BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 17 2009, pages 2181–2187
doi:10.1093/bioinformatics/btp370

Sequence analysis

Reliable prediction of protein thermostability change upon double mutation from amino acid sequence

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

[1]Department of Computer Science and Information Engineering, 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, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan

Received on March 13, 2009; revised on June 10, 2009; accepted on June 11, 2009

Advance Access publication June 17, 2009

Associate Editor: Burkhard Rost

ABSTRACT

Summary: The accurate prediction of protein stability change upon mutation is one of the important issues for protein design. In this work, we have focused on the stability change of double mutations and systematically analyzed the wild-type and mutant residues, patterns in amino acid sequence and locations of mutants. Based on the sequence information of wild-type, mutant and three neighboring residues, we have presented a weighted decision table method (WET) for predicting the stability changes of 180 double mutants obtained from thermal ($\Delta\Delta G$) denaturation. Using 10-fold cross-validation test, our method showed a correlation of 0.75 between experimental and predicted values of stability changes, and an accuracy of 82.2% for discriminating the stabilizing and destabilizing mutants.

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

Contact: michael-gromiha@aist.go.jp

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

1 INTRODUCTION

Prediction of protein stability upon mutation is one of the most important tasks in protein engineering for understanding the mechanisms responsible for protein stability as well as for designing temperature sensitive protein mutants. For the past several years this problem has been addressed with the development of different methods for discriminating the stabilizing and destabilizing mutants and predicting the stability change of proteins upon point mutations (Gromiha, 2007). These methods are broadly classified into three major categories: (i) development of distance and torsion potentials (Gilis and Rooman, 1996; Parthiban *et al.*, 2007; Yin *et al.*, 2007), (ii) statistical analysis on protein stability (Gromiha *et al.*, 1999a; Guerois *et al.*, 2002; Saraboji *et al.*, 2006) and (iii) machine learning techniques (Capriotti *et al.*, 2008; Cheng *et al.*, 2006; Huang *et al.*, 2007b; Masso and Vaisman, 2008). Most of the methods used the information from the 3D structures of proteins and recently several methods have been developed for predicting protein mutant stability from amino acid sequence (Capriotti *et al.*, 2008; Huang *et al.*, 2007b).

Studies on protein stability upon double mutations are important to understand the stability change due to the formation of ion pairs, disulfide bonds, hydrophobic packing, removal of salt bridge etc. by substituting both the residues in the formation/removal of specific interactions. Earlier, stability of double mutants has been well addressed with the concept of 'additivity'. A double mutant is said to be additive if the sum of the stabilities of the two single mutants is similar to the stability change upon double mutation. From the studies on gene V protein, Skinner and Terwilliger (1996) showed that the stability of hydrophobic core mutants are additive when the regions of structures influenced by the mutants are not overlapping with each other. Horovitz (1996) reported that if a free energy change of double mutant differs from the sum of the changes in free energy due to the single mutations then the residues at the two positions are coupled and such coupling reflects either direct or indirect interactions between the residues. The studies on the double mutations of charged residues in barnase showed that the changes in free energy of the single mutants are not additive because of the coulombic interaction energy (Serrano *et al.*, 1990). Eijsink's group extensively analyzed the stability of protein mutants with the introduction of disulfide bonds, multiple substitutions to reduce the entropy of the unfolding state, the selection of mutations for increased protein stability and proposed a simple model for relating non-additivity with the occurrence of independent partial unfolding processes that occur in parallel at elevated temperatures (Burg and Eijsink, 2002; Burg *et al.*, 1998; Mansfeld *et al.*, 1997; Vriend *et al.*, 1998). Wells (1990) analyzed the stability changes of both single and double mutants in a set of proteins and reported that the stability of a protein upon double mutants is non-additive for the mutants, which are spatially close to each other in the protein structure. Recently, Istomin *et al.* (2008) showed that the mutational effects tend to be non-additive if two structurally well-separated mutated residues belong to the same rigid cluster within the wild-type protein, and additive if they are located within different clusters. These studies demonstrated that the methods developed for predicting protein stability change upon point mutation are not suitable for predicting the stability of all double mutants.

