J Comput Aided Mol Des (2012) 26:339–347
DOI 10.1007/s10822-012-9560-3

Real value prediction of protein folding rate change upon point mutation

Liang-Tsung Huang · M. Michael Gromiha

Received: 21 February 2011 / Accepted: 2 March 2012 / Published online: 18 March 2012

Abstract Prediction of protein folding rate change upon amino acid substitution is an important and challenging problem in protein folding kinetics and design. In this work, we have analyzed the relationship between amino acid properties and folding rate change upon mutation. Our analysis showed that the correlation is not significant with any of the studied properties in a dataset of 476 mutants. Further, we have classified the mutants based on their locations in different secondary structures and solvent accessibility. For each category, we have selected a specific combination of amino acid properties using genetic algorithm and developed a prediction scheme based on quadratic regression models for predicting the folding rate change upon mutation. Our results showed a 10-fold cross validation correlation of 0.72 between experimental and predicted change in protein folding rates. The correlation is 0.73, 0.65 and 0.79, respectively in strand, helix and coil segments. The method has been further tested with an extended dataset of 621 mutants and a blind dataset of 62 mutants, and we observed a good agreement with experiments. We have developed a web server for predicting the folding rate change upon mutation and it is available at http://bioinformatics.myweb.hinet.net/fora.htm.

Keywords Protein folding rate · Mutation · Amino acid properties · Long-range contact · Quadratic regression model

Electronic supplementary material The online version of this article (doi:10.1007/s10822-012-9560-3) contains supplementary material, which is available to authorized users.

L.-T. Huang
Department of Biotechnology, Mingdao University, Changhua 523, Taiwan

M. M. Gromiha (✉)
Department of Biotechnology, Indian Institute of Technology (IIT) Madras, Chennai 600 036, Tamilnadu, India
e-mail: gromiha@iitm.ac.in

Introduction

Protein folding is a process by which a polypeptide chain of amino acid residues folds into a specific three-dimensional structure. Another related parameter to protein folding is protein folding rate, which is a measure to understand the tendency of folding (slow/fast) from unfolded state to its native three-dimensional structure. Studies of protein folding rates enhance our understanding of the variations in protein folding kinetics, which may lead to several pathologies such as prion and Alzheimer diseases [1–4]. Hence, several investigations have been carried out to measure protein folding rates as well as to predict protein folding rates using structural information and/or just from amino acid sequence [5–8].

The experimental data on folding rates of proteins and mutants have been accumulated in two major kinetic databases, such as protein folding database, PFD [9] and Protein Folding Kinetics Database, kineticDB [10]. In addition, we have collected the data for several new mutants and set up a dataset with 467 unique mutants, which can be used for understanding/predicting the folding rate change upon amino acid substitution [11].

The data available in kinetic databases as well as those reported in the literature has been used to predict protein folding rates. The classical methods include the concepts of contact order [12], long-range order [13], total contact distance [14], cliquishness [15], and multiple contact index [16]. Further, several methods have been reported to predict protein folding rates from amino acid sequence. These

methods include the relationship with amino acid properties [17–20], predicted secondary structures [21], predicted inter-residue contacts [22], amino acid composition [23, 24], secondary structure length [25], hybrid sequence representation [26], and flexibility and solvent accessibility [27]. The details of computational methods for predicting protein folding rates have been reviewed recently [11, 28].

All the above mentioned studies are focused on predicting the folding rates of proteins without any modifications of amino acids. In fact the substitution of amino acid residues in a protein alters the folding, stability, specificity and functions of proteins [29–34]. The influence of amino acid substitutions on protein stability has been studied extensively and several methods have been proposed to predict protein stability change upon mutation [24, 35–44]. On the other hand, prediction of Φ values based on the change in the logarithm of folding rate by the change in stability due to mutation has been reported [45, 46]. However, studies on the influence of amino acid mutations to protein folding rates are still at infant stage.

In our earlier work, we have set up a dataset of 467 mutants and developed a method to discriminate the mutants that accelerate or decelerate the folding process [11]. In this work, we have focused on the real value prediction of protein folding rate change upon amino acid substitution. We have systematically analyzed the influence of 49 physical, chemical, energetic and conformational properties of amino acid residues as well as long-range contacts to the folding rate change of 467 protein mutants. The dataset has been classified into different groups based on the location of mutants in helix, strand and coil regions as well as at different ranges of solvent accessibility. We have proposed a method based on quadratic regression models for predicting the folding rate change upon point mutation and our method showed a 10-fold cross-validation correlation of 0.72 between experimental and predicted folding rates. We have developed a web server for predicting the folding rate change of protein mutants and it is available at http://bioinformatics.myweb.hinet.net/fora.htm .

Materials and methods

Experimental folding rates

We have constructed a dataset (F467) of protein mutants with experimental k_f values and relevant features. F467 was obtained with the following conditions: (i) all single mutants, (ii) two-state folding proteins and (iii) k_f values are extrapolated to zero concentration (i.e. water). The folding rate change upon single mutation was calculated by $\Delta k_f = k_f^{mutant} - k_f^{wild}$, where k_f^{mutant} and k_f^{wild} are k_f values for mutant and wild-type residues, respectively. F467 consisted of 467 unique mutants from 15 different proteins and the detailed information about the mutants is given in Supplementary Tables S1 and S2. Further, we have derived three datasets (162S, 155H and 150C) from F467, which contained 162, 155 and 150 mutants in strand, helix and coil regions, respectively.

