Biophysical Chemistry 125 (2007) 462–470

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

Prediction of protein mutant stability using classification and regression tool

Liang-Tsung Huang [a], K. Saraboji [b], Shinn-Ying Ho [c,d], Shiow-Fen Hwang [a], M.N. Ponnuswamy [b], M. Michael Gromiha [e,*]

[a] *Institute of Information Engineering and Computer Science, Feng-Chia University, Taichung, 407, Taiwan*
[b] *Department of Crystallography and Biophysics, University of Madras, Guindy Campus, Chennai — 600 025, India*
[c] *Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan*
[d] *Institute of Bioinformatics, National Chiao Tung University, Hsinchu 300, Taiwan*
[e] *Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio IT Research Building, 2 42 Aomi, Koto ku, Tokyo 135 0064, Japan*

Received 1 September 2006; received in revised form 19 October 2006; accepted 23 October 2006
Available online 20 November 2006

Abstract

Prediction of protein stability upon amino acid substitutions is an important problem in molecular biology and the solving of which would help for designing stable mutants. In this work, we have analyzed the stability of protein mutants using two different datasets of 1396 and 2204 mutants obtained from ProTherm database, respectively for free energy change due to thermal ($\Delta\Delta G$) and denaturant denaturations ($\Delta\Delta G^{H_2O}$). We have used a set of 48 physical, chemical energetic and conformational properties of amino acid residues and computed the difference of amino acid properties for each mutant in both sets of data. These differences in amino acid properties have been related to protein stability ($\Delta\Delta G$ and $\Delta\Delta G^{H_2O}$) and are used to train with classification and regression tool for predicting the stability of protein mutants. Further, we have tested the method with 4 fold, 5 fold and 10 fold cross validation procedures. We found that the physical properties, shape and flexibility are important determinants of protein stability. The classification of mutants based on secondary structure (helix, strand, turn and coil) and solvent accessibility (buried, partially buried, partially exposed and exposed) distinguished the stabilizing/destabilizing mutants at an average accuracy of 81% and 80%, respectively for $\Delta\Delta G$ and $\Delta\Delta G^{H_2O}$. The correlation between the experimental and predicted stability change is 0.61 for $\Delta\Delta G$ and 0.44 for $\Delta\Delta G^{H_2O}$. Further, the free energy change due to the replacement of amino acid residue has been predicted within an average error of 1.08 kcal/mol and 1.37 kcal/mol for thermal and chemical denaturation, respectively. The relative importance of secondary structure and solvent accessibility, and the influence of the dataset on prediction of protein mutant stability have been discussed.

Keywords: Thermal stability; Free energy change; Amino acid substitution; Protein mutants; Amino acid property; Classification and regression tool, CART

1. Introduction

Prediction of protein stability upon amino acid substitutions is one of the challenging tasks in molecular biology and a reliable prediction method would also help in designing stable protein mutants. Several methods have been proposed for understanding the factors influencing the stability of protein mutants and for predicting protein stability changes upon mutation. It has been reported that hydrophobicity, hydrogen bonds, ion pairs and other non covalent interactions are important for the stability of proteins upon amino acid substitutions and the stability depends on the location of mutants with respect to secondary structure and solvent accessibility [1–9].

Gromiha et al. [10] collected the experimental stability data for more than 18,000 protein mutants and developed a database, ProTherm, which is available freely for academic users. Based on this database several methods have been developed for predicting protein stability upon mutation, such as those based on energetic criterion, stability scale for the 20 amino acid residues, contact potentials, neural networks, support vector machines, average assignment method etc. [11–17]. Guerois et al. [17] developed a computer algorithm, FOLDEF for estimating the stability of proteins upon amino acid substitutions.

[*] Corresponding author. Tel.: +81 3 3599 8046; fax: +81 3 3599 8081.
E-mail address: michael-gromiha@aist.go.jp (M.M. Gromiha).

0301-4622/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2006.10.009

Zhou and Zhou [16] derived a stability scale of 20 amino acid residues from a database of 1023 mutation experiments and utilized the same for predicting protein mutant stability. Bordner and Abagyan [15] developed an empirical energy function, which includes terms representing the energy contributions of the folded and denatured proteins and used the energy function to predict protein mutant stability. Khatun et al. [13] tested the ability of contact potentials to accurately and transferably predict stability changes of proteins upon mutations. Capriotti et al. [12] developed a method based on support vector machines for protein mutant stability prediction. Saraboji et al. [11] proposed an average assignment method for predicting the stability of protein mutants.

Recently, the secondary structure and solvent accessibility parameters have been used to classify the mutants and it was observed that this classification improved the accuracy of prediction. Gilis and Rooman [18,19] separated the mutants based on solvent accessibility and used torsion and distance potentials with different weighting factors for predicting the stability of proteins upon buried and exposed mutations. Gromiha et al. [8,20,21] analyzed the relationship between amino acid properties and protein stability upon buried, partially buried and exposed mutants. They showed that hydrophobicity plays a major role to the stability of buried mutations whereas hydrogen bonds, other polar interactions and hydrophobic interactions are important to the stability of partially buried mutations. Further, hydration entropy and strain energy are influencing the stability of exposed mutations. Capriotti et al. [14] have developed a neural network method for predicting the stability of protein mutants in which the information about solvent accessibility was added as one of the input neurons. Further, the analysis of T4 and human lysozyme mutants showed that the classification of secondary structure and solvent accessibility are important to understand their stability [22]. These studies demonstrate the importance of classifying the mutant stability data based on secondary structure and/or solvent accessibility for understanding/predicting the stability of protein mutants.

