BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 17 2010, pages 2121–2127
doi:10.1093/bioinformatics/btq350

Structural bioinformatics

Advance Access publication July 8, 2010

First insight into the prediction of protein folding rate change upon point mutation

Liang-Tsung Huang[1] and M. Michael Gromiha[2,*]

[1]Department of Information Communication, Mingdao University, Changhua 523, Taiwan and [2]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: The accurate prediction of protein folding rate change upon mutation is an important and challenging problem in protein folding kinetics and design. In this work, we have collected experimental data on protein folding rate change upon mutation from various sources and constructed a reliable and non-redundant dataset with 467 mutants. These mutants are widely distributed based on secondary structure, solvent accessibility, conservation score and long-range contacts. From systematic analysis of these parameters along with a set of 49 amino acid properties, we have selected a set of 12 features for discriminating the mutants that speed up or slow down the folding process. We have developed a method based on quadratic regression models for discriminating the accelerating and decelerating mutants, which showed an accuracy of 74% using the 10-fold cross-validation test. The sensitivity and specificity are 63% and 76%, respectively. The method can be improved with the inclusion of physical interactions and structure-based parameters.

Availability: http://bioinformatics.myweb.hinet.net/freedom.htm

Contact: michael-gromiha@aist.go.jp

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

Received on April 16, 2010; revised on June 4, 2010; accepted on June 25, 2010

1 INTRODUCTION

Protein folding is a process by which a polypeptide chain of amino acid residues folds into a specific 3D structure. The rate of protein folding is a measure to understand the tendency of folding (slow/fast) from unfolded state to its native 3D structure, and it varies several orders of magnitude, ranging from microseconds to an hour (Bogatyreva *et al*., 2009; Fulton *et al*., 2005; Jackson, 1998). 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's diseases. Hence, several investigations have been carried out to measure protein folding rates, and the data are accumulated in different databases, such as protein folding database (PFD; Fulton *et al*., 2005) and protein folding kinetics database, kineticDB (Bogatyreva *et al*., 2009). These databases provide annotated structural, methodological, kinetic and thermodynamic data of proteins and mutants.

Currently, experimental data for protein folding rates are available for less than 100 proteins, which demand the necessity of computational methods to predict protein folding rates from 3D structures of proteins and/or just from amino acid sequence. Plaxco *et al*. (1998) proposed the first method based on contact order to relate protein folding rates with the total number of contacts in protein structures. Gromiha and Selvaraj (2001) introduced the concept of long-range order, which relates long-range contacts (LRC, contact between two residues that are close in space and far in the sequence) in protein structures with protein folding rate. Consequently, several parameters such as total contact distance (Zhou and Zhou, 2002), cliquishness (Micheletti, 2003), multiple contact index (Gromiha, 2009), etc. have been developed to understand protein folding rates using 3D structures of proteins.

On the other hand, several methods have been proposed for predicting protein folding rates from amino acid sequence with/without the structural class information. These methods include the relationship with amino acid properties (Gromiha, 2003; Gromiha, 2005; Gromiha *et al*., 2006; Huang and Tian, 2006), predicted secondary structures (SSs) (Ivankov and Finkelstein, 2004), predicted inter-residue contacts (Punta and Rost, 2005), amino acid composition (Ma *et al*., 2006; Huang and Gromiha, 2008), SS length (Huang *et al*., 2007a) and hybrid sequence representation (Jiang *et al*., 2009). Gromiha and Selvaraj (2008) reviewed the details of computational methods for predicting protein folding rates.

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 drastically alters the folding, stability, specificity and functions of proteins (Gromiha *et al*., 1999a; Gromiha *et al*., 2009; Kumar and Gromiha, 2006; Lopez *et al*., 2007; Porter *et al*., 2004; Prabakaran *et al*., 2001). The problems related to protein stability and function upon point mutations have been addressed in earlier studies (Chea and Livesay, 2007; Gromiha, 2007; Holliday *et al*., 2009; Pugalenthi *et al*., 2008; Sankararaman *et al*., 2010). Consequently, several methods have been proposed to predict protein stability and function due to amino acid replacements (Bordner and Abagyan, 2004; Bromberg *et al*., 2008; Capriotti *et al*., 2005; Carlsson *et al*., 2009; Cheng *et al*., 2006; Dehouck *et al*., 2009; Gao *et al*., 2009; Guerois *et al*., 2002; Huang and Gromiha, 2008; Huang and Gromiha, 2009; Yin *et al*., 2007). However, the influence of amino acid mutations to protein

*To whom correspondence should be addressed.

folding rates has not yet been explored. Hence, it is essential and important to understand and predict protein folding rates upon amino acid substitutions.