Descriptions for the prediction model

Properties of amino acids and amino acid pairs

We used a set of 49 diverse amino acid properties (physical–chemical, energetic and conformational), which fall into various clusters analyzed by Tomii and Kanehisa [47] in the present study. The amino acid properties (i.e. vector $\boldsymbol{p}$) were normalized between 0 and 1 using the equation:

$$\boldsymbol{p}^a = \frac{\boldsymbol{p}^a_{ori} - \boldsymbol{p}_{\min}}{\boldsymbol{p}_{\max} - \boldsymbol{p}_{\min}}, \quad (1)$$

where $\boldsymbol{p}^a$ and $\boldsymbol{p}^a_{\text{ori}}$ are, respectively, the normalized and original property vectors of amino acid a; and $\boldsymbol{p}_{\min}$ and $\boldsymbol{p}_{\max}$ are, respectively, the minimum and maximum vectors for each property. The numerical and normalized values for all the 49 properties used in this study along with their brief descriptions have been explained in our earlier articles [48, 49] and are available at http://www.cbrc.jp/~gromiha/fold_rate/property.html.

Furthermore, we quantified various physicochemical and biochemical properties of amino acid pairs from AAindex database version 9.1 [50, 51]. Firstly, the amino acid substitution of each mutant was calculated to 87 properties from the substitution matrices of AAindex2 section. Secondly, the difference of statistical contact potentials between the mutant and wild-type protein sequences was calculated from AAindex3 section by the following equation: $\Delta CP = CP_{\text{mutant}} - CP_{\text{wild}}$, where CP_{mutant} is the potential of the mutant residue in contact with the first neighboring residue along the N- or C-terminus; CP_{wild} is the potential of the wild type residue in contact with the first neighboring residue along the N- or C-terminus. Finally, all the properties were normalized using Eq. (1).

Secondary structure (SS) and solvent accessibility (SA)

Secondary structure and solvent accessibility are important parameters to predict protein mutant stability and binding sites in protein complexes [52]. We have utilized these parameters in the present work and we obtained the information for all the wild type residues from DSSP, Dictionary of Secondary Structure of Proteins [53]. The occurrence of mutants based on solvent accessibility and secondary structure are presented in Supplementary Figure S1. Moreover, the

distribution of mutants at different ranges of solvent accessibility is displayed in Supplementary Figure S2.

Long-range contact (LRC)

The residues in a protein molecule are represented by their α-carbon atoms. Using the C_α coordinates, a sphere of radius 8Å is fixed around each residue and the residues occurring in this volume are identified. The composition of surrounding residues is analyzed in terms of the location at the sequence level and the contributions from ±3 or ±4 residues are termed as medium range contacts and >±4 residues are treated as long-range contacts [54, 55]. The mutant distribution based on long-range contacts is shown in Supplementary Figure S3.

Quadratic regression models (QRMs)

We have developed regression models [24] for predicting the real value of protein folding rate change. These models are denoted by a quadratic form:

$$\widehat{y} = b_0 + \sum_{j=1}^{p} b_j x_j + \sum_{j=1}^{p} \sum_{k=j}^{p} b_{jk} x_j x_k + e, \tag{2}$$

where $\widehat{y}$ is the output variable of folding rate change; b_0, b_j and b_{jk} are regression coefficients; x_j and x_k are relevant features; p is the total number of features; uncontrolled factors and errors are modeled by e. Given n independent instances $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)$, the model becomes an n-by-m system of equations and the estimates of the model coefficients are determined by using the least square method.

The method establishes the relationship between input and output variables by polynomial equations, which are non-linear but low-order model. Further inference can be made by well-known regression analysis.

Performance measurement and validation

In this work, we have used both correlation coefficient and mean absolute error (MAE) to assess the prediction performance of folding rate change. The correlation coefficient between the experimental and predicted values (Δk_f) is calculated using

$$R = \frac{n\sum_{i=1}^{n} y_i\hat{y}_i - \left(\sum_{i=1}^{n} y_i \sum_{i=1}^{n} \hat{y}_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]\left[n\sum_{i=1}^{n} \hat{y}_i^2 - \left(\sum_{i=1}^{n} \hat{y}_i\right)^2\right]}}, \tag{3}$$

where R is the correlation coefficient, n, y_i, and $\widehat{y}_i$ are the total number of mutants, experimental and predicted folding rate change, respectively; and i varies from 1 to n. The mean absolute error (MAE) is defined as the absolute difference between experimental and predicted values:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y}_i|. \tag{4}$$

The present method was validated by n-fold cross-validation test. The n-fold cross-validation partitions samples into n subsamples chosen randomly with approximately equal size. For each sub-sample, the method trains a regression model from the remaining data and uses it to predict the folding rate change of the sub-sample. The procedure is repeated n times to obtain the mean measure. In this work, both 10-fold and 20-fold cross-validation tests were implemented.

Additionally, we calculated and displayed the confidence interval of the prediction to evaluate the level of certainty. The confidence level (CI) of the k-th predicted value with a given probability of $(1-\alpha)$ is given by the following equation:

$$\mathrm{CI}_k = \widehat{y}_k \pm t_{\frac{\alpha}{2},n-2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x_k-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]}, \tag{5}$$

where $\widehat{y}_k$ is the k-th predicted value of folding rate change; $t_{\frac{\alpha}{2},n-2}$ is the inverse of Student's t cumulative distribution function (CDF) with $\frac{\alpha}{2}$ probability and n-2 degrees of freedom; and $\widehat{\sigma}^2$ is the mean square error (MSE) of $\widehat{y}$ with respect to y.