The stability of protein structures is dictated by several non covalent interactions and the amino acid properties carry sufficient information for understanding protein stability. In our earlier works, we have explored the importance of amino acid properties for understanding protein folding rates, stability and transition state structures [8,23,24]. In this work, we have used a set of 48 physical, chemical energetic and conformational properties of amino acid residues and analyzed the relationship between amino acid properties and protein stability using two large sets of data. We observed that the physical properties, shape and flexibility are the major determinants for protein stability. Further, we have developed a method using classification and regression tool for predicting the stability of protein mutants at different secondary structures and solvent accessibility. We observed that the present method could predict the stability of protein mutants at an accuracy of 81% and 80% for $\Delta\Delta G$ and $\Delta\Delta G^{H_2O}$. The relative importance of solvent accessibility and secondary structure, and the effect of data size have been discussed.

2. Materials and methods

2.1. Datasets

We have used two sets of experimental stability data, (i) free energy change of 1396 mutants obtained from thermal denaturation ($\Delta\Delta G$) and (ii) free energy change of 2204 mutants obtained with chemical denaturation ($\Delta\Delta G^{H_2O}$) for the present study. All these datasets have been obtained from ProTherm database (http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html [10,25]) with the following conditions: (i) all single mutants, (ii) known three dimensional structures, (iii) any secondary structure and solvent accessibility and (iv) data obtained from thermal denaturation for $\Delta\Delta G$ and chemical denaturation (urea and GdnHCl) for $\Delta\Delta G^{H_2O}$. The secondary structure information was obtained from the DSSP, Dictionary of Secondary Structures in Proteins [26]. We have used the program ASC [27] for calculating the solvent accessibility of each residue and normalized with its respective extended state accessibility as explained in our earlier article [8]. For a protein of unknown structure, available online methods such as PHD [28] and RVP net [29] are used for predicting the secondary structure and solvent accessibility, respectively. The majority of the mutants in this dataset are located in helical segments (42%) followed by strand (24%) and coil (23%) regions. With respect to solvent accessibility the distribution of mutants at different ranges, such as, buried (0–2%), partially buried (2–20%), partially exposed (20–50%) and exposed (>50%) are, respectively, 23%, 23%, 26% and 28%. We have not considered the experimental conditions, such as pH, temperature, buffers, ions, additives etc. as well as the occurrence of mutants with excess heat capacity as these information are unknown for a new mutant. All datasets used in the present study are available from the corresponding author.

2.2. Amino acid properties

In the present study, we used a set of 48 diverse amino acid properties (physical-chemical, energetic and conformational), which fall into various clusters analyzed by Tomii and Kanehisa [30]. This set of properties has been used in our previous works for understanding protein stability, transition state structures of proteins, predicting protein folding and unfolding rates etc. [8,24,31–33]. The list of 48 properties used in the present study and their brief descriptions are presented in Table 1.

2.3. Computational procedure

The mutation induced changes in property values $\Delta P(i)$ was computed using the equation:

$$\Delta P(i) = P_{\mathrm{mut}}(i) - P_{\mathrm{wild}}(i), \tag{1}$$

where, $P_{\mathrm{mut}}(i)$ and $P_{\mathrm{wild}}(i)$ are, respectively, the property value of the ith mutant and wild type residues, and i varies from 1 to N, total number of mutants. The computed difference in property values $\Delta P(i)$ has been used to predict the stability of protein mutants as stabilizing/destabilizing based on the increase/decrease in both stability and property values.

2.4. Classification and regression tool

We have used classification and regression tool (CART), which is an implementation based on classification and regression tree algorithm [34], for predicting the stability of protein mutants with 48 amino acid properties. It provides numerical values in the output, which can be compared with experimental stability data.

CART constructs a binary decision tree based on training dataset, starting at the tree root. The dataset will be progressively split into smaller subsets which satisfy a given condition. The splitting procedure is made in accordance with squared residuals minimization criterion which implies that expected sum variances for two resulting nodes should be minimized:

$$\underset{x_i \le x_i^R, i=1,\dots,M}{\operatorname{argmin}} \left[P_l \mathrm{Var}(Y_l) + P_r \mathrm{Var}(Y_r)\right] \tag{2}$$

where P_l, P_r are fractions of samples in the left and right nodes; Var(Y_l), Var(Y_r) are variances of response vectors for corresponding left and right child nodes; $x_i \le x_i^R$ is the optimal splitting condition which satisfies the criterion (Eq. (1)) with x_j^R the best splitting value of variable x_j from M variables in learning samples. The splitting procedure goes on until: (1) only one observation or more than two observations with the identical values exist in each of the child nodes, or (2) the number of levels exceeds the limit set by system. Then the procedure of building the maximal tree is terminated.

CART is a nonparametric type of regression fitting approach, which is suitable for unknown distributions of data. Another advantage is that CART deals effectively with large datasets and the issues of higher dimensionality. We have applied CART to predict the stability of protein mutants and analyzed the results obtained for different classifications based on secondary structure and solvent accessibility.