In this work, we have collected the experimental data on protein folding rates upon point mutation and compiled a dataset. It contains many mutants that are not included in PFD (Fulton *et al*., 2005) and kineticDB (Bogatyreva *et al*., 2009), and the newly compiled dataset would serve as a useful resource for further studies. We have developed a method using quasi-regression models and various features such as amino acid properties, SS, solvent accessibility (SA), conservation score and LRC's for discriminating the effect of folding rates (increase or decrease; in other words, accelerating or decelerating mutants). Our method showed an accuracy of 74% using the 10-fold cross-validation test with a specificity of 76% and sensitivity of 63% in a dataset of 467 mutants. We have developed a web server for distinguishing the accelerating and decelerating mutants and it is available at http://bioinformatics.myweb.hinet.net/freedom.htm.

2 METHODS

2.1 Protein mutant dataset

In this work, we have originally constructed a dataset (F467) of protein mutants with experimental k_f values and relevant features. The detailed information about the mutants is given in Supplementary Table S1. F467 consists of 467 unique mutants from 15 different proteins. Most of the proteins have 15 to 68 mutants and the number of mutants in each protein is shown in Supplementary Figure S1. The data were obtained from the careful search of published reports in the literature as well as from two freely accessible kinetic databases, PFD (Fulton *et al*., 2005) and kineticDB (Bogatyreva *et al*., 2009). The original references for all the mutants in F467 dataset are listed in Supplementary Table S2.

F467 is obtained with the following conditions: (i) all single mutants, (ii) kinetic type two; and (iii) k_f values are extrapolated to zero concentration (i.e. water). The folding rate change upon single mutation is calculated by $\Delta k_f = k_f^{\text{mutant}} - k_f^{\text{wild}}$, where k_f^{mutant} and k_f^{wild} are k_f values for mutant and wild-type residues, respectively. The Δk_f ranges between -1193.60/s and 1481.34/s, and the number of accelerating and decelerating mutants is 79 and 338, respectively. It may be noted that some mutants cause aggregation/precipitation and the proteins may not fold at all. Under these circumstances, the protein folding kinetic databases miss the data on protein folding rates upon mutations. Such data are not considered in the present study.

2.2 SS and SA

SS and SA are important parameters to predict protein mutant stability and binding sites in protein complexes (Gromiha *et al*., 1999b). We have utilized these parameters in the present work and we obtained the information for all the wild type residues from the Dictionary of SS of Proteins (DSSP; Kabsch and Sander, 1983).

2.3 LRC

The residues in a protein molecule are represented by their α-carbon atoms. Using the C_α coordinates, a sphere of radius 8 Å 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 to ± 4 residues as medium range contacts and more than ± 4 residues are treated as LRCs (Gromiha and Selvaraj, 1997; Gromiha and Selvaraj, 2004).

2.4 Conservation score

We have used the program Scorecons for computing the conservation score for all the mutants in the considered proteins (Valdar, 2002). The target sequence has been searched against non-redundant protein sequences deposited in SWISS-PROT 57.0 (Boutet *et al*., 2007) using BLAST (Altschul *et al*., 1990). We have used the default parameters (BLOSUM62 and the threshold of 10) for BLAST search. Using the results obtained with BLAST, we have performed multiple sequence alignment with MAFFT program (Katoh *et al*., 2005). The aligned sequences have been utilized to compute the conservation score for all the amino acid residues.

2.5 Amino acid properties

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 (1996). The amino acid properties (i.e. vector **p**) were normalized between 0 and 1 using the expression

$$\mathbf{p}^a = \frac{\mathbf{p}^a_{\text{ori}} - \mathbf{p}_{\min}}{\mathbf{p}_{\max} - \mathbf{p}_{\min}}, \tag{1}$$

where $\mathbf{p}^a$ and $\mathbf{p}^a_{\text{ori}}$ are the normalized and original property vectors, respectively, of amino acid a; and $\mathbf{p}_{\min}$ and $\mathbf{p}_{\max}$ the minimum and maximum vectors, respectively, 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 (Gromiha *et al*., 1999c; Gromiha *et al*., 2000) and are available at http://www.cbrc.jp/~gromiha/fold_rate/property.html.