Although this work is focused on the real value prediction of folding rate change, we also assessed the discrimination performance of folding rate change by using several measures (sensitivity, specificity and accuracy). Sensitivity shows the performance of the method to correctly identify the true positives (TP) whereas specificity reveals the ability of the method to exclude true negatives (TN). Accuracy indicates the overall performance.

$$\mathrm{Sensitivity} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}), \tag{6}$$

$$\mathrm{Specificity} = \mathrm{TN}/(\mathrm{TN}+\mathrm{FP}), \tag{7}$$

$$\mathrm{Accuracy} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}), \tag{8}$$

where FN and FP are the false negatives and false positives, respectively.

Results and discussions

Amino acid properties and protein folding rate change

We have analyzed the relationship between amino acid properties and protein folding rate change upon mutation

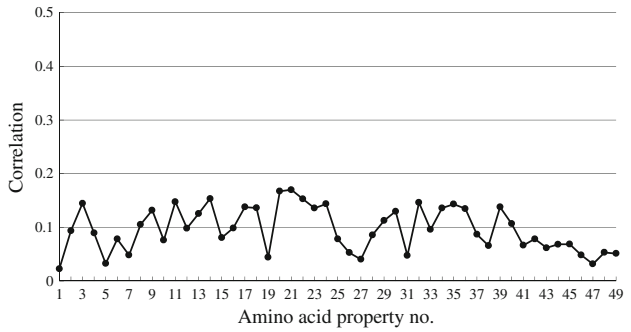

Fig. 1 Highest absolute correlation between experimental Δk_f and 49 properties on dataset F467

using wild type, mutant and neighboring residues of the mutant site. The correlation coefficients obtained with 49 amino properties are shown in Fig. 1. Most of the properties show low correlation coefficient and the r-value varies from 0.02 to 0.17. This result shows the folding rate change by mutation is not attributed with any single amino acid property. Hence, we have given different levels of priority to the properties according the r-value and selected the properties as input parameters (Collection 1) for predicting protein folding rate change.

Long-range contact (LRC) and protein folding rate change

We have analyzed the relationship between long-range contact and folding rate change, and the results are shown in Fig. 2. Apparently, the information about long-range contacts has higher influence than any of the considered amino acid properties. However, the contribution is not very large and the highest correlation is 0.363 for the interval 41–50 followed by 0.279–0.235 for the intervals >50 and 11–20, respectively. We included the long-range contacts in these three intervals as probable features

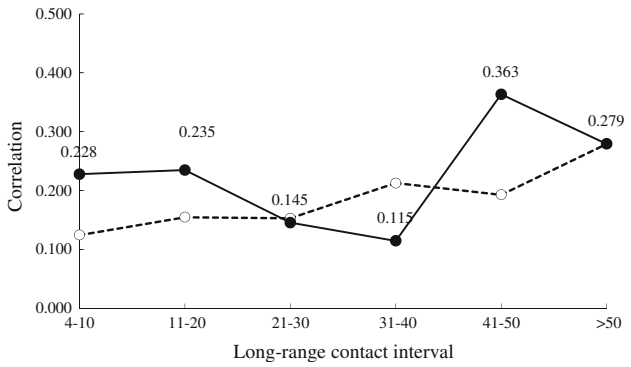

Fig. 2 Highest absolute correlation between experimental Δk_f and long-range contact (LRC). The *black* and *white circles* denote LRC for individual intervals and cumulative intervals, respectively

(Collection 2) for predicting the folding rate upon mutation.

Development of prediction models for folding rate change

We have utilized the parameters, amino acid properties (Collection 1), long-range contacts at different intervals (Collection 2), secondary structure information (helix, strand and coil) and different ranges of solvent accessibility (0–2 %, 2–20 %, 20–50 % and >50 %) for predicting the folding rate change. Quadratic regression models have been constructed for different secondary structures using the various combinations of parameters and optimized by genetic algorithm to achieve the highest prediction performance. The optimization procedure of input parameters has been described in the supplementary material.

In Table 1, we list the selected input parameters for predicting folding rate change of mutants in helical, strand and coil regions. In order to avoid over-fitting we have restricted the number of input parameters. There are seven parameters in which six of them are based on amino acid sequence and long-range contact is obtained with structural information. The amino acid properties are obtained from three aspects, wild type, mutant and neighboring residues. For the dataset 162S, we utilized pH_i, α_c and $-T\Delta S_h$ of wild type residue, N_s of mutant residue and average of K^0 and ΔG_c for three neighboring residues and we observed that the features are different for the three secondary structures. Interestingly, the parameter α_c (power to be at the C-terminal) is included in both 162S and 150C, showing the importance of it for the prediction of mutants with strand and coil structures. Similarly, the parameter H_t (Thermodynamic transfer hydrophobicity) appears twice and improves the prediction performance of mutants in the 155H dataset.