Table 1
Accuracy of predicting the stability of protein mutants using 48 various amino acid properties

Property no.	Property name	Accuracy (%)	
		$\Delta\Delta G$	$\Delta\Delta G^{\mathrm{H_2O}}$
47	s	73.14	78.36
48	f	70.92	77.77
46	v	67.48	71.64
19	C_a	65.62	70.15
38	$-T\Delta S_h$	64.04	66.02
8	B_1	63.61	62.80
13	E_l	63.32	56.26
27	V^0	63.25	67.97
39	ΔC_{ph}	63.18	61.03
33	ΔASA	63.11	64.79
2	H_t	62.11	61.98
29	N_1	61.68	53.95
9	R_f	61.53	55.85
3	H_p	61.25	53.22
16	P_β	60.89	57.53
23	N_s	60.32	51.09
31	$\mathrm{ASA_D}$	59.89	67.56
21	B_r	59.89	53.04
30	H_{gm}	59.89	51.91
7	M_w	59.46	67.24
22	R_a	59.24	53.58
36	G_{hN}	58.88	51.95
10	μ	58.17	61.07
34	ΔG_h	57.95	52.36
41	ΔH_c	57.74	62.16
35	G_{hD}	57.59	54.45
11	H_{nc}	57.59	52.77
14	E_t	57.45	48.82
4	P	56.09	65.02
5	pHi	54.08	51.32
45	$-T\Delta S$	54.01	55.58
24	α_n	52.44	49.27
6	pK′	49.07	46.78
28	N_m	48.57	50.54
15	P_α	48.50	49.77
43	ΔG	47.28	43.83
44	ΔH	46.99	47.73
1	K^0	46.56	39.47
32	ASA_N	46.49	57.12
37	ΔH_h	46.35	40.88
12	E_{sm}	46.35	37.75
18	P_c	45.99	49.27
20	F	44.20	48.00
26	α_m	43.91	46.92
25	α_c	43.63	46.14
17	P_t	43.20	45.33
40	ΔG_c	42.91	48.68
42	$-T\Delta S_c$	42.26	40.29

Notes to Table:

K^0, compressibility; H_t, thermodynamic transfer hydrophobicity; H_p, surrounding hydrophobicity; P, polarity; pHi, isoelectric point; pK′, equilibrium constant with reference to the ionization property of COOH group; M_w, molecular weight; B_1, bulkiness; R_f, chromatographic index; μ, refractive index; H_{nc}, normalized consensus hydrophobicity; E_{sm}, short- and medium-range nonbonded energy; E_1 long-range nonbonded energy; E_t, total nonbonded energy (E_{sm}+ E_1); P_α, P_β, P_t, and Pc are, respectively, α-helical, β-structure, turn, and coil tendencies; C_a, helical contact area; F, mean r.m.s. fluctuational displacement; B_r, buriedness; R_a, solvent-accessible reduction ratio; N_s, average number of surrounding residues; α_n, α_c, and α_m are, respectively, power to be at the N-terminal, C-terminal, and middle of α-helix; $V°$, partial specific volume; N_m and N_1 are, respectively, average medium-and long-range contacts; H_{gm}, combined surrounding hydrophobicity (globular and membrane); $\mathrm{ASA_D}$, $\mathrm{ASA_N}$, and ΔASA are, respectively, solvent accessible surface area for denatured, native, and unfolding; ΔG_h, G_{hD}, and G_{hN} are, respectively, Gibbs free energy change of hydration for unfolding, denatured, and native protein; ΔH_h, unfolding enthalpy change of hydration; $-T\Delta S_h$, unfolding entropy change of hydration; ΔC_{ph}, unfolding hydration heat capacity change; ΔG_c, ΔH_c, and $-T\Delta S_c$ are, respectively, unfolding Gibbs free energy, unfolding enthalpy, and unfolding entropy changes of chain; ΔG, ΔH, and $-T\Delta S$ are respectively, unfolding Gibbs free energy change, unfolding enthalpy change, and unfolding entropy change; V, volume (number of nonhydrogen side-chain atoms); s, shape (position of branch point in a side chain); f, flexibility (number of side-chain dihedral angles).

K^0 in m^3/mol/Pa (×10^{-15}); H_t, H_p, H_{nc}, H_{gm}, ΔG_h, G_{hD}, G_{hN}, ΔH_h, $-T\Delta S_h$, ΔG_c, ΔH_c, $-T\Delta S_c$, ΔG, ΔH, and $-T\Delta S$ in kcal/mol; P in Debye; pHi and pK′ in pH units; E_{sm}, E_1, and E_t, in kcal/mol/atom; B_1, C_α, $\mathrm{ASA_D}$, $\mathrm{ASA_N}$, and ΔASA in Å^2; F in Å; $V°$ in m^3/mol (×10^{-6}); ΔC_{ph} in cal/mol/K; and the rest are dimensionless quantities.

Table 2
Prediction results based on the classification with ASA and secondary structure by self-consistency test

	ASA range	Helix				Strand				Turn				Coil			
		number of data	r	Accuracy (%)	MAE	number of data	r	Accuracy (%)	MAE	number of data	r	Accuracy (%)	MAE	number of data	r	Accuracy (%)	MAE
$\Delta\Delta G$	0–2	199	0.8932	90.95	0.6102	73	0.8779	87.67	0.4415	29	0.9601	100.00	0.2543	23	0.9247	91.30	0.6243
	2–20	137	0.9476	87.59	0.4112	63	0.6942	90.48	1.2419	23	0.9767	95.65	0.2030	34	0.9830	100.00	0.1197
	20–50	149	0.8031	85.91	0.6658	102	0.9569	95.10	0.4267	22	0.8454	81.82	0.3235	96	0.8684	89.58	0.3457
	>50	176	0.8763	89.20	0.2584	40	0.9551	95.00	0.1847	75	0.9446	92.00	0.1626	155	0.9147	86.45	0.2612
Weighted average			0.8797	88.65	0.4878		0.8764	92.09	0.5805		0.9379	91.19	0.2549		0.9086	89.28	0.2990
$\Delta\Delta G^{H_2O}$	0–2	162	0.8142	90.12	0.9189	244	0.8408	86.89	1.0256	12	0.8517	100.00	0.7083	56	0.8148	92.86	0.9417
	2–20	170	0.7621	85.88	0.9997	248	0.8393	92.74	0.8522	61	0.8766	85.25	0.5175	155	0.8080	92.90	0.7445
	20–50	183	0.7152	89.07	0.6657	139	0.8830	90.65	0.4596	74	0.9362	91.89	0.3777	196	0.6997	84.69	0.8067
	>50	201	0.7832	86.57	0.4094	56	0.9201	94.64	0.1772	102	0.8599	86.27	0.4334	138	0.7787	86.23	0.3817
Weighted average			0.7678	87.85	0.7303		0.8553	90.39	0.7793		0.8863	91.92	0.5501		0.7623	88.25	0.6953

ASA: accessible surface area (solvent accessibility).
MAE: mean absolute error.