The average property value of a segment in a protein sequence, $\mathbf{p}_{\text{ave}}$, was computed using the standard formula

$$\mathbf{p}_{\text{ave}} = \frac{\sum_{i=1}^{r} \mathbf{p}_i}{r}, \tag{2}$$

where $\mathbf{p}_i$ is the property vector of i-th residue and the summation is over r, the total number of residues in the segment.

2.6 Quadratic regression models

Regression models are prediction methods, which establish the relationship between input and output variables by the polynomial equation. In this study, we proposed regression models for discriminating folding rate change by a quadratic form

$$y = b_0 + \sum_{j=1}^{p} b_j x_j + \sum_{j=1}^{p} \sum_{k=j}^{p} b_{jk} x_j x_k + e, \tag{3}$$

where y is the output variable of folding rate change; b_0, b_j and b_{jk} are regression coefficients; x_j and x_k the relevant features; p the total number of features; uncontrolled factors and errors are modeled by e. Given n independent observations $(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^n, y^n)$, the model becomes an n-by-m system of equations

$$\begin{bmatrix} y^1 \\ \vdots \\ y^n \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{x}^1) & \cdots & f_m(\mathbf{x}^1) \\ \vdots & \ddots & \vdots \\ f_1(\mathbf{x}^n) & \cdots & f_m(\mathbf{x}^n) \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_m \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}, \tag{4}$$

where y^g is the output variable value of folding rate change for the g-th ($g = 1, \ldots, n$) observation; $\mathbf{x}^g = (x_1^g, \ldots, x_p^g)$ the input vector of p variable values for the g-th ($g = 1, \ldots, n$) observation; $f_h(\mathbf{x}^g)$ the h-th ($h = 1, \ldots, m$) transferred term of the model (e.g. terms are $f_1(\mathbf{x}^1) = 1$, $f_2(\mathbf{x}^1) = x_1^1$ and $f_3(\mathbf{x}^1) = x_2^1$), and the value of m is calculated as $m = 1 + p + [p(p+1)/2]$; c_h and e_g the corresponding coefficient and error values, respectively. Therefore, the estimates of the model coefficients are determined by using the least-square method, which minimizes the statistics derived from errors.

The present method has several main advantages of predicting folding rate change. (i) The quadratic regression model (QRM) is a non-linear but low-order model. Thus, it builds more complex and accurate models using relatively few observations. (ii) The lower complexity of polynomial equations is helpful to reduce computational cost and time. (iii) Further inference can be carried out by well-known regression analysis.

2.7 Performance measurement

The discrimination of the folding rate change (accelerating/decelerating) can be regarded as one of the binary classification problems. Several measures of prediction performance are regularly used for such classifications.

In this work, we have used overall accuracy (Q2), sensitivity (SE), specificity (SP) and Matthews correlation coefficient (MCC; Baldi *et al*., 2000) for distinguishing the folding rate change of mutants. These measures have been calculated by using the following expressions:

$$Q2 = \frac{TP+TN}{TP+FP+TN+FN} \times 100\%, \quad SE = \frac{TP}{TP+FN} \times 100\%,$$

$$SP = \frac{TN}{TN+FP} \times 100\%, \text{ and}$$

$$MCC = \frac{TP \times TN + FP \times FN}{\sqrt{(TN+FN) \times (TN+FP) \times (TP+FN) \times (TP+FP)}},$$

where TP, TN, FP and FN refer to the number of true positives, true negatives, false positives and false negatives respectively. Further, receiver operating characteristic (ROC) curves, which plot the true positive rate against the false positive rate, are provided for showing the tradeoff between sensitivity and specificity (Sonego *et al*., 2008). The area under the curve (AUC) is calculated to summarize a curve in a single quantity.

We have carried out both jack-knife (leave-one-out cross-validation) and *n*-fold cross-validation tests for validating the present method. Jack-knife test systematically constructs the coefficients of a prediction model by leaving one-out observation at a time from the dataset, and then predicts the folding rate change of the omitted mutant. For *n*-fold cross-validation test, the dataset is divided into *n* subsets chosen randomly with approximately equal size. A prediction model is built with $n-1$ subsets of data and the remaining subset of mutants is used for predicting the folding rate change. The procedure has been repeated for *n* times to obtain the mean measure.