Protein structures are stabilized by various non-covalent interactions such as hydrophobic, electrostatic, Van der Waals and hydrogen bonds (Dill, 1990; Pace, 1990; Ponnuswamy and Gromiha, 1994). Dosztanyi *et al.* (1997) reported that the mutations of residues involved in non-covalent cross-links of the structure like stabilization center elements have greater influence on stability than the rest of the residues. The importance of these non-covalent

*To whom correspondence should be addressed.

interactions to protein stability has been revealed by site directed mutagenesis experiments (Itzhaki *et al.*, 1995; Matthews, 1995; Shortle *et al.*, 1990; Yutani *et al.*, 1987). On the other hand, free-energy calculations on protein mutants have been carried out to understand/predict the contributions from different interactions and the stability change of proteins upon mutation (Bash *et al.*, 1987; Guerois *et al.*, 2002; Tidor and Karplus, 1991). These free-energy calculations require the 3D structure of a protein and extensive computations. As the gap between known amino acid sequences and 3D structures is exponentially increasing it is necessary to develop methods for predicting mutant protein stability from sequence. Montanucci *et al*. (2008) addressed this problem and developed a support vector machines based method to predict whether a set of mutations can enhance the stability of a given protein. However, this method mainly focused on multiple mutants and two-state predictions. All other sequence based methods predict the stability change of a protein only upon point mutation.

Hence, it is necessary to develop a reliable method for discriminating the stabilizing and destabilizing mutants as well as predicting the stability change of proteins upon double mutations. In this work, we have systematically analyzed the stability of double mutants available in ProTherm database and developed a weighted decision table method (WET) for predicting the stability changes of 180 double mutants obtained from thermal ($\Delta\Delta G$) denaturation. The method is mainly based on the information on wild-type residue, mutant residue and three neighboring residues on both sides of the mutant residues for the two sites. Using 10-fold cross-validation test, our method showed a correlation of 0.75 between experimental and predicted values of stability changes, and an accuracy of 82.2% for discriminating the stabilizing and destabilizing mutants. Furthermore, a web server has been developed for prediction and it is available at http://bioinformatics.myweb.hinet.net/wetstab.htm.

2 METHODS

2.1 Protein mutant dataset

In this work, we have collected a set of double mutants from ProTherm database (Gromiha *et al.*, 1999a; Kumar *et al.*, 2006) (http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html) in which the free energy changes of mutants were obtained with thermal denaturation. We removed the redundant data and averaged out the free energy changes of mutants when multiple data were reported in the literature for the same mutants with same experimental conditions. The final non-redundant dataset consists of 180 mutants acquired from 27 different proteins. The $\Delta\Delta G$ ranges between -8.9kcal/mol and 6.2 kcal/mol and the number of stabilizing and destabilizing mutants is 93 and 87, respectively. The datasets used in the present study are available at http://bioinformatics.myweb.hinet.net/wetstab.htm. The occurrence of stabilizing and destabilizing mutants in different secondary structures and at various ranges of solvent accessibility is given in Supplementary Table S1. Most of the data were obtained with reversible process (71%) and the reversibility is not reported for 24% of the data.

2.2 Sequence based attribute sets

We have utilized two different sets of attributes M_1 and M_2 as input variables to represent a double mutant. M_1 refers to the attributes for the first mutant and M_2 for the second mutant. Each attribute set consists of eight attributes, including (i) wild-type residue (attribute 1) and (iii) mutant residue (attribute 2) and (iii) three neighboring residues of the mutation site along both directions (attributes 3–8). In addition, attribute set M_{12} refers to all attributes at both sites, namely, those attributes in both M_1 and M_2. The schematic view of the attributes is shown in Supplementary Figure S1. Earlier studies have shown that three neighboring residues can efficiently predict stability changes on single mutants (Cheng *et al.*, 2006; Huang *et al.*, 2007a; Huang *et al.*, 2007b) and the wild-type and mutant residues are important attributes for discriminating stability changes (Huang *et al.*, 2006). The stability of buried mutants is influenced with the type of mutation and that of exposed ones is governed by neighboring and surrounding residues (Gromiha *et al.* 1999b). Furthermore, the free energies of pair-wise interactions of residues depends on residue composition with an exception of disulfide bonds (Dosztanyi *et al.*, 2005).