2.5. Accuracy of distinguishing stability of protein mutants

The accuracy of distinguishing the stability of mutants (stabilizing/destabilizing) has been determined by using the following expression:

$$\text{Accuracy}(\%) = p{*}100.0/N \tag{3}$$

where, p is the total number of correctly discriminated residues (both the difference in property and stability, increase/decrease upon mutations) and N is the total number of data used for discrimination.

2.6. Single correlation

The correlation between the experimental and assigned stability (ΔT_m) has been calculated using the familiar expression:

$$r = \frac{N\sum XY - (\sum X \sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}} \tag{4}$$

where r is the correlation coefficient, N, X, and Y are the number of data, experimental and assigned stability, respectively.

2.7. Mean absolute error

The mean absolute error (MAE) is defined as the absolute difference between predicted and experimental stability values:

$$\text{MAE} = \frac{1}{N}\sum_i |X_i - Y_i| \tag{5}$$

where, X_i and Y_i are the experimental and predicted stability values, respectively and i varies from 1 to N, N being the total number of mutants.

2.8. Self-consistency and n-fold cross-validation tests

The present method was validated by both self-consistency and n-fold cross-validation tests. Self-consistency included all the stability data for training the CART model and prediction was made for all the mutants. n-fold cross-validation partitions samples into n sub-samples chosen randomly with approximately equal size. For each sub-sample, the method fits a tree to the remaining data and uses it to predict the stability of the sub-sample. The procedure has been repeated for n times to obtain the accuracy, correlation and MAE.

Table 3
Prediction results based on the classification of ASA and secondary structure by 5-fold cross-validation test

	ASA range	Helix				Strand				Turn				Coil			
		Number of data	r	Accuracy (%)	MAE	Number of data	r	Accuracy (%)	MAE	Number of data	r	Accuracy (%)	MAE	Number of data	r	Accuracy (%)	MAE
$\Delta\Delta G$	0–2	199	0.7825	87.18	1.0086	73	0.4450	90.00	1.3213	29	0.8707	100.00	0.6538	23	0.9360	85.00	1.5218
	2–20	137	0.8488	80.00	0.8395	63	0.3872	83.33	2.5068	23	0.6173	90.00	1.6597	34	0.5672	93.33	1.2550
	20–50	149	0.4921	72.41	1.4411	102	0.7409	79.00	1.6189	22	0.2022	75.00	0.9461	96	0.6549	90.53	0.7928
	>50	176	0.4235	78.29	0.6598	40	0.7060	85.00	0.7956	75	0.3998	68.00	0.8759	155	0.5964	74.19	0.7192
Weighted average			0.6352	80.00	0.9782		0.5780	83.73	1.6235		0.4958	83.96	1.0210		0.6368	82.20	0.8612
$\Delta\Delta G^{H_2O}$	0–2	162	0.4488	83.75	1.8572	244	0.6744	79.58	1.6809	12	0.8182	100.00	1.7375	56	0.5149	87.27	2.0714
	2–20	170	0.4694	74.12	1.8281	248	0.6145	88.57	1.5737	61	0.2117	73.33	1.4435	155	0.5463	89.68	1.2825
	20–50	183	0.3147	77.78	1.2528	139	0.4513	82.96	1.0989	74	0.4156	80.00	1.3255	196	0.1956	78.97	1.4037
	>50	201	0.4052	75.00	0.7970	56	0.6780	85.45	0.5624	102	0.3248	70.00	0.9658	138	0.1402	71.11	0.9071
Weighted average			0.4072	77.48	1.3982		0.6079	83.99	1.4333		0.3479	83.88	1.4851		0.3141	80.88	1.3121

ASA: accessible surface area (solvent accessibility).
MAE: mean absolute error.

3. Results and discussion

3.1. Relationship between amino acid properties and protein mutant stability

We have computed the accuracy of predictions for the stability of protein mutants using the difference in amino acid properties upon mutation and the stability change due to amino acid substitution. The results obtained for the two sets of data ($\Delta\Delta G$ and $\Delta\Delta G^{H_2O}$) are shown in Table 1. We found that the properties shape (s) and flexibility (f) predicted the stability of protein mutants at an accuracy of more than 70% for $\Delta\Delta G$ whereas helical contact area (C_α), volume (v), shape (s) and flexibility (f) predicted the stability of protein mutants at high accuracy for $\Delta\Delta G^{H_2O}$. Interestingly, s predicted the stability of protein mutants with the highest accuracy of more than 72% in these two sets of data. This result indicates that shape (position of branch point in a side chain) is one of the major determinants to protein mutant stability, which is consistent with other studies that shape plays an important role to the stability of thermophilic proteins [35,36] as well as for explaining the stability of protein mutants [37].