Prediction performance

The performance of our prediction method has been assessed with the correlation between experimental and predicted folding rate change and the results obtained with 10-fold and 20-fold cross-validation tests are presented in Table 2. We observed that the present method could predict the folding range change with a correlation coefficient of 0.72. The correlation coefficients for the mutants in helical, strand and coil regions are 0.65, 0.73 and 0.79, respectively. Additionally, Table 2 shows several measures (accuracy, sensitivity and specificity) to assess the performance of discriminating folding change rate (accelerating or decelerating) upon mutation. The average accuracy is 67.4 % and the mutants in strand showed the highest specificity (77.7 %) and accuracy (74.7 %) than those in other types of secondary structures.

Table 1 Input parameters used for predicting the folding rate change of proteins upon mutations

Dataset	Input parameter			
	Sequence based			Structure based
	Wild type	Mutant	Neighbor	
162S[a] (Strand)	pH_i, $\boldsymbol{\alpha_c}$, $-T\Delta S_h$	N_s	K^0, ΔG_c	LRC_5
155H[b] (Helix)	$\boldsymbol{H_t}$, ΔG_{ph}	α_m	$\boldsymbol{H_t}$, R_f, μ	LRC_2
150C[c] (Coil)	F	E_t	H_p, B_r, $\boldsymbol{\alpha_c}$, G_{hN}	LRC_6

[a] pH_i: Isoelectric point; α_c: Power to be at the C-terminal of α-helix; $-T\Delta S_h$: Unfolding entropy change of hydration; N_s: Average number of surrounding residues; K^0: Compressibility; ΔG_c: Unfolding Gibbs free energy of chain; LRC_5: Long-range contacts at interval 41–50

[b] H_t: Thermodynamic transfer hydrophobicity; ΔC_{ph}: Unfolding hydration heat capacity change; α_m: Power to be at the middle of α-helix; R_f: Chromatographic index; μ: Refractive index; LRC_2: Long-range contacts at interval 11–20

[c] F: Mean root-mean-square fluctuational displacement; E_t: Total nonbonded energy; H_p: Surrounding hydrophobicity; B_r: Buriedness; G_{hN}: Gibbs free energy change of hydration for native protein; LRC_6: Long-range contacts at interval >50. More details are available in Gromiha et al. [48]

Table 2 Prediction performance for different datasets based on secondary structure using 10-fold cross-validation test

Results obtained by 20-fold cross-validation test are shown in parentheses; [1]Correlation coefficient based on absolute change (Δk_f); [2]Correlation coefficient based on logarithmic change ($\Delta \ln(k_f)$)

Measure	Secondary structure			Weighted average
	Strand	Helix	Coil	
Number	162	155	150	
R^1	0.73 (0.72)	0.65 (0.65)	0.79 (0.77)	0.72 (0.72)
R^2	0.62	0.54	0.44	0.54
MAE	30.7	193.8	49.4	90.8
Sensitivity (%)	42.9	59.5	42.9	48.4
Specificity (%)	77.7	70.3	63.1	70.6
Accuracy (%)	74.7	67.7	59.3	67.4

Figure 3 shows the experimental and predicted folding rate changes of the mutants in three different secondary structures. We removed 6 mutants from F467 to emphasize the range of experimental Δk_f between $-1{,}000$ and $+1{,}000$. Also, the 95 % confidence interval has been calculated, showing the degree of uncertainty to the predicted observation.

Friel et al. [56] measured the folding rates of 32 mutants in Im7 protein. We have applied our method for predicting the folding rate change upon mutation, which are kept in the test set. The results are presented in Fig. 4. Interestingly, the folding rates of most of the mutants are well predicted and the correlation coefficient between experimental and predicted folding rate change is 0.80.

Influence of solvent accessibility for predicting the folding rate change upon mutation

We have analyzed the influence of solvent accessibility for predicting the folding rate change upon mutation by classifying them at various ranges of solvent accessibility. The results obtained with four classifications based on buried (<2 %), partially buried (2–20 %), partially exposed (20–50 %) and exposed (>50 %) mutants are presented in Table 3. We observed that the correlation is good only for partially exposed/buried regions and the r-value is less than 0.60 in buried and exposed regions. Further, the SA based sub-classification of mutants in secondary structures did not increase the correlation in any of the three secondary structures (Table 4). Conversely, the classification reduced the correlation in strand and coil regions. This analysis revealed that the solvent accessibility is not an important factor for the predicting the folding rates of mutant proteins. Similar result has also been reported in the classification of accelerating and decelerating mutants. This trend is opposite to the general behavior of solvent accessibility that it is one of the major parameters for predicting the thermal stability of protein mutants [57, 58] and the binding sites of protein complexes [59].

Comparison with other methods

Machine learning algorithms are widely employed in different applications of bioinformatics. The present work is the first method for predicting protein folding rate change upon mutation. Hence, we have made an attempt to compare the present method with other machining learning algorithms, such as support vector machines [60] (using a

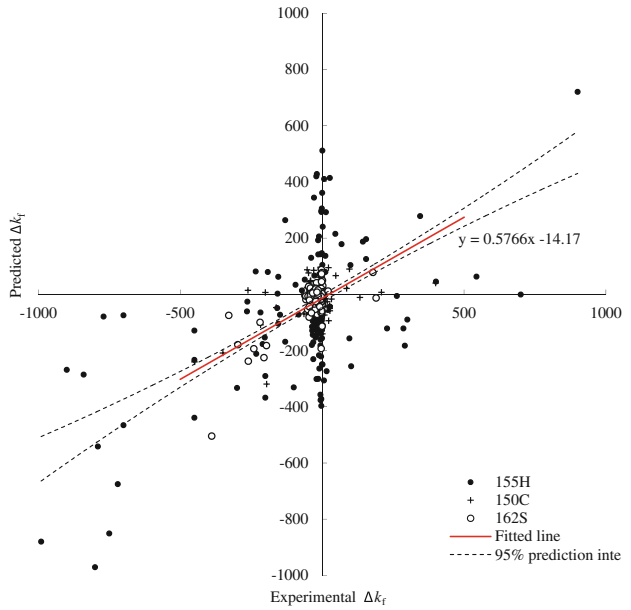

Fig. 3 Relationship between experimental and predicted Δk_f values for the mutants in three secondary structures. 460 data points lie in the range between −1,000 and 1,000 are shown