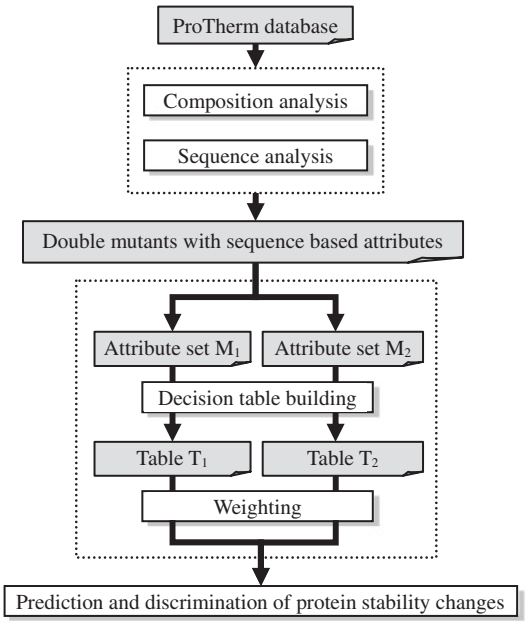

Fig. 1. The framework for analyzing and predicting stability changes of protein mutants.

2.3 Analytical framework

Figure 1 sketches our framework for analyzing and predicting stability changes of protein mutants. Firstly, the mutant dataset obtained from the thermodynamic database is analyzed in terms of composition, as well as sequence of amino acid residues. Secondly, two individual decision table models, T_1 and T_2 are built from the datasets with attribute sets M_1 and M_2, respectively. Whenever a prediction is required, it is made by weighting the results obtained from the two models.

In addition to predicting the numerical values of stability changes it can also handle the problem of discriminating the stabilizing and destabilizing mutants. Furthermore, various performance scores obtained from different attribute sets and test procedures have been assessed and compared.

2.4 Weighted decision table method

One of the simplest ways of representing the prediction model from machine learning is to make it the same as the training set. The one-to-one correspondence between prediction model and training data helps to understand the meaning of the model and especially in decision trees. In this study, the inducer of decision table majority algorithm (Kohavi, 1995) was implemented to build such a representation, called decision table majority, for the prediction of stability changes.

A decision table majority mainly consists of two components: a schema and a body. The schema is a set of attributes that are included into the table; and the body is a list of instances with values defined by the attributes in

the schema. The set of instances with the same values for attributes in the schema is named a cell.

Let A be a set of attributes given, and A' be a subset of A. For a set of instances Iover attributes in A, the induction algorithm aims to choose a optimal schema A^* such that

$$A^* = arg\min_{A' \subseteq A} E(D(A',I),f), \quad (1)$$

where $D(A',I)$ is the decision table majority with the schema A' and the body consisting of all instances in I projected on A'; $E(D,f)$ gives an error estimation between the decision table majority D and a given target function f. After the optimal schema A^* is chosen, the inducted prediction model is represented by the decision table majority $D(A^*,I)$.

Therefore, the optima schema is obtained based on a wrapper approach, which uses classifiers to evaluate the performance of attribute subsets. The algorithm is illustrated below:

function INDUCER-OF-DECISION-TABLE-MAJORITY (I, A) **returns** $D(A^*,I)$
  inputs: I, set of instances; A, set of attributes.
  outputs: $D(A^*,I)$, the decision table majority for I over A^*; A^*, the optimal schema.
  variables: *A_candidate*, set of candidate attributes; *A_confirmed*, set of attributes confirmed.
  if I is empty **then return** the default class for the predicate
  else
    A_candidate ← A
    A_confirmed ← Ø
    repeat
      for each attribute a_i of *A_candidate* **do**
        $A_i \leftarrow \{a_i\} \cup$ *A_confirmed*
        evaluate the error estimation to the decision table majority for I over A_i
        $a^* \leftarrow$ the attribute a_i that reduces the error estimation
      A_candidate ← remove attribute a^* from *A_candidate*
      A_confirmed ← add attribute a^* into *A_confirmed*
    until *A_candidate* is empty or error can not be reduced for a default number of consecutive examinations
    $A^* \leftarrow$ *A_confirmed*
  return $D(A^*,I)$

Once a decision table majority is available, the outcome of a query instance can be predicted by the following criteria. Let L be the set of instances in the cell that matches the query instance over the attributes in the schema.