Table 4
Average accuracy, correlation and MAE of predicting protein mutant stability with self-consistency, 4-fold and 5-fold cross-validation methods

	$\Delta\Delta G$			$\Delta\Delta G^{H_2O}$		
	r	Accuracy (%)	MAE	r	Accuracy (%)	MAE
4-fold	0.5884	80.59	1.1010	0.4159	80.22	1.3816
5-fold	0.6093	81.08	1.0794	0.4401	80.10	1.3684

MAE: mean absolute error.

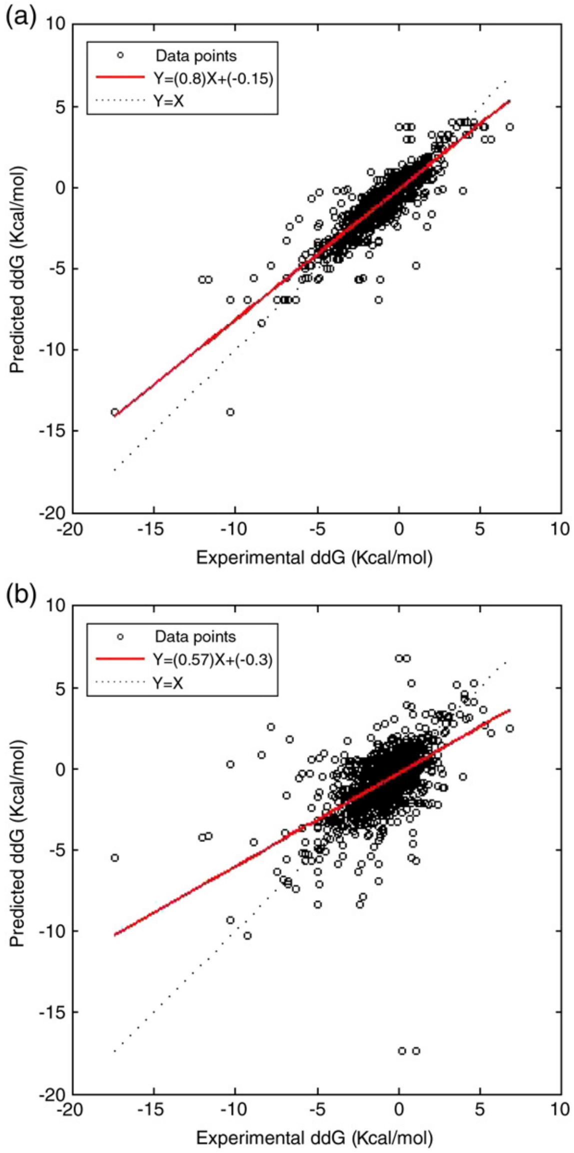

Fig. 1. a: Relationship between experimental and predicted $\Delta\Delta G$ (self-consistency test) in a set of 1396 mutants ($r=0.90$). b: Relationship between experimental and predicted $\Delta\Delta G$ (5-fold cross-validation method) in a set of 1396 mutants ($r=0.59$).

3.2. Prediction of protein stability upon amino acid substitutions

The information about the difference of amino acid properties upon mutation have been used to predict the stability of protein mutants through CART. We have used CART for predicting the stability of protein mutants based on their secondary structure and solvent accessibility and the method has been optimized with 4-fold, 5-fold and 10-fold cross-validation procedures. The results obtained with self-consistency test are shown in Table 2. From this table, we observed that the data has been trained with the weighted average accuracy in the range of 88–92% for the two measures of stability. The weighted average correlation lies between 0.76 and 0.94. The weighted average mean absolute error between the experimental and computed free energy change is 0.25–0.78 kcal/mol, which shows the average deviation of about 40%.

The results obtained with 5-fold cross-validation test are presented in Table 3. We found that for the data $\Delta\Delta G$, the weighted accuracy for predicting the stability of protein mutants is 81%; the buried strand and turn mutants can be predicted with 90–100% accuracy whereas the exposed turn mutants are predicted with the accuracy of 68%. We obtained similar trend for $\Delta\Delta G^{H_2O}$. This might be due to the fact that buried mutants are dominated by hydrophobic interactions whereas other interactions including hydrogen bonds, electrostatic and van der Waals interactions along with hydrophobic interactions are important for the stability of exposed mutations [8]. Further, the high accuracy of buried mutations is attributed with the constraints in the interior of the protein. The correlation in 16 sets of data lies in the range of 0.20–0.94 and the average correlation is 0.61. The MAE between experimental and predicted $\Delta\Delta G$ is 1.08 kcal/mol. Further, we noticed that the results obtained with $\Delta\Delta G^{H_2O}$ data are moderately worse than that obtained with $\Delta\Delta G$.

The comparison between experimental and predicted $\Delta\Delta G$ for a set of 1396 mutants is shown in Fig. 1(a and b) and we observed a good relationship between them. The main cause of the outliers might be due to the inclusion of same mutants obtained with different experimental conditions, which are not considered in the present work.

Table 5
Prediction results for the classification based on secondary structure or ASA by self-consistency test

Group	$\Delta\Delta G$				$\Delta\Delta G^{H_2O}$			
	Number of data	r	Accuracy (%)	MAE	Number of data	r	Accuracy (%)	MAE
Helix	661	0.7939	81.09	0.7849	716	0.6685	84.22	1.0139
Strand	278	0.7967	89.57	1.0459	687	0.7607	86.32	1.0339
Turn	149	0.9101	83.89	0.4275	249	0.8308	83.94	0.5897
Coil	308	0.8227	87.99	0.5436	545	0.6530	82.39	0.9156
Weighted average		0.8132	84.60	0.7455		0.7119	84.39	0.9477
ASA 0–2	324	0.8074	88.27	0.8600	480	0.8048	86.88	1.1778
ASA 2–20	257	0.8212	89.11	0.7342	635	0.7683	86.61	0.9240
ASA 20–50	369	0.7907	84.01	0.8158	592	0.7614	88.18	0.6682
ASA>50	446	0.7503	79.15	0.4541	497	0.8144	84.51	0.4294
Weighted average		0.7873	84.38	0.6955		0.7848	86.62	0.7990

ASA: accessible surface area (solvent accessibility).
MAE: mean absolute error.