3 RESULTS

3.1 Distribution of mutants based on SA, conservation score and LRCs

We have analyzed the distribution of protein mutants based on various features used in the present study. The variation of SA for the mutants in F467 dataset is shown in Supplementary Figure S2. We noticed that the mutants are located at various ranges of SA. Most of the mutants (47.8%) are located in buried/partially buried region (SA<=20%) whereas the partially exposed ($20\% < SA \leq 50\%$) and exposed ($SA > 50\%$) regions accommodate 32.8% and 19.5% of mutants, respectively.

The analysis on LRCs shows that the mutants are widely distributed with the normalized LRCs between 0 and 1 (Supplementary Fig. S3). The observed conservation score for all the 467 mutants showed that the mutants are not populated with any specific conservation score. On the other hand, the mutants are more or less equally distributed between 0.3 and 1.0 with the intervals of 0.1 (Supplementary Fig. S4).

The systematic analysis of various features in all the mutants used in the present study reveals that the dataset is not biased with any specific features and it will be a useful resource for further investigations on protein folding rates.

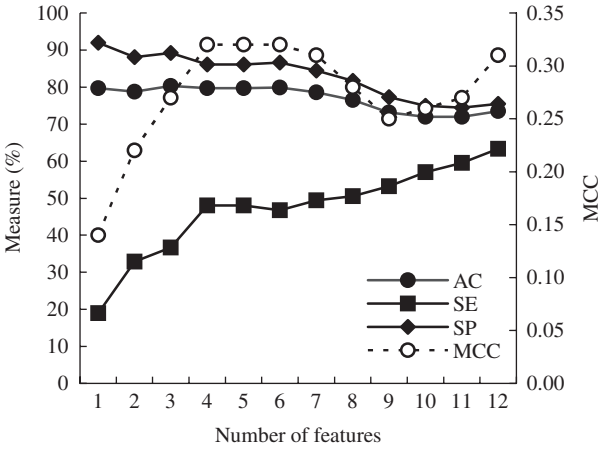

Fig. 1. Improvement of discrimination performance with numbers of features.

3.2 Discrimination of accelerating and decelerating proteins upon point mutations

We have used various sequence and structure-based parameters and QRMs for discriminating the accelerating and decelerating mutants. The parameters wild type and mutant residues, 49 amino acid properties, conservation score, SS, SA and LRCs have been used as the initial set of parameters.

Further, these parameters have been optimized to obtain the highest accuracy of discrimination with minimum number of parameters. We have utilized genetic algorithm for feature selection and the technical details are described in Supplementary Material. The prediction performance with the increase of features is shown in Figure 1. Interestingly, the combination of 12 selected features showed the highest accuracy/MCC as well as the balance between sensitivity and specificity for discriminating the accelerating and decelerating mutants. It can also be seen from Figure 1 that a single feature showed high specificity, whereas the sensitivity is poor. The final selected parameters are given in Table 1. The present method utilizes only 12 parameters for discriminating 467 mutants. Among 49 amino acid properties 10 of them are used for discrimination [E_l and ΔG_{ph} of wild type residue, E_l , $-T\Delta S_h$ and ΔH of mutant residue, and average of M_w, E_l, N_s, s and ΔG, unfolding Gibbs free energy change (Oobatake and Ooi, 1993) for the three neighboring residues on both directions].

Interestingly, the information on E_1 (long-range non-bonded energy) is used for wild type, mutant and neighboring residues, showing the importance of long-range interactions for understanding the folding rate change upon mutation.

We have analyzed the importance of each feature by computing the discrimination accuracy without it. The results are shown in Supplementary Table S3. Our result showed that the omission of a single feature decreases the accuracy by 1–6% and specifically, the parameters E_l and N_s highly reduced the prediction accuracy.

Machine learning algorithms are commonly used for classifying proteins based on their structure and function (Sonego *et al*., 2007). The discrimination results obtained with the present method

Table 1. Features used for discriminating the accelerating and decelerating mutants

Feature number	Parameter name	Residue position
1	E_l	Wild residue
2	ΔG_{ph}	
3	E_l	Mutant residue
4	$-T\Delta S_h$	
5	ΔH	
6	M_w	Three neighbors
7	E_l	
8	N_s	
9	ΔG	
10	s	
11	LRC	Wild residue
12	Conservation score	

Table 2. Comparison of our method with other methods by the 10-fold cross-validation test