(i) If $L \neq \emptyset$, return the majority class of instances in L for the nominal class (or the average class for numeric classes).
(ii) Otherwise ($L = \emptyset$), return the majority (or average) class of instances in the decision table majority.

In this study, the decision table majority has been applied to the collected double mutant dataset with attribute sets M_1, M_2 and M_{12}, respectively. We define the stability changes as a numeric class. Thus, for discrimination, the mutant is directly assigned as stabilizing or destabilizing by its predicted value. Furthermore, we have proposed a WET based on a bagging scheme (Breiman, 1996), which makes a prediction by combining the outcome of individual models and decreases the expected error by reducing the variance component. WET builds two independent models to the attribute sets M_1 and M_2, respectively, and then determines the outcomes individually for a query instance. The individually predicted values are weighted to yield a single value (also see Fig. 1). Here, we gave the same value of 0.5 for both weights, that is, under the assumption that both sites have an approximately equal effect.

The proposed WET method provides the simplest representation of decision tables, by which the prediction model can be further examined and validated. Other major advantages include the following: (i) decision tables are non-parametric based models, which make no assumption on data distribution, and hence such errors can be avoided, (ii) the optimal schema decreases the number of attributes included in the table, which reduces the computational complexity of the task, (iii) the optimal schema may provide information about the relative importance between given attributes and (iv) the weights of two independent models can be adjusted according to the relative effects of two mutation sites. It is also convenient for introducing related knowledge into the prediction model and helpful for accurate predictions.

2.5 Single correlation and mean absolute error

Two general measures are used to assess the prediction performance of stability changes. The correlation between the experimental and assigned stability ($\Delta\Delta G$) has been calculated using the familiar expression:

$$R = \left[N\sum_{i=1}^{N} X_i Y_i - \left(\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i\right)\right] \Bigg/ \sqrt{\left[N\sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i\right)^2\right]\left[N\sum_{i=1}^{N} Y_i^2 - \left(\sum_{i=1}^{N} Y_i\right)^2\right]}, \quad (2)$$

where R is the correlation coefficient, N, X_i and Y_i are the total number of mutants, experimental and predicted stability, respectively; and i varies from 1 to N. The mean absolute error (MAE) is defined as the absolute difference between predicted and experimental stability values:

$$\mathrm{MAE} = \sum_{i=1}^{N} |X_i - Y_i| / N. \quad (3)$$

2.6 Accuracy of distinguishing the stability of protein mutants

The discrimination of the stability changes (stabilizing/destabilizing) can be regarded as a binary classification problem. Several measures of accuracy are regularly used and described as follows.

The accuracy of distinguishing the stability of mutants has been determined by using the following expression: $[(TP+TN)/(TP+TN+FP+FN) \times 100\%]$, where TP, TN, FP and FN refer to the number of true positives, true negatives, false positives and false negatives, respectively. Other measures includes sensitivity $[TP/(TP+FN) \times 100\%]$, specificity $[TN/(TN+FP) \times 100\%]$, accuracy for positives $[TP/(TP+FP) \times 100\%]$ and accuracy for negatives $[TN/(TN+FN) \times 100\%]$.

In addition, 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. The area under the curve (AUC) is calculated to summarize a curve in a single quantity. Generally speaking, the larger the area is, the better performance the curve has.

2.7 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 includes all the stability data for training the prediction model of decision tables, and prediction has been 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 models a decision table from the remaining data and uses it to predict the stability of the sub-sample. The procedure is repeated n times to obtain the mean measure.

3 RESULTS AND DISCUSSIONS

3.1 Analysis of neighboring residues for stabilizing and destabilizing mutant proteins

We have analyzed the preference of neighboring residues for stabilizing and destabilizing mutant proteins and the results obtained for frequently occurring nearest neighboring residues is presented in Tables S2 and S3. From this table, we observed that the segments

Table 1. Prediction performance of stability changes ($\Delta\Delta G$) for different test procedures and attribute sets on a set of 180 mutants by 10-fold cross-validation

Performance measure	Method			
	DT & M_1	DT & M_2	DT & M_{12}	WET
R	0.71	0.73	0.73	0.75
MAE	1.27	1.22	1.22	1.17
Accuracy (%)	75.6	75.0	75.0	82.2
Sensitivity (%)	65.6	61.3	61.3	76.3
Specificity (%)	86.2	89.7	89.7	88.5