3.3. Effect of the data obtained with excess heat capacity

The influence of excess heat capacity on protein stability has been analyzed by carrying out the computations using the data associated with ΔC_p. We obtained a set of 510 mutants with excess heat capacity and used the same dataset for predicting $\Delta\Delta G$. We observed that the stability of protein mutants in regular secondary structures (helix and strand) are better predicted for the data obtained with excess heat capacity. This result reveals that the mutations in helical and strand segments are attributed with amino acid properties and our method could reliably predict the stability. On the other hand mutations in turn and coil regions showed moderately less correlation compared with that obtained with the whole dataset of 1396 mutants.

3.4. Influence of n-fold cross-validation data

The average accuracy, correlation and MAE obtained with 4-fold and 5-fold cross-validation procedures are presented in Table 4. We observed that the difference in accuracy between 4-fold and 5-fold cross-validation methods is marginal. Similar trend is also observed for correlation and MAE.

3.5. Relative importance of secondary structure and solvent accessibility to predict protein mutant stability

We have analyzed the relative influence of secondary structure and solvent accessibility for predicting the stability of protein mutants and the results obtained with self-consistency and 5-fold cross-validation methods are presented in Tables 5 and 6. We observed that the accuracy of correctly assigning the mutants as stabilizing or destabilizing for the mutants in helical, strand, turn and coil segments are, respectively 81%, 90%, 84 and 88% for $\Delta\Delta G$. The 5-fold cross-validation test showed the accuracy of 71%, 80%, 68% and 79% for the mutants in these regions. The correlation lies in the range of 0.44 to 0.64.

On the other hand, the classification based on solvent accessibility showed the accuracy of 82%, 82%, 70% and 68% for the buried, partially buried, partially exposed and exposed mutations, respectively. The correlation lies in the range of 0.25 to 0.60. We observed a similar trend in $\Delta\Delta G^{H_2O}$. However, the average weighted correlation is weaker than that obtained with $\Delta\Delta G$.

From these results, we noticed that the performance of the prediction method with the classification of secondary structure

Table 6
Prediction results based on the classification of ASA or secondary structure by 5-fold cross-validation test

Group	$\Delta\Delta G$				$\Delta\Delta G^{H_2O}$			
	Number of data	r	Accuracy (%)	MAE	Number of data	r	Accuracy (%)	MAE
Helix	661	0.6322	71.21	1.0380	716	0.3494	77.34	1.3508
Strand	278	0.4795	79.64	1.6162	687	0.5481	79.71	1.3903
Turn	149	0.4397	67.59	1.0265	249	0.4564	74.29	0.9740
Coil	308	0.6358	78.69	0.8050	545	0.3230	76.33	1.2675
Weighted average		0.5820	74.15	1.1005		0.4171	77.48	1.2998
ASA 0–2	324	0.6036	81.88	1.1656	480	0.5538	81.25	1.6661
ASA 2–20	257	0.6040	81.57	1.1865	635	0.3951	77.01	1.5440
ASA 20–50	369	0.2501	70.41	1.4026	592	0.2207	78.14	1.1912
ASA >50	446	0.4209	67.87	0.7051	497	0.1722	67.47	0.8846
Weighted average		0.4519	74.32	1.0850		0.3326	76.09	1.3271

ASA: accessible surface area (solvent accessibility).
MAE: mean absolute error.