Measure	Method				
	Present	SVM	RBF	ANN	K*
Accuracy (%)	73.5 (71.7)	83.1 (83.1)	82.4 (83.3)	77.1 (77.9)	74.7 (75.2)
Sensitivity (%)	63.3 (58.2)	– (–)	12.7 (16.5)	19.0 (20.3)	21.5 (21.5)
Specificity (%)	75.5 (74.5)	100 (100)	96.6 (96.9)	88.9 (89.7)	85.6 (86.1)
MCC	0.31 (0.27)	– (–)	0.16 (0.22)	0.09 (0.11)	0.07 (0.08)

The results obtained with jack-knife test are shown in parentheses; SVM, support vector machine; RBF, radial basis function network; ANN, artificial neural network; K*, instance-based learner.

as well as other widely used methods, support vector machines (Chang and Lin, 2001), radial basis function networks (Moody and Darken, 1989), artificial neural networks (Rumelhart *et al*., 1986), and K* instance-based learner (Cleary and Trigg, 1995) are presented in Table 2. Technical details of these methods are described in Supplementary Material. We noticed that the present method discriminated the accelerating and decelerating mutants with an accuracy of 74% and 72% using the 10-fold cross-validation and jack-knife tests, respectively. The sensitivity is 63% and 58% with the specificity of 76% and 75%, respectively. It is noteworthy that the specificity and sensitivity are close to each other and the average is the balance between these two terms.

We have also analyzed the prediction performance of other popular methods, such as neural networks, support vector machines, radial basis function networks and k-nearest neighbors using the same features. The results obtained with these methods indicated that these machine learning methods are biased with high specificity (86–100%), whereas the sensitivity is just 0–22%. Further, support vector machines with various kernels predicted the mutants as decelerating ones, the most abundant class in the dataset.

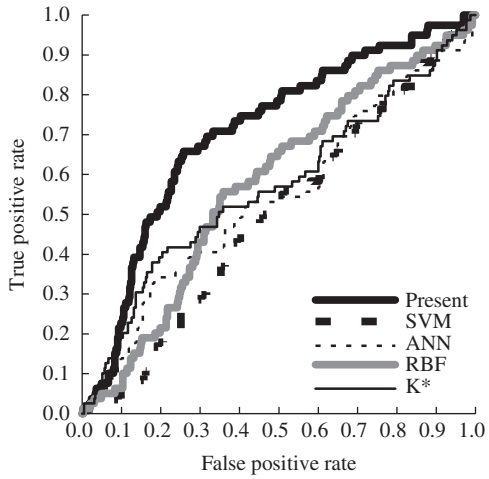

Fig. 2. ROC curves for different methods obtained with the 10-fold cross-validation test.

In Figure 2, we show the ROC curves obtained with the present method along with those obtained with other methods. We found that our method shows an AUC value of 0.71, whereas the AUC values are 0.50, 0.54, 0.58 and 0.57 for support vector machines, neural networks, radial basis function networks and k-nearest neighbor methods, respectively.

3.3 Prediction performance at different SSs and various ranges of SA

We have analyzed the prediction accuracy at different SSs and various ranges of SA, and the results are presented in Supplementary Table S4. We noticed that the mutants at strand regions are better predicted than those at helical and coil regions. It has been reported that the neighboring residues influence the helical and coil segments, whereas the strands are highly influenced with LRCs (Gromiha and Selvaraj, 2004). Hence the inclusion of LRCs and the property, long-range interaction energy enhanced the discrimination accuracy of mutants at strands.

We have analyzed the influence of SA to discriminate the decelerating and accelerating mutants. The inclusion of SA as a feature did not improve the discrimination performance, and in fact it decreased the accuracy by 3%. This result indicates that the change of protein folding rates upon mutations did not depend on the location of residues and the replacement of amino acid residues at different locations may have similar effects on protein folding rates.

The classification of mutants based on SA is also included in Supplementary Table S4. We noticed that the accuracy lies between 67% and 78% for the mutants in all the regions. Further, the mutants in buried, partially buried and partially exposed regions are well discriminated between accelerating and decelerating mutants. It is noteworthy that the buried mutants are influenced with hydrophobic interactions, whereas the exposed residues tend to form hydrogen bonding, electrostatic and van der Waals interactions (Gromiha *et al*., 1999b). The mutants in exposed regions are predicted with low accuracy and sensitivity. The inclusion of all these interactions may be necessary to correctly distinguish the accelerating and decelerating mutants in exposed regions.