R: correlation coefficient; MAE: mean absolute error; DT: decision table majority method; M_1: mutant 1; M_2: mutant 2; M_{12}: mutants 1 and 2.

with motifs F*V (* denotes any mutation site), L*L, K*A and A*E stabilize the proteins upon amino acid substitutions. Further analysis on nearest neighboring residues of the mutation sites in all the considered mutant proteins showed that the pair of residues with hydrophobic–hydrophobic, hydrophobic–polar, polar–hydrophobic and polar–polar motifs stabilize the proteins with the frequency of 48.9%, 19.4%, 21.5% and 8.6%, respectively. This result indicates that the hydrophobic environment increases the stability of protein mutants. On the other hand, the frequencies of occurrence of destabilizing mutants are 24.3%, 24.3%, 26.5% and 17.8%, respectively. It can be seen from Supplementary Table S2 that the mutants with motifs K*A, G*P, K*Q and L*Q destabilize the proteins upon double mutation.

3.2 Prediction of protein stability changes upon double mutation

We have utilized the information on wild-type residue, mutant residue and three neighboring residues on both sides of double mutants to discriminate the stabilizing and destabilizing mutants as well as predicting the stability change upon mutation. The results are presented in Table 1. Our method could discriminate the stabilizing and destabilizing mutants with an accuracy of 82.2% and the predicted stability change of the protein upon double mutation showed a correlation of 0.75 with experimental data. On the other hand, the data trained with a simple artificial neural network model with one hidden layer and two nodes showed an accuracy of 59.4% and correlation of 0.33.

In Figure 2, we show the relationship between experimental and predicted stability change of 180 protein mutants obtained with 10-fold cross-validation method and the dark line shows the linear fit between them. We have also estimated ROC curves for different attribute sets (Fig. 3). We observed that the AUC values are 0.81, 0.83, 0.83 and 0.87 for M_1, M_2, M_{12} and WET, respectively.

We have compared the performance of our method using the information about only one mutant (either M_1 or M_2) as well as with both mutants. In addition, decision table majority method and WET has been compared for the performance. We observed that the WET method with both M_1 and M_2 information has the best performance, showing the close relationship between experimental and predicted protein stability change upon double mutation.

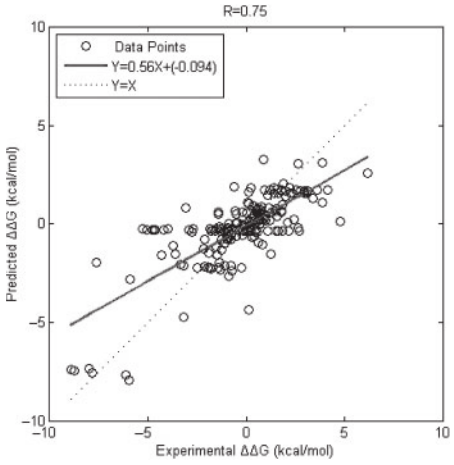

Fig. 2. Relationship between experimental and predicted $\Delta\Delta G$ based on Relationship between experimental and predicted $\Delta\Delta G$ based on WET with 10-fold cross-validation test on a set of 180 mutants (correlation coefficient is 0.75).

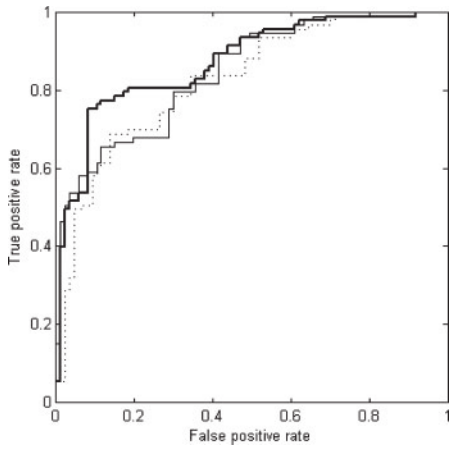

Fig. 3. Receiver operating characteristic curves for different attribute sets with 10-fold cross-validation test (dotted, thin and thick lines for M_1, M_2/M_{12} and WET, respectively).