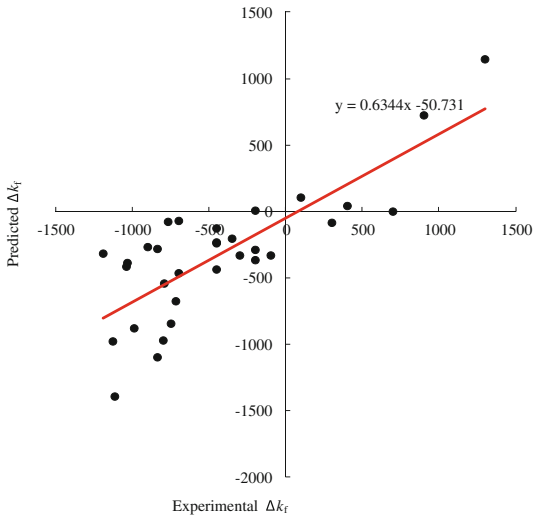

Fig. 4 Relationship between experimental and predicted Δk_f values for protein 1IMQ (r = 0.80)

parameter v to control the number of support vectors, called v-support vector regression [61]), radial basis function networks [62], artificial neural networks [63], K* instance-based learner [64], linear regression and ridge linear regression [65]. The details of the learning methods have been described in the supplementary material. The results obtained with 10-fold and 20-fold cross-validation tests for four popular machine learning methods are presented in Table 5. We observed that the present method has the highest correlation coefficient of 0.71 whereas other

Table 3 Prediction results using the data at various ranges of solvent accessibility

ASA range	Number	R	MAE
0–2	94	0.57	121.67
2–20	129	0.87	59.68
20–50	153	0.65	75.53
>50	91	0.59	128.94
Weighted average		0.68	90.85

MAE mean absolute error based on absolute change (Δk_f)

Table 4 Prediction results based on the classification of solvent accessibility and secondary structure by 10-fold cross-validation test

ASA range	Secondary structure					
	Strand		Helix		Coil	
	R	MAE	R	MAE	R	MAE
0–2	0.34	33.50	0.52	293.95	0.72	63.13
2–20	0.83	26.42	0.84	141.76	0.94	41.97
20–50	0.78	33.09	0.65	166.88	0.20	41.24
>50	0.28	28.21	0.61	190.12	0.42	69.31
Weighted average	0.63	30.71	0.65	193.80	0.53	49.41

MAE mean absolute error based on absolute change (Δk_f)

methods showed a correlation between −0.05 and 0.37. This analysis revealed the importance of our method for predicting the folding rate change upon mutation.

Prediction of protein folding rate change with large and blind data sets

Naganathan and Munoz [66] collected the folding rates of 800 mutants in a set of two-state proteins. We have carefully surveyed the data and included an addition data of 154 mutants, which have complete information for our method (Supplementary Table S3). The final dataset contains 621 mutants in which 559 are used for training/cross-validation (C559) and the rest of them (62 mutants) are kept for blind test (B62). The blind data are marked in Supplementary Tables S1 and S3.

We have made an attempt to predict protein folding rates using the properties of amino acids amino acid pairs. First, we trained our model with all the 621 mutants using 173 amino acid properties obtained from AAindex database. These properties have been examined using genetic algorithm (details are given in supplementary material) and extracted a set of 7 properties, which are given below: amino acid properties (1) wild type *pHi*, (isoelectric point), (2) wild type B_1, (bulkiness) and (3) *pK'* (equilibrium constant with reference to the ionization property of COOH group) of 3-neighboring residues; amino acid pairs

Table 5 Comparison of the present method with four other methods using 10-fold cross-validation test

Measure	Method						
	Present	SVM (*v*-SVR)	RBF	ANN	K*	LR	RLR
R for 162S	0.73 (0.72)	−0.12 (−0.14)	−0.25 (−0.14)	0.26 (0.22)	0.07 (0.05)	0.31 (0.31)	0.28 (0.29)
R for 155H	0.65 (0.65)	−0.10 (−0.31)	0.11 (0.10)	0.50 (0.51)	0.28 (0.27)	0.002 (−0.01)	0.28 (0.28)
R for 150C	0.76 (0.75)	−0.18 (−0.20)	0.01 (0.02)	0.34 (0.10)	0.35 (0.34)	0.07 (0.04)	0.14 (0.16)
Average	0.71 (0.71)	−0.13 (−0.21)	−0.05 (−0.01)	0.37 (0.28)	0.23 (0.22)	0.13 (0.12)	0.24 (0.24)

The results by 20-fold cross-validation test are shown in parentheses

SVM (*v*-SVR): Support vector machine using *v*-support vector regression

RBF radial basis function network, *ANN* artificial neural network, *K** Instance-based learner, *LR* linear regression, *RLR* ridge linear regression