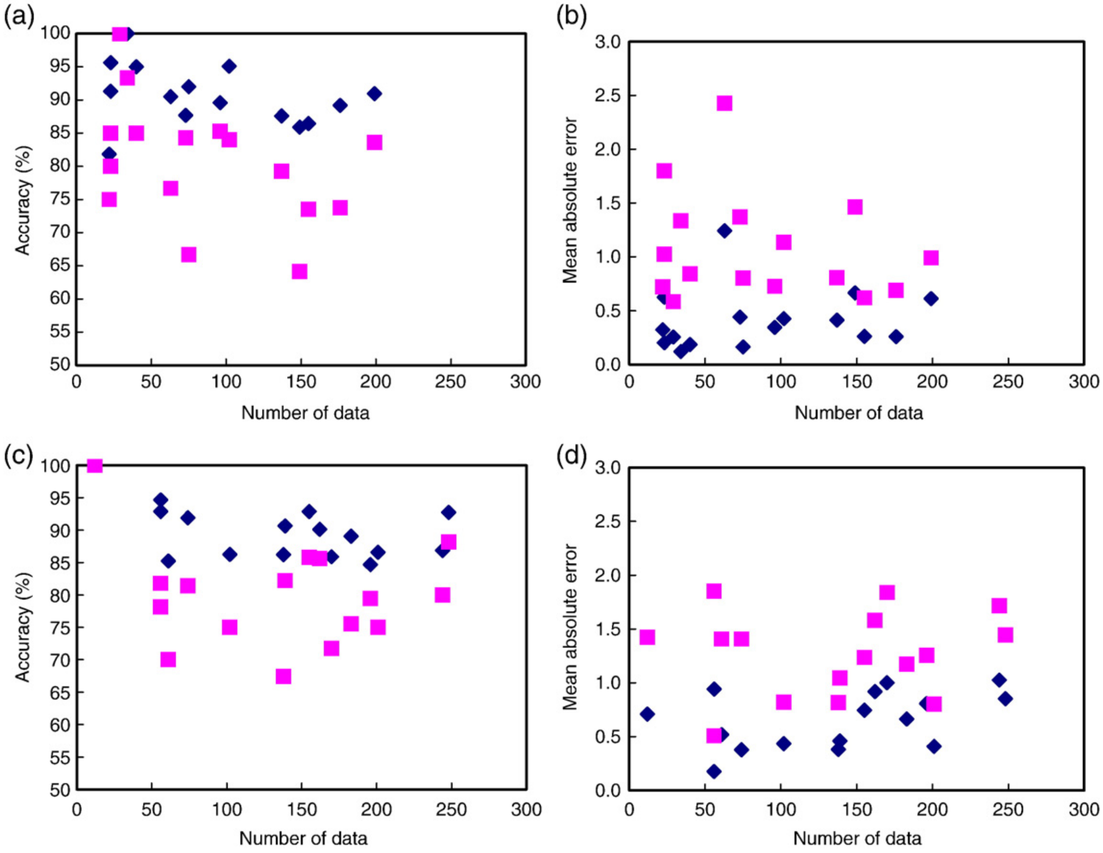

Fig. 2. Relationship between the number of data and prediction performance: (a) accuracy and (b) mean absolute error for $\Delta\Delta G$ and (c) and (d) for $\Delta\Delta G^{H_2O}$. The diamonds and squares represent the results obtained with 5-fold cross-validation and self-consistency tests, respectively.

is better than or similar to that with solvent accessibility. This observation is similar to our previous findings on human and T4 lysozymes that secondary structure carries more or similar information than solvent accessibility for assigning the stability of protein mutants [22].

3.6. Number of data versus accuracy

We have examined the bias on the present results whether the prediction accuracy depends on the population of data in each class. This has been revealed from the relationship between the number of data for each class and their average accuracy. Fig. 2 shows a relationship between number of data and accuracy (and MAE) for the two measures of stability, $\Delta\Delta G$ and $\Delta\Delta G^{H_2O}$ obtained with self consistency and 5-fold cross-validation method. We observed that the class with 137 $\Delta\Delta G$ data (helix, 2–20% ASA; see Table 3) predicted the stability of protein mutants (accuracy: 88% and 80%, correlation: 0.95 and 0.85 and MAE: 0.41 and 0.84 kcal/mol) better than that with 75 data at exposed turn mutants (accuracy: 86 and 68%, correlation: 0.94 and 0.40 and MAE: 0.16 and 0.88 kcal/mol). Similar results are also obtained with $\Delta\Delta G^{H_2O}$. We observed a poor correlation between number of data and accuracy or MAE and it is in the range of −0.4 to 0.1. These results showed that the accuracies obtained in this work are not biased with the number of data in each class.

3.7. Comparison with other methods

We have compared the predictive ability of our method with other methods in the literature. Although direct comparison is not appropriate due to the difference of datasets used in for training and test as well as the information used to develop the model it would provide the information about performance of different methods. Gilis and Rooman [18,19] derived distance and torsion potentials using 10 proteins and reported the correlation of 0.80 and 0.67 for 121 buried and 106 surface mutations between the predicted and experimental $\Delta\Delta G$. Khatun et al. [13] developed contact potentials and showed a correlation of 0.66 and 0.46, respectively for training and validation tests in a dataset of 1356 mutations, Zhou and Zhou [16] used a finite ideal gas reference state for the statistical potential and reported a correlation of 0.55 for 1023 mutants in 35 proteins. Here, the mutations that have decreased number of atoms were only used to avoid strains associated small-to-large mutations. Capriotti et al. [12] developed support vector machine based method, which predicts protein stability with 80% accuracy. The correlation and MAE are respectively, 0.71 and 1.3 kcal/mol. However, multiple occurrences of same mutations were observed in the dataset. The present method could predict the stability of protein mutants with an average accuracy of 90% using self consistency and 78% with 5-fold cross validation test. The average correlation and MAE are 0.90

and 0.49 kcal/mol, respectively for self consistency, and 0.59 and 1.01 kcal/mol for 5-fold cross validation test. This analysis shows that the performance of our method is similar to or better than other methods in the literature.

4. Conclusions

We have analyzed the stability of protein mutants using two large databases and 48 various amino acid properties. We found that the properties shape and flexibility are the major determinants to protein stability. Further, we proposed a method based on classification and regression tool for predicting the stability of proteins upon amino acid substitutions. We observed that the classification based on secondary structure and solvent accessibility significantly improved the correlation and the accuracy of assigning the stability of protein mutants. This classification showed an average accuracy of 81–90% for correctly assigning the protein mutants in two different datasets of $\Delta\Delta G$ and $\Delta\Delta G^{H_2O}$. The correlation is significantly high and ranges from 0.42–0.90. We suggest that the present method could be used as an effective tool for predicting the stability of protein mutants.

Acknowledgements

We thank the reviewers for constructive comments on our manuscript. K.S. acknowledges the Council of Scientific and Industrial Research (CSIR), Govt. of India, for the award of Senior Research Fellowship.

References

[1] S.E. Jackson, M. Moracci, N. elMasry, C.M. Johnson, A.R. Fersht, Effect of cavity-creating mutations in the hydrophobic core of chymotrypsin inhibitor 2, Biochemistry 32 (1993) 11259–11269.

[2] D. Shortle, W.E. Stites, A.K. Meeker, Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease, Biochemistry 29 (1990) 8033–8041.

[3] K. Yutani, K. Ogasahara, T. Tsujita, Y. Sugino, Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit, Proc. Natl. Acad. Sci. U. S. A. 84 (1987) 4441–4444.