3.3 Role of secondary structure and solvent accessibility

We have analyzed the performance of our method based on the location of the mutants at different secondary structures and various ranges of solvent accessibility. The results obtained for different combinations of mutants are presented in Table 2 along with the number of mutants in each category. We observed that the stability of mutations at regular structures (HH or SS) are better predicted than that in loops (CC). Furthermore, the stability of mutations in strand segments showed the accuracy of 89%, which is better than that in helical segments (78%). This might be due to the presence of residues in β-strand segments, which are close in space but far away in sequence (Gromiha and Selvaraj, 2004).

The analysis on solvent accessibility showed that the stability of mutants in hydrophobic core regions can be predicted with an accuracy of 87%. The accuracies for partially buried and exposed

Table 2. Influence of secondary structure and solvent accessibility for predicting stability change of double mutation on 10-fold cross-validation

	Number of data	Number of correct prediction	Accuracy (%)
Secondary structure			
HH	60	47	78.3
SS	36	32	88.9
CC	11	8	72.7
HS	15	12	80.0
SC	20	18	90.0
HC	38	31	81.6
Solvent accessibility			
BB	68	59	86.8
BP	25	23	92.0
BE	12	9	75.0
PP	30	25	83.3
PE	26	17	65.4
EE	19	15	78.9

H: Helix; S: strand; C: coil; B: buried (0–20% ASA); P: partially buried (20–50% ASA); E: exposed (>50% ASA).

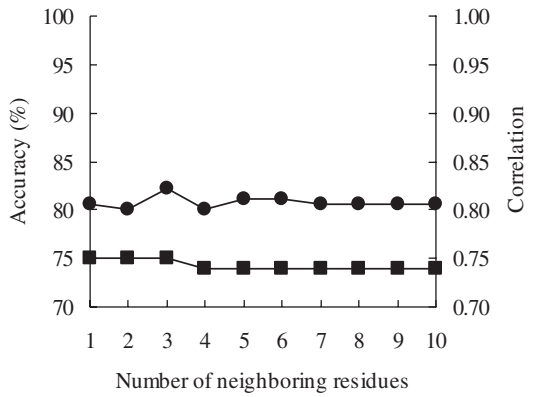

Fig. 4. Variation of accuracy (circles) and correlation (squares) with different number of neighboring residues.

mutations are 83% and 79%, respectively. The stability of buried mutants mainly depends on the wild-type and mutant residues due to the dominance of non-specific interactions (Gromiha *et al.*, 1999b) and it might be a reason for obtaining high accuracy for core mutants.

3.4 Influence of neighboring residues for protein stability change prediction

We have analyzed the influence of neighboring residues for discriminating the stabilizing and destabilizing mutants and predicting the stability change upon double mutations. The results obtained for different window lengths are shown in Figure 4.

The inclusion of three neighboring residues showed the highest accuracy and correlation for discrimination and prediction, respectively. The performance did not improve with additional neighboring residues.

Table 3. Comparison between the present method and other methods on a set of 180 mutants

Performance measure	Method					
	iPTREE	ANN	SVM	LR	RBFN	WET
R	0.55	0.54	0.62	0.62	0.24	0.75
MAE	1.63	2.58	2.75	1.99	1.79	1.17
Accuracy (%)	71.7	68.3	76.7	72.2	55.6	82.2
Sensitivity (%)	53.8	65.6	74.2	66.7	20.4	76.3
Specificity (%)	90.8	71.3	79.3	78.2	93.1	88.5

R: correlation coefficient; MAE: mean absolute error; iPTREE (Huang *et al.*, 2007b); ANN: artificial neural networks (Capriotti *et al.*, 2004); SVM: support vector machines (Capriotti *et al.*, 2005; Cheng *et al.*, 2006); LR: linear regression models; RBFN: radial basis function network (Grosfils *et al.*, 2008); WET: weighted decision table (present work).