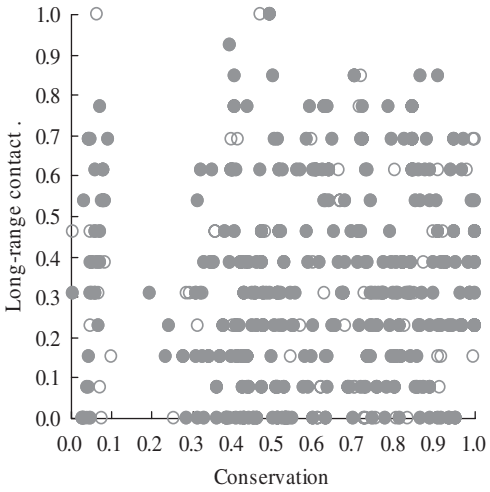

Fig. 3. Distribution of conservation score and long-rang contact in F467 dataset. Black and dark circles denote correctly and wrongly predicted mutants, respectively.

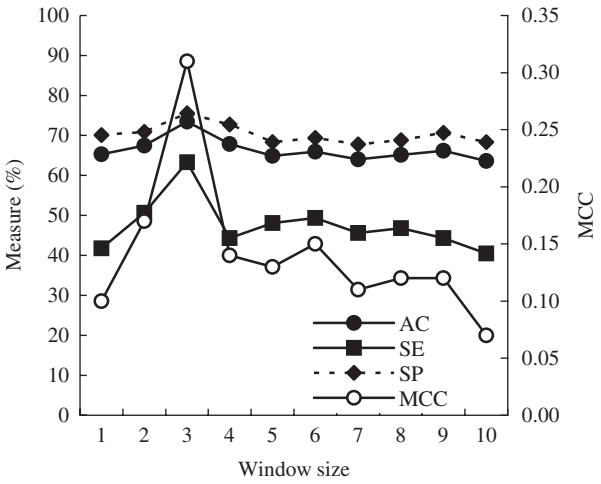

Fig. 4. Variation of discrimination performance with different number of neighboring residues.

3.4 Variation of conservation score and LRCs for discriminating the accelerating and decelerating mutants

We have analyzed the influence of LRCs and conservation score for discriminating the mutants, which enhance or slow down the folding process. The sensitivity, specificity, accuracy and MCC of discrimination are given in Supplementary Table S5. We noticed that the mutants with more number of LRCs and high conservation score are well predicted with the accuracy in the range 75–80%. In Figure 3, we plotted the relationship between conservation score and LRCs with correctly/wrongly predicted mutants. We observed that the mutants with high conservation (>0.6) and more number of LRCs (>0.6) are correctly predicted with our model. Among the 55 mutants, 42 of them are correctly predicted with the accuracy of 76.4%. Further, the sensitivity of mutants with normalized LRCs of >0.75 is 100%.

3.5 Influence of neighboring residues for discriminating the accelerating and decelerating mutants

We have analyzed the influence of neighboring residues for discriminating the accelerating and decelerating mutants. The results obtained for different window lengths are shown in Figure 4.

The inclusion of three neighboring residues for the features, M_w, E_l, N_s, ΔG and s showed the best performance in terms of sensitivity, specificity, accuracy and MCC. The performance did not improve with additional neighboring residues.

3.6 Examining the performance of the method

We have examined the performance of the prediction method with two typical examples, barnase (1BNR, 1BNR*) and acyl coenzyme protein (2ABD) with 19 and 30 mutants, respectively. These proteins belong to $\alpha+\beta$ class of proteins. The mutants have wide range of LRCs and the conservation scores lie between 0.3 and 1.0. We have utilized our model and predicted the change of folding rate upon mutation. We obtained the accuracy, sensitivity, specificity and MCC of 90%, 100%, 88% and 0.72, respectively for the protein 1BNR, which has 16 decelerating and 3 accelerating mutants. The folding rates of 2ABD mutants showed the presence of equal number of accelerating and decelerating mutants. Our method could distinguish the mutants with an accuracy of 73% (sensitivity, 0.73 and specificity, 0.73) and MCC of 0.47. The variation of conserved scores and LRCs for these proteins are shown in Supplementary Figure S5.

Further, we have analyzed the discrimination performance in different mutation pairs (e.g. A to G and G to A). We observed that five pairs of mutations have appreciable number of data (at least in one pair) in the F467 dataset. The discrimination accuracy obtained for all these five mutation pairs are presented in Supplementary Table S6. Our method could discriminate the accelerating and decelerating mutants with high accuracy in most of the considered pairs. The only exception is the mutation of P to A, which showed an accuracy of 50%. The mutation A to G with 26 data and V to A with 54 data discriminated the accelerating and decelerating mutants with an accuracy of 81% and 76%, respectively.