[4] S.R. Trevino, K. Gokulan, S. Newsom, R.L. Thurlkill, K.L. Shaw, V.A. Mitkevich, A.A. Makarov, J.C. Sacchettini, J.M. Scholtz, C.N. Pace, Asp79 makes a large, unfavorable contribution to the stability of RNase Sa, J. Mol. Biol. 354 (2005) 967–978.

[5] C.F. Lee, G.I. Makhatadze, K.B. Wong, Effects of charge-to-alanine substitutions on the stability of ribosomal protein L30e from Thermococcus celer, Biochemistry 44 (2005) 16817–16825.

[6] J.M. Schwehm, C.A. Fitch, B.N. Dang, E.B. Garcia-Moreno, W.E. Stites, Changes in stability upon charge reversal and neutralization substitution in staphylococcal nuclease are dominated by favorable electrostatic effects, Biochemistry 42 (2003) 1118–1128.

[7] J. Funahashi, K. Takano, Y. Yamagata, K. Yutani, Positive contribution of hydration structure on the surface of human lysozyme to the conformational stability, J. Biol. Chem. 277 (2002) 21792–21800.

[8] M.M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations, Protein Eng. 12 (1999) 549–555.

[9] B.W. Matthews, Studies on protein stability with T4 lysozyme, Adv. Protein Chem. 46 (1995) 249–278.

[10] M.M. Gromiha, J. An, H. Kono, M. Oobatake, H. Uedaira, A. Sarai, ProTherm: thermodynamic database for proteins and mutants, Nucleic Acids Res. 27 (1999) 286–288.

[11] K. Saraboji, M.M. Gromiha, M.N. Ponnuswamy, Average assignment method for predicting the stability of protein mutants, Biopolymers 82 (2006) 80–92.

[12] E. Capriotti, P. Fariselli, R. Calabrese, R. Casadio, Predicting protein stability changes from sequences using support vector machines, Bioinformatics 21 (Suppl 2) (2005) ii54–ii58.

[13] J. Khatun, S.D. Khare, N.V. Dokholyan, Can contact potentials reliably predict stability of proteins? J. Mol. Biol. 336 (2004) 1223–1238.

[14] E. Capriotti, P. Fariselli, R. Casadio, A neural-network-based method for predicting protein stability changes upon single point mutations, Bioinformatics 20 (Suppl 1) (2004) I63–I68.

[15] A.J. Bordner, R.A. Abagyan, Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations, Proteins 57 (2004) 400–413.

[16] H. Zhou, Y. Zhou, Stability scale and atomic solvation parameters extracted from 1023 mutation experiments, Proteins 49 (2002) 483–492.

[17] R. Guerois, J.E. Nielsen, L. Serrano, Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations, J. Mol. Biol. 320 (2002) 369–387.

[18] D. Gilis, M. Rooman, Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence, J. Mol. Biol. 272 (1997) 276–290.

[19] D. Gilis, M. Rooman, Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials, J. Mol. Biol. 257 (1996) 1112–1126.

[20] M.M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, Importance of surrounding residues for protein stability of partially buried mutations, J. Biomol. Struct. Dyn. 18 (2000) 281–295.

[21] M.M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, Relationship between amino acid properties and protein stability: buried mutations, J. Protein Chem. 18 (1999) 565–578.

[22] K. Saraboji, M.M. Gromiha, M.N. Ponnuswamy, Relative importance of secondary structure and solvent accessibility to the stability of protein mutants. A case study with amino acid properties and energetics on T4 and human lysozymes, Comput. Biol. Chem. 29 (2005) 25–35.

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

[24] M.M. Gromiha, S. Selvaraj, Important amino acid properties for determining the transition state structures of two-state protein mutants, FEBS Lett. 526 (2002) 129–134.

[25] K.A. Bava, M.M. Gromiha, H. Uedaira, K. Kitajima, A. Sarai, ProTherm, version 4.0: thermodynamic database for proteins and mutants, Nucleic Acids Res. 32 (2004) D120–D121.

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

[27] F. Eisenhaber, P. Argos, Improved strategy in analytical surface calculation for molecular system-handling of singularities and computational efficiency, J. Comput. Chem. 14 (1993) 1272–1280.

[28] B. Rost, C. Sander, R. Schneider, PHD—an automatic mail server for protein secondary structure prediction, Comput. Appl. Biosci. 10 (1994) 53–60.

[29] S. Ahmad, M.M. Gromiha, A. Sarai, RVP-net: online prediction of real valued accessible surface area of proteins from single sequences, Bioinformatics 19 (2003) 1849–1851.

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

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

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

[33] M.M. Gromiha, S. Selvaraj, A.M. Thangakani, A statistical method for predicting protein unfolding rates from amino acid sequence, J. Chem. Inf. Model 46 (2006) 1503–1508.

[34] L. Breiman, Classification and Regression Trees, Wadsworth International Group, Belmont, Calif., 1984.

[35] M.M. Gromiha, M. Oobatake, A. Sarai, Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins, Biophys. Chem. 82 (1999) 51–67.

[36] S. Thorvaldsen, E. Ytterstad, T. Fla, Property-dependent analysis of aligned proteins from two or more populations, in: T. Jiang, U.-C. Yang, Y.-P.P. Chen, L. Wong (Eds.), 4th Asia-Pacific Bioinformatics Conference, Imperial college press, UK, 2006, pp. 169–178.

[37] W.F. van Gunsteren, A.E. Mark, Prediction of the activity and stability effects of site-directed mutagenesis on a protein core, J. Mol. Biol. 227 (1992) 389–395.