3.5 Performance of methods developed for predicting protein stability change of point mutants

We have analyzed the performance of different methods developed for predicting the stability change of proteins upon single mutations. We have estimated the $\Delta\Delta G$ for both positions and added the stability values for overall prediction for double mutations. The results obtained with five different methods are given in Table 3. These methods are based on decision tree models (Huang *et al.*, 2007b), (ii) artificial neural networks (Capriotti *et al.*, 2004), (iii) support vector machines (Capriotti *et al.*, 2005; Cheng *et al.*, 2006), (iv) linear regression models and (v) radial basis function network (Grosfils *et al.*, 2008). We observe that the accuracy lies in the range of 55–77% for most of the methods. The method RBFN has the highest specificity of 93% whereas the sensitivity is very low (20%). SVM showed the accuracy of 77% and the correlation is 0.62. The present method discriminated the stabilizing and destabilizing mutants with an accuracy of 82% and the correlation between the experimental and predicted stability change upon double mutations is 0.75. The best performance of the present method is because of the fact that it could adequately take into account both the additive and non-additive mutants whereas the methods developed for single mutants are not able to predict the stability of non-additive mutants.

3.6 Prediction on the web

We have developed a web server, WET–STAB, for predicting protein stability changes upon double mutations from amino acid sequence and it is freely available at http://bioinformatics.myweb.hinet.net/wetstab.htm. The input options for the server are illustrated in Figure 5a. It takes the information on wild-type, mutant and three neighboring residues for M_1 and M_2 as input. As an example, we provided the data for the mutants N116D and R119M in 2LZM. After submitting this query, WET-STAB brings out the prediction results within a minute. In the out put, we display the input information and predicted $\Delta\Delta G$ values using three different methods (i) M_1 using DT, (ii) M_2 using DT and (iii) M_{12} using WET method. In addition, the discrimination of stability changes (stabilizing and destabilizing) is also given. In this example WET showed the predicted $\Delta\Delta G$ of 0.147 kcal/mol, which is similar to experimental observation, 0.15 kcal/mol (Fig. 5b).

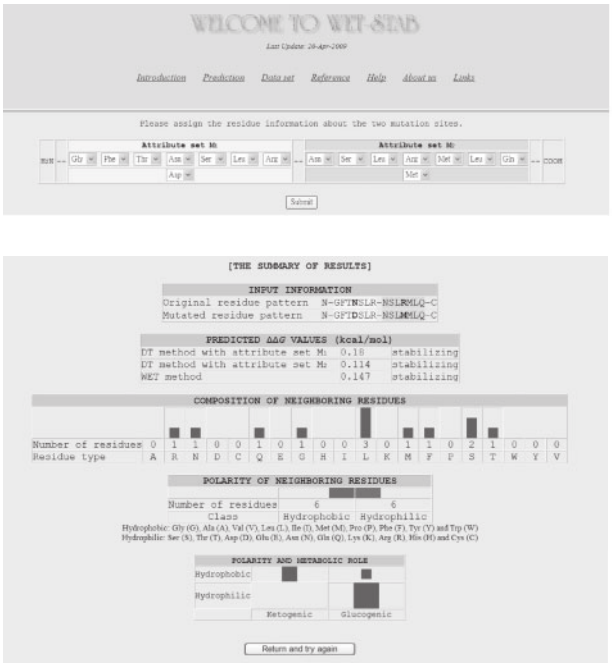

Fig. 5. (**a**) Snapshot of the prediction page showing the residue information about two mutation sites (**b**) Snapshot of the result page generated by WET–STAB.

4 CONCLUSIONS

We have developed a WET for discriminating the stabilizing and destabilizing mutants and predicting the stability change upon double mutation by utilizing the information on wild-type, mutant and three neighboring residues of both mutation sites. Our method showed an accuracy and correlation of 82% and 0.75, respectively, for discrimination and prediction. Furthermore, we have analyzed the influence of neighboring residues and the ability of methods developed for predicting the stability changes of single mutants. A web server has been developed for discrimination and prediction purposes.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for constructive comments and Dr. Paul Horton for critical reading of the manuscript.

Conflict of Interest: none declared.

REFERENCES

Bash,P.A. *et al.* (1987) Free energy calculations by computer simulation. *Science*, **236**, 564–568.

Breiman,L. (1996) Bagging predictors, *Machine Learn.*, **24**, 123–140.

Burg,B.V.D. and Eijsink,V.G. (2002) Selection of mutations for increased protein stability. *Curr. Opin. Biotechnol.*, **13**, 333–337.