3.7 Discrimination on the web

We have developed a web server, FREEDOM (folding rate change prediction using regression models), for discriminating the accelerating and decelerating mutants and it is freely available at http://bioinformatics.myweb.hinet.net/freedom.htm. Figure 5a shows the necessary input parameters (wild type, mutant and three neighboring residues, normalized LRC and conservation score) for predicting the folding rate change upon mutation for the mutant D23A in 2CI2. FREEDOM utilizes QRMs and returns the predicted results within a minute. The output (Fig. 5b) contains the information about the input sequence and the predicted folding rate change (accelerating or decelerating). In this example, FREEDOM showed that the mutant accelerated the folding process, which is in agreement with the experimental observation. Further, we have included the characteristic features of wild type residue and three neighboring residues in terms of composition, polarity and metabolic role.

Fig. 5. (**a**) Snapshot showing the necessary input parameters for predicting the folding rate change upon mutation. (**b**) Prediction results obtained by FREEDOM and the characteristic features of wild type and three neighboring residues.

3.8 Merits, limitations and possible improvements

The present method is the first one for discriminating the accelerating and decelerating mutants. The compiled dataset from various resources can be used for further analysis and prediction of protein folding rates upon mutations. The present work is focused on various amino acid properties, conservation score, SS, SA and LRCs. The problem is similar to the prediction of protein stability upon mutations (Gromiha, 2007). Hence, for the proteins with known 3D structures, different interactions (Guerois *et al*., 2002) and potentials (Dehouck *et al*., 2009; Parthiban *et al*., 2006) can be derived and applied for predicting protein folding rates upon mutation. Further, the method can be refined with other sequence-based parameters (Capriotti *et al*., 2005; Huang *et al*., 2007b). These sequence and structure-based methods can be extended to predict the numerical values for the changes in protein folding rates upon mutations in addition to discrimination. The works on these directions are on progress.

4 CONCLUSIONS

We have compiled the first dataset for protein folding rates upon amino acid substitutions. The mutants are located in various ranges of SA, distributed in different SSs, accommodated in variable and conserved positions, and connected with variety of LRCs. We have utilized the information on conservation score and LRC along with 10 amino acid properties for discriminating the accelerating and decelerating mutants.

Our method showed an accuracy of 74% with the sensitivity and specificity of 63% and 76%, respectively. The method has been implemented on a web server and the discrimination results are available at http://bioinformatics.myweb.hinet.net/freedom.htm. This is the first method to discriminate the accelerating and decelerating mutants and it can be refined with the inclusion of various physical interactions and predicting the changes in folding rates quantitatively.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
Baldi,P. *et al.* (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.
Bogatyreva,N.S. *et al.* (2009) KineticDB: a database of protein folding kinetics. *Nucleic Acids Res.*, **37**, D342–D346.
Bordner,A.J. and Abagyan,R.A. (2004) Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins*, **57**, 400–413.
Boutet,E. *et al.* (2007) UniProtKB/Swiss-Prot. *Methods Mol. Biol.*, **406**, 89–112.
Bromberg,Y. *et al.* (2008) SNAP predicts effect of mutations on protein function. *Bioinformatics*, **24**, 2397–2398.
Capriotti,E. *et al.* (2005) Predicting protein stability changes from sequences using support vector machines. *Bioinformatics*, **21** (Suppl. 2), ii54–ii58.
Carlsson,J. *et al.* (2009) Investigation and prediction of the severity of p53 mutants using parameters from structural calculations. *FEBS J.*, **276**, 4142–4155.
Chang,C.-C. and Lin,C.-J. (2001) LIBSVM: a library for support vector machines. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (last accessed date June 4, 2010).
Chea,E. and Livesay,D.R. (2007) How accurate and statistically robust are catalytic site predictions based on closeness centrality? *BMC Bioinformatics*, **8**, 153.
Cheng,J. *et al.* (2006) Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, **62**, 1125–1132.
Cleary,J.G. and Trigg,L.E. (1995) K*: an instance-based learner using an entropic distance measure. *In Proceedings of the 12th International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 108–114.
Dehouck, Y. *et al.* (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, **25**, 2537–2543.
Fulton,K.F. *et al.* (2005) PFD: a database for the investigation of protein folding kinetics and stability. *Nucleic Acids Res.*, **33**, D279–D283.
Gao,S. *et al.* (2009) Prediction of function changes associated with single-point protein mutations using support vector machines (SVMs). *Hum. Mutat.*, **30**, 1161–1166.
Gromiha,M.M. *et al.* (1999a) ProTherm: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **27**, 286–288.
Gromiha,M.M. *et al.* (1999b) Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. *Protein Eng.*, **12**, 549–555.
Gromiha,M.M. *et al.* (1999c) Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys. Chem.*, **82**, 51–67.
Gromiha,M.M. *et al.* (2000) Importance of surrounding residues for protein stability of partially buried mutations. *J. Biomol. Struct. Dyn.*, **18**, 281–295.
Gromiha,M.M. (2003) Importance of native-state topology for determining the folding rate of two-state proteins. *J. Chem. Inf. Comput. Sci.*, **43**, 1481–1485.
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. (2007) Prediction of protein stability upon point mutations. *Biochem. Soc. Trans.*, **35**, 1569–1573.
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 Selvaraj,S. (1997) Influence of medium and long range interactions in different structural classes of globular proteins. *J. Biol. Phys.*, **23**, 151–162.
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.
Gromiha,M.M. and Selvaraj,S. (2004) Inter-residue interactions in protein folding and stability. *Prog. Biophys. Mol. Biol.*, **86**, 235–277.
Gromiha,M.M. and Selvaraj,S (2008) Bioinformatics approaches for understanding and predicting protein folding rates. *Curr. Bioinform.*, **3**, 1–9.
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. *et al.* (2009) TMFunction: database for functional residues in membrane proteins. *Nucleic Acids Res.*, **37**, D201–D204.