Table 6 Performance of present method in a large dataset of 621 mutants

Measure	C559			C559 + B62	
	10-fold	20-fold	Blind[2]	10-fold	20-fold
R[1]	0.58	0.62	0.79	0.63	0.63
MAE[1]	0.73	0.71	0.72	0.73	0.72
Sensitivity (%)	34.4	32.3	45.5	40.2	38.3
Specificity (%)	90.1	90.9	90.2	90.3	90.3
Accuracy (%)	80.5	80.9	82.3	81.6	81.3

[1] Correlation coefficient and mean absolute error are based on logarithmic change ($\Delta\ln(k_f)$); [2]Trained on C559 and tested with B62

obtained from the amino acid substitution matrices: (4) context-dependent optimal substitution matrices [67] and (5) structure-conservation scoring tables [68]; the difference of statistical contact potentials in contact with the first neighboring residue along the (6) N-terminus [69] and (7) C-terminus [70].

The 10-fold and 20-fold cross-validation accuracies for C559 and C559 + B62 are presented in Table 6. We observed that the results are similar to those obtained with a dataset of F467 mutants and structural information. We have also examined the validity of our model using a blind data set of B62 mutants trained with C559 mutants, which showed an average correlation of 0.79 between experimental and predicted change in protein folding rates.

Folding rate versus logarithm of protein folding rate change

We have trained our model for predicting the $\Delta\ln(k_f)$ upon mutation and the results are presented in Tables 2 and 6. We noticed that the results are comparable to those obtained with change in folding rates. Further inspection of Table 2 showed that the correlation coefficients (R) between experimental and predicted $\Delta\ln(k_f)$ are less than those obtained with Δk_f. This might be due to the accumulation of large number of data with narrow range of change in folding rates. This result reveals the necessity of further refinement to account the small changes in folding rates upon mutation.

Limitation of our method

The present method is developed with all the available data of folding rate change upon mutation. Careful inspection of the data showed the presence of vast amount of data on protein folding rate change in a narrow range of −500–500/s and −300–300/s. The model obtained with the data in this range reduced the correlation between experimental and predicted change in folding rates (Supplementary Figure S4). Hence, further refinements are necessary to account such small change. Secondly, the predicted results of $\Delta\ln(k_f)$ showed that the correlation coefficient is less than that obtained with $\Delta(k_f)$. Finally, the method is mainly for predicting the change in real value folding rates of proteins. Hence, the accuracy of discriminating the accelerating and decelerating mutants is less that that obtained with the model specifically developed for discrimination [11].

Prediction on the web

We have developed a web server for predicting the real value of protein folding rate change upon point mutation (FORA). Figure 5a shows a snapshot of the input page with necessary information (sequence and structure). The sequence information includes wild type, mutant and three neighboring residues, and structural information involves secondary structure and long-range contact of the wild type residue. We have chosen F50V mutant of chymotrypsin inhibitor 2 (PDB code: 2CI2) as an example and the output page is shown in Fig. 5b. It contains the input parameters and predicted folding rate change along with the characteristics of residues, such as composition, polarity and metabolic role. In this case, the predicted folding rate change is −35.4971/s, which agrees well with the experimental observation (−35.56/s). The server is freely available online at http://bioinformatics.myweb.hinet.net/fora.