Burg,B.V.D. *et al.* (1998) Engineering an enzyme to resist boiling. *Proc. Natl Acad. Sci. USA*, **95**, 2056–2060.

Capriotti,E. *et al.* (2004) A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics*, **20** (Suppl. 1), I63–I68.

Capriotti,E. *et al.* (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, W306–W310.

Capriotti,E. *et al* (2008) A three-state prediction of single point mutations on protein stability changes, *BMC Bioinformatics*, **9** (Suppl. 2), S6.

Cheng,J. *et al.* (2006) Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, **62**, 1125–1132.

Dill,K.A. (1990) Dominant forces in protein folding. *Biochemistry*, **29**, 7133–7155.

Dosztanyi,Z., *et al.* (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.

Dosztanyi, Z. *et al.* (1997) Stabilization centers in proteins: identification, characterization and predictions. *J. Mol. Biol.*, **272**, 597–612.

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

Gromiha,M.M. (2007) Prediction of protein stability upon point mutations. *Biochem. Soc. Trans.*, **35**, 1569–1573.

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. and Selvaraj,S. (2004) Inter-residue interactions in protein folding and stability. *Prog. Biophys. Mol. Biol.*, **86**, 235–277.

Grosfils,A. *et al.* (2008) Neural networks to predict protein stability changes upon mutation. In *Proceedings of the 17th International Federation of Automatic Control World Congress, The International Federation of Automatic Control*. Seoul, Korea, pp. 12619–12624.

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.

Horovitz,A. (1996) Double-mutant cycles: a powerful tool for analyzing protein structure and function, *Fold Des.*, **1**, R121–R126.

Huang,L.T. *et al.* (2007a) Sequence analysis and rule development of predicting protein stability change upon mutation using decision tree model. *J. Mol. Model.*, **13**, 879–890.

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

Huang,L.T. *et al.* (2006) Knowledge acquisition and development of accurate rules for predicting protein stability changes. *Comput. Biol. Chem.*, **30**, 408–415.

Istomin,A.Y. *et al.* (2008) New insight into long-range nonadditivity within protein double-mutant cycles. *Proteins*, **70**, 915–924.

Itzhaki,L.S. *et al.* (1995) The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.*, **254**, 260–288.

Kohavi,R. (1995) The power of decision tables. *Lecture Notes Comp. Sci.*, **912**, 174–189.

Kumar,M.D. *et al.* (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–D206.

Mansfeld,J. *et al.* (1997) Extreme stabilization of a thermolysin-like protease by an engineered disulfide bond. *J. Biol. Chem.*, **272**, 11152–11156.

Masso,M. and Vaisman,I.I. (2008) Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics*, 24, 2002–2009.

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

Montanucci,L. *et al.* (2008) Predicting protein thermostability changes from sequence upon multiple mutations. *Bioinformatics*, **24**, i190–i195.

Pace,C.N. (1990) Conformational stability of globular proteins. *Trends Biochem. Sci.*, **15**, 14–17.

Parthiban,V. *et al.* (2007) Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. *Proteins*, **66**, 41–52.

Ponnuswamy,P.K. and Gromiha,M.M. (1994) On the conformational stability of oligonucleotide duplexes and tRNA molecules. *J. Theor. Biol.*, **169**, 419–432.

Saraboji,K., *et al.* (2006) Average assignment method for predicting the stability of protein mutants. *Biopolymers*, **82**, 80–92.

Serrano,L. *et al.* (1990) Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. *Biochemistry*, **29**, 9343–9352.

Shortle,D. *et al.* (1990) Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry*, **29**, 8033–8041.

Skinner,M.M. and Terwilliger,T.C. (1996) Potential use of additivity of mutational effects in simplifying protein engineering. *Proc. Natl Acad. Sci. USA*, **93**, 10753–10757.

Tidor,B. and Karplus,M. (1991) Simulation analysis of the stability mutant R96H of T4 lysozyme. *Biochemistry*, **30**, 3217–3228.

Vriend,G. *et al*. (1998) Early steps in the unfolding of thermolysin-like proteases. *J. Biol. Chem.*, **273**, 35074–35077.

Wells,J.A. (1990) Additivity of mutational effects in proteins. *Biochemistry*, **29**, 8509–8517.

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

Yutani,K. *et al.* (1987) 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. USA*, **84**, 4441–4444.