Guerois,R. *et al.* (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.

Holliday,G.L. *et al.* (2009) Understanding the functional roles of amino acid residues in enzyme catalysis. *J. Mol. Biol.*, **390**, 560–577.

Huang,J.T. and Tian,J. (2006) Amino acid sequence predicts folding rate for middle-size two-state proteins. *Proteins*, **63**, 551–554.

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. (2009) Reliable prediction of protein thermostability change upon double mutation from amino acid sequence. *Bioinformatics*, **25**, 2181–2187.

Huang,J.T. *et al.* (2007a) Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics. *Proteins*, **67**, 12–17.

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

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.

Jackson,S.E. (1998) How do small single-domain proteins fold? *Fold Des.*, **3**, R81–R91.

Jiang,Y. *et al.* (2009) Prediction of protein folding rates from primary sequences using hybrid sequence representation. *J. Comput. Chem.*, **30**, 772–783.

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

Katoh,K. *et al.* (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.

Kumar,M.D. and Gromiha,M.M. (2006) PINT: protein-protein interactions thermodynamic database. *Nucleic Acids Res.*, **34**, D195–D198.

Lopez,G. *et al.* (2007) FireDB—a database of functionally important residues from proteins of known structure. *Nucleic Acids Res.*, **35**, D219–D223.

Ma,B.G. *et al.* (2006) Direct correlation between proteins' folding rates and their amino acid compositions: an ab initio folding rate prediction. *Proteins*, **65**, 362–372.

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

Moody,J. and Darken,C. (1989) Fast learning in networks of locally-tuned processing units. *Neural Comput.*, **1**, 281–294.

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

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.

Porter,C.T. *et al.* (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.

Prabakaran,P. *et al.* (2001) Thermodynamic database for protein-nucleic acid interactions (ProNIT). *Bioinformatics*, **17**, 1027–1034.

Pugalenthi,G. *et al.* (2008) Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. *Biochem. Biophys. Res. Commun.*, **367**, 630–634.

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

Rumelhart,D.E. *et al.* (1986) Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1: Foundations*. MIT Press, Cambridge, MA, pp. 318–362.

Sankararaman,S. *et al.* (2010) Active site prediction using evolutionary and structural information. *Bioinformatics*, **26**, 617–624.

Sonego,P. *et al.* (2008) ROC analysis: applications to the classification of biological sequences and 3D structures. *Brief. Bioinform.*, **9**, 198–209.

Sonego,P. *et al.* (2007) A protein classification benchmark collection for machine learning. *Nucleic Acids Res.*, **35**, D232–D236.

Tomii,K. and Kanehisa,M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.*, **9**, 27–36.

Valdar,W.S. (2002) Scoring residue conservation. *Proteins*, **48**, 227–241.

Yin,S. *et al.* (2007) Eris: an automated estimator of protein stability. *Nat. Methods*, **4**, 466–467.

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