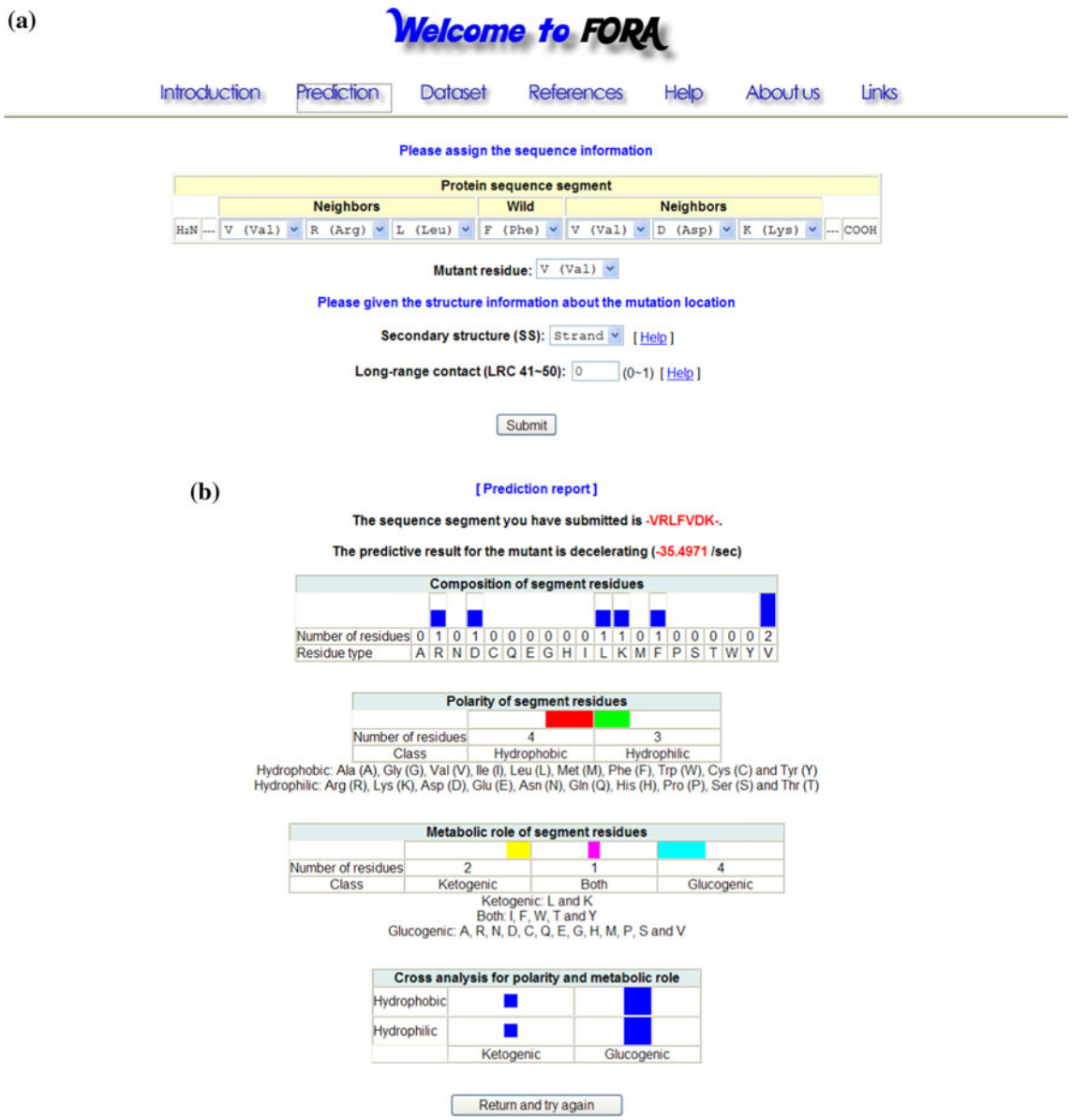

Fig. 5 (**a**) A snapshot showing the input page of the server for real value prediction of folding rate change upon single mutation. (**b**) The output page with input parameters predicted folding rate change and characteristics of amino acid residues

htm, and it is linked with a related web server FREEDOM [11], which discriminates the accelerating and decelerating mutants (http://bioinformatics.myweb.hinet.net/freedom.htm).

Conclusions

We have systematically analyzed the relationship between various amino acid properties and folding rate change of proteins upon point mutations. Our results suggested that unlike protein mutant stability none of the properties could explain the change of protein folding rate upon amino acid substitution. However, the combination of amino acid properties along with long-range contacts showed a correlation of 0.72 between experimental and predicted folding rate change upon mutation in a dataset of 467 mutants. The present method was also tested with a larger dataset of 612 mutants and a blind set of 62 mutants, and we observed a good agreement with experimental protein folding rates. The classification of mutants based on secondary structure improved the correlation in strand and coil regions. Unexpectedly, the classification based on solvent accessibility did not improve the correlation. We have developed a web server for predicting the folding rate change upon mutation, which is freely available for the users. We suggest that our method could be an effective tool for predicting the folding rate change of mutant proteins.

Acknowledgments We thank the anonymous reviewers for their constructive comments to improve the manuscript.

