensitive to such effects. Finally, the detailed nature of the model led to an increase in the number of parameters, as both events are mutually exclusive. Hence, in order to increase the prediction accuracy despite of various constraints imposed by the experimental dataset, we had to compromise with the parameter count.

3.6 Web server development

A web server named ‘Folding RaCe’ has been developed for the prediction of change in folding rates upon point mutations. The user is required to provide the PDB ID or protein structure file in PDB format, chain ID, the position and the mutation to be introduced (e.g. F15A). The output includes the change in folding rate upon the mutation along with the details of secondary structure, ASA and sequential position of the mutant. It is freely available at http://www.iitm.ac.in/bioinfo/proteinfolding/foldingrace.html.

3.7 Conclusion

We present a multiple linear regression approach to predict the changes in folding rates upon point mutations in two-state proteins. It involves a combination of rigorous classification of mutations incorporating both structural (secondary structure and ASA) and sequence-based aspects of proteins, and heuristic and knowledge based feature selection of amino acid descriptors. The model could predict with a correlation coefficient of 0.73, an absolute error of 0.42 s^{-1} and an accuracy of 81.2% with the experimental changes in rates after cross validation and outperforms the only available method FORA. The symmetric mean absolute percentage error rates (sMAPE) for the cross-validation and test datasets is 31.16 and 49.08%, respectively. We have developed a web server named ‘*Folding RaCe*’ that incorporates the features discussed. The structural implications of outliers have also been discussed. We expect this method and server to be of importance to experimentalists in identifying the position and nature of mutations that change the rates to the desired extent that can have both folding-mechanistic and functional implications.

Acknowledgements

We thank Protein Bioinformatics Laboratory, Department of Biotechnology, Bioinformatics Infrastructure Facility, P.G. Senapathy Center for Computing Resources and Indian Institute of Technology, Madras for computational facilities. PC thanks Saket Gurukar for helping her during web server development. We would like to thank the anonymous reviewers for their constructive comments.

Funding

Department of Science and Technology (SR/SO/BB-0036/2011) for research grant to MMG and Ministry of human resource and development (MHRD) for HTRA scholarship to P.C.

Conflict of Interest: none declared.

References

Bigelow,C.C. (1967) On the average hydrophobicity of proteins and the relation between it and protein structure. *J. Theor. Biol.*, **16**, 187–211.

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

Cheng,X. *et al.* (2013) Swfoldrate: Predicting protein folding rates from amino acid sequence with sliding window method. *Proteins Struct. Funct. Bioinf.*, **81**, 140–148.

Cornette,J.L. *et al.* (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.*, **195**, 659–685.

Dehouck,Y. *et al.* (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics*, **12**, 151.

De Sancho,D. *et al.* (2009) Protein folding rates and stability: How much is there beyond size? *J. Am. Chem. Soc.*, **131**, 2074–2075.

Fersht,A.R. *et al.* (1992) The folding of an enzyme: I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.*, **224**, 771–782.

Grewal,P.S. (1987) Numerical Methods of Statistical Analysis. Sterling Publishers, New Delhi.

Gromiha,M.M. (2005) A statistical model for predicting protein folding rates from amino acid sequence with structural class information. *J. Chem. Inf. Model.*, **45**, 494–501.

Gromiha,M.M. *et al.* (2006) FOLD-RATE: prediction of protein folding rates from amino acid sequence. *Nucleic Acids Res.,* **34**, W70–W74.

Gromiha,M.M. (2009) Multiple contact network is a key determinant to protein folding rates. *J. Chem. Inf. Model.*, **49**, 1130–1135.

Gromiha,M.M. and Huang,L.-T. (2011) Machine learning algorithms for predicting protein folding rates and stability of mutant proteins: comparison with statistical methods. *Curr. Protein Pept. Sci.*, **12**, 490–502.

Gromiha,M.M. and Selvaraj,S. (2001) 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**, 27–32.

Hall,M. *et al.* (2009) The WEKA Data mining software: an update. *SIGKDD Explor. Newsl.*, **11**, 10–18.

Henry,E.R. and Eaton,W.A. (2004) Combinatorial modeling of protein folding kinetics: free energy profiles and rates. *Chem. Phys.*, **307**, 163–185.

Huang,L.-T. *et al.* (2007) iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics,* **23,** 1292–1293.

Huang,L.-T. and Gromiha,M.M. (2008) Analysis and prediction of protein folding rates using quadratic response surface models. *J. Comput. Chem.*, **29**, 1675–1683.

Huang,L.-T. and Gromiha,M.M. (2010) First insight into the prediction of protein folding rate change upon point mutation. *Bioinformatics,* **26,** 2121–2127.

Huang,L.-T. and Gromiha,M.M. (2012) Real value prediction of protein folding rate change upon point mutation. *J. Comput. Aided. Mol. Des.*, **26**, 339–347.

Ivankov,D.N. and Finkelstein,A. V. (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc. Natl Acad. Sci. USA,* **101,** 8942–8944.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kawashima,S. and Kanehisa,M. (2000) AAindex: Amino Acid index database. *Nucleic Acids Res.,* **28,** 374.

Kelley,L.A. and Sternberg,M.J.E. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.*, **4**, 363–371.

Lin,G. *et al.* (2010) SeqRate: sequence-based protein folding type classification and rates prediction. *BMC Bioinformatics*, **11**, 1–9.

Makarov,D.E. *et al.* (2002) How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. *Proc. Natl Acad. Sci. USA,* **99,** 3535–3539.

McGill,R. *et al.* (1978) Variations of box plots. *Am. Stat.*, **32**, 12–16.

Micheletti,C. (2003) Prediction of folding rates and transition-state placement from native-state geometry. *Proteins Struct. Funct. Bioinformatics*, **51**, 74–84.

Muñoz,V. and Eaton,W.A. (1999) A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA,* **96,** 11311–11316.

Naganathan,A.N. (2013) A rapid, ensemble and free energy based method for engineering protein stabilities. *J. Phys. Chem. B*, **117**, 4956–4964.

Naganathan,A.N. and Muñoz,V. (2010) Insights into protein folding mechanisms from large scale analysis of mutational effects. *Proc. Natl Acad. Sci. U.S.A.,* **107,** 8611–8616.

Naganathan,A.N. and Muñoz,V. (2004) Scaling of folding times with protein size. *J. Am. Chem. Soc.*, **127**, 480–481.

Nagano,K. (1973) Logical analysis of the mechanism of protein folding: I. Predictions of helices, loops and *β*-structures from primary structure. *J. Mol. Biol.*, **75**, 401–420.

Oobatake,M. and Ooi,T. (1977) An analysis of non-bonded energy of proteins. *J. Theor. Biol.*, **67**, 567–584.

Oobatake,M. and Ooi,T. (1993) Hydration and heat stability effects on protein unfolding. *Prog. Biophys. Mol. Biol.*, **59**, 237–284.

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

Parthiban,V. *et al.* (2006) CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.,* **34,** W239–W242.

Plaxco,K.W. *et al.* (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.*, **277**, 985–994.

Punta,M. and Rost,B. (2005) Protein folding rates estimated from contact predictions. *J. Mol. Biol.*, **348**, 507–512.

Šali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Schrödinger,L. (2010) The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC, New York.

Schymkowitz,J. *et al.* (2005) The FoldX web server: an online force field. *Nucleic Acids Res.,* **33,** W382–W388.

Thirumalai,D. (1995) From minimal models to real proteins: time scales for protein folding kinetics. *J. Phys. I*, **5**, 1457–1467.

Witten,I.H. and Frank,E. (2005) Data Mining: Practical Machine Learning Tools and Techniques. 2nd edn. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Yin,S. *et al.* (2007) Modeling backbone flexibility improves protein stability estimation. *Structure*, **15**, 1567–1576.

Zhang,Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics*, **9**, 40.

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

Zimmerman,J.M. *et al.* (1968) The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.*, **21**, 170–201.