References

1. Apetri AC, Surewicz K, Surewicz WK (2004) J Biol Chem 279(17):18008
2. Capriotti E, Casadio R (2007) Bioinformatics 23(3):385
3. Jenkins DC, Pearson DS, Harvey A, Sylvester ID, Geeves MA, Pinheiro TJ (2009) Eur Biophys J 38(5):625
4. Hart T, Hosszu LL, Trevitt CR, Jackson GS, Waltho JP, Collinge J, Clarke AR (2009) Proc Natl Acad Sci USA 106(14):5651
5. Maxwell KL, Wildes D, Zarrine-Afsar A, De Los Rios MA, Brown AG, Friel CT, Hedberg L, Horng JC, Bona D, Miller EJ, Vallee-Belisle A, Main ER, Bemporad F, Qiu L, Teilum K, Vu ND, Edwards AM, Ruczinski I, Poulsen FM, Kragelund BB, Michnick SW, Chiti F, Bai Y, Hagen SJ, Serrano L, Oliveberg M, Raleigh DP, Wittung-Stafshede P, Radford SE, Jackson SE, Sosnick TR, Marqusee S, Davidson AR, Plaxco KW (2005) Protein Sci 14(3):602
6. Jackson S (1998) Fold Des 3(4):R81
7. Gromiha MM, Huang LT (2011) Curr Protein Pept Sci 12(6):490
8. Gromiha MM (2010) Protein Bioinformatics: From Sequence to Function. Academic Press, Singapore
9. Fulton KF, Devlin GL, Jodun RA, Silvestri L, Bottomley SP, Fersht AR, Buckle AM (2005) Nucleic Acids Res 33(Database issue):D279
10. Bogatyreva NS, Osypov AA, Ivankov DN (2009) Nucleic Acids Res 37(Database issue):D342
11. Huang L-T, Gromiha MM (2010) Bioinformatics 26(17):2121
12. Plaxco KW, Simons KT, Baker D (1998) J Mol Biol 277(4):985
13. Gromiha MM, Selvaraj S (2001) J Mol Biol 310(1):27
14. Zhou H, Zhou Y (2002) Biophys J 82(1 Pt 1):458
15. Micheletti C (2003) Proteins 51(1):74
16. Gromiha MM (2009) J Chem Inf Model 49(4):1130
17. Gromiha MM (2003) J Chem Inf Comput Sci 43(5):1481
18. Gromiha MM (2005) J Chem Inf Model 45(2):494
19. Huang JT, Tian J (2006) Proteins 63(3):551
20. Gromiha MM, Thangakani AM, Selvaraj S (2006) Nucleic Acids Res 34(Web Server issue):W70
21. Ivankov DN, Finkelstein AV (2004) Proc Natl Acad Sci USA 101(24):8942
22. Punta M, Rost B (2005) J Mol Biol 348(3):507
23. Ma BG, Guo JX, Zhang HY (2006) Proteins 65(2):362
24. Huang LT, Gromiha MM (2008) J Comput Chem 29(10):1675
25. Huang JT, Cheng JP, Chen H (2007) Proteins 67(1):12
26. Jiang Y, Iglinski P, Kurgan L (2009) J Comput Chem 30(5):772
27. Gao J, Zhang T, Zhang H, Shen S, Ruan J, Kurgan L (2010) Proteins 78(9):2114
28. Gromiha Selvaraj (2008) Curr Bioinforma 3(1):1
29. Prabakaran P, An J, Gromiha MM, Selvaraj S, Uedaira H, Kono H, Sarai A (2001) Bioinformatics 17(11):1027
30. Porter CT, Bartlett GJ, Thornton JM (2004) Nucleic Acids Res 32(Database issue):D129
31. Lopez G, Valencia A, Tress M (2007) Nucleic Acids Res 35(Database issue):D219
32. Kumar MD, Gromiha MM (2006) Nucleic Acids Res 34(Database issue):D195
33. Gromiha MM, Yabuki Y, Suresh MX, Thangakani AM, Suwa M, Fukui K (2009) Nucleic Acids Res 37(Database issue):D201
34. Gromiha MM, An J, Kono H, Oobatake M, Uedaira H, Sarai A (1999) Nucleic Acids Res 27(1):286
35. Guerois R, Nielsen JE, Serrano L (2002) J Mol Biol 320(2):369
36. Bordner AJ, Abagyan RA (2004) Proteins 57(2):400
37. Capriotti E, Fariselli P, Calabrese R, Casadio R (2005) Bioinformatics 21(Suppl 2):ii54
38. Cheng J, Randall A, Baldi P (2006) Proteins 62(4):1125
39. Yin S, Ding F, Dokholyan NV (2007) Nat Methods 4(6):466
40. Bromberg Y, Yachdav G, Rost B (2008) Bioinformatics 24(20):2397
41. Huang LT, Gromiha MM (2009) Bioinformatics 25(17):2181
42. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M (2009) Bioinformatics 25(19):2537
43. Carlsson J, Soussi T, Persson B (2009) FEBS J 276(15):4142
44. Gao S, Zhang N, Duan GY, Yang Z, Ruan JS, Zhang T (2009) Hum Mutat 30(8):1161
45. Munoz V, Eaton WA (1999) Proc Natl Acad Sci USA 96(20):11311
46. Weikl TR (2005) Proteins 60(4):701
47. Tomii K, Kanehisa M (1996) Protein Eng 9(1):27
48. Gromiha MM, Oobatake M, Sarai A (1999) Biophys Chem 82(1):51
49. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A (2000) J Biomol Struct Dyn 18(2):281
50. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M (2008) Nucleic Acids Res 36(Database issue):D202
51. Kawashima S, Ogata H, Kanehisa M (1999) Nucleic Acids Res 27(1):368
52. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A (1999) Protein Eng 12(7):549
53. Kabsch W, Sander C (1983) Biopolymers 22(12):2577
54. Gromiha MM, Selvaraj S (1997) J of Biol Phys 23(3):151
55. Gromiha MM, Selvaraj S (2004) Prog Biophys Mol Biol 86(2):235
56. Friel CT, Capaldi AP, Radford SE (2003) J Mol Biol 326(1):293
57. Gromiha MM (2007) Biochem Soc Trans 35(Pt 6):1569
58. Gromiha MM, Huang LT (2011) Curr Protein Pept Sci 12(6):490
59. Ahmad S, Gromiha MM, Sarai A (2004) Bioinformatics 20(4):477
60. Chang C–C, Lin C-J (2001):http://www.csie.ntu.edu.tw/∼cjlin/libsvm
61. Wu T-F, Lin C-J, Weng R (2004) J Mach Learn Res 5:975
62. Moody J, Darken C (1989) Neural Comput 1(2):281
63. Rumelhart DE, Hinton GE, Williams RJ (1986) In Parallel distributed processing: explorations in the microstructure of cognition, vol. 1. Cambridge, MA, USA, MIT Press, pp 318
64. Cleary JG, Trigg LE (1995) In Proceedings of the 12th international conference on machine learning. San Francisco: Morgan Kaufmann, pp 108
65. Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning : data mining, inference, and prediction, 2nd edn. Springer, New York, NY
66. Naganathan AN, Munoz V (2010) Proc Natl Acad Sci USA 107(19):8611
67. Koshi JM, Goldstein RA (1995) Protein Eng 8(7):641
68. Luthy R, McLachlan AD, Eisenberg D (1991) Proteins 10(3):229
69. Tobi D, Shafran G, Linial N, Elber R (2000) Proteins 40(1):71
70. Miyazawa S, Jernigan R (1985) Macromolecules 18(3):534