Sequence Based Prediction of Protein Mutant Stability and Discrimination of Thermophilic Proteins

M. Michael Gromiha[1], Liang-Tsung Huang[2], and Lien-Fu Lai[3]

[1] 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
michael-gromiha@aist.go.jp
[2] Department of Computer Science and Information Engineering, MingDao University, Changhua 523, Taiwan
larry@mdu.edu.tw
[3] Department of Computer Science and Information Engineering, National Changhua University of Education, Changhua 500, Taiwan

Abstract. Prediction of protein stability upon amino acid substitution and discrimination of thermophilic proteins from mesophilic ones are important problems in designing stable proteins. We have developed a classification rule generator using the information about wild-type, mutant, three neighboring residues and experimentally observed stability data. Utilizing the rules, we have developed a method based on decision tree for discriminating the stabilizing and destabilizing mutants and predicting protein stability changes upon single point mutations, which showed an accuracy of 82% and a correlation of 0.70, respectively. In addition, we have systematically analyzed the characteristic features of amino acid residues in 3075 mesophilic and 1609 thermophilic proteins belonging to 9 and 15 families, respectively, and developed methods for discriminating them. The method based on neural network could discriminate them at the 5-fold cross-validation accuracy of 89% in a dataset of 4684 proteins and 91% in a test set of 707 proteins.

Keywords: Protein stability, rule generator, discrimination, prediction, thermophilic proteins, neural network, machine learning techniques.

1 Introduction

One of the most important tasks in protein engineering is to understand the mechanisms responsible for protein stability changes affected by single point mutations, which can be employed for constructing temperature sensitive mutants and used to identify a wide spectrum of drug resistance conferring mutations. Another related task is to understand the important factors for the extreme stability of thermophilic proteins and discriminating them from mesophilic ones.

M. Chetty, A. Ngom, and S. Ahmad (Eds.): PRIB 2008, LNBI 5265, pp. 1–12, 2008.

Several methods have been proposed for predicting the stability of proteins upon amino acid substitutions. These methods are mainly based on distance and torsion potentials [1,2], multiple regression techniques [3], energy functions [4], contact potentials [5], neural networks [6], support vector machines, SVMs [7,8], average assignment [9], classification and regression tool [10], backbone flexibility [11] etc. Further, it has been reported that the discrimination of stabilizing and destabilizing mutants is more important than its magnitude in many cases [6]. Most of these methods used the information from the three-dimensional structures of proteins for discrimination/prediction. On the other hand, prediction accuracy using amino acid sequence is significantly lower than that with structural data [12].

Several attempts have been made to understand the factors influencing the stability of thermophilic proteins using three-dimensional structural information as well as from amino acid sequence. It has been reported that increase in number of salt bridges and side chain-side chain interactions [13], counterbalance between packing and solubility [14], aromatic clusters [15], contacts between the residues of hydrogen bond forming capability [16,17], ion pairs [18], cation-π interactions [19,20], non-canonical interactions [21], electrostatic interactions of charged residues and the dielectric response [22,23], amino acid coupling patterns [24], main-chain hydrophobic free energy [25] and hydrophobic residues [26] in thermophilic proteins enhanced the stability. In addition, the amino acid sequences of genomes have been used for understanding the stability of thermophilic proteins. Das and Gerstein [27] reported that intra-helical salt bridges are prevalent in thermophiles. Fukuchi and Nishikawa [28] showed that the amino acid composition on protein surface may be an important factor for understanding the stability. Ding et al. [29] revealed the preferences of dipeptides in thermophilic proteins for extreme stability. Berezovsky et al. [30] found that the proteomes of thermophilic proteins are enriched in hydrophobic and charged amino acids at the expense of polar ones.

In spite of these studies, it is necessary to build a system, which derives stability rules for any input data and convert them into prediction. In this work, we have developed a classification rule generator to provide an online service for relating protein stability changes from the information about the mutated residue, three neighboring residues and the mutant residue. The rules can be interpreted to understand and predict protein stability changes upon point mutations. We have developed a method based on decision tree for discriminating /predicting protein mutant stability just from amino acid sequence. Using the information of a short window of seven residues (three residues on both directions of the mutant site) our method discriminated the stabilizing and destabilizing mutants with an accuracy of 82% and predicted the stability changes with a correlation of 0.70. Further, we have analyzed the performance of different algorithms, such as Bayes rules, neural network, SVM, decision trees etc for discriminating mesophilic and thermophilic proteins. We found that the 5-fold cross-validation accuracy is almost similar in most of the machine learning algorithms and the accuracy of discriminating mesophilic and thermophilic proteins using neural networks is marginally better than other methods. It could discriminate them at an accuracy of 93% and 89%, respectively, for self-consistency and 5-fold cross-validation tests in a dataset of 4684 proteins.

2 Materials and Methods

We have used different sets of data for predicting protein stability upon point mutations, and discriminating mesophilic and thermophilic proteins. Likewise, different methods have been used for these two studies.

2.1 Datasets

For the study on protein mutant stability, we have constructed a dataset of 1859 non-redundant single mutants from 64 proteins using ProTherm, the thermodynamic database for proteins and mutants available on the web [31,32]. We have removed the duplicate mutants that have same mutated and mutant residues, residue number, experimental conditions (pH and temperature, T) and $\Delta\Delta G$ values. Further, we retained only one data (the average value) for the mutants in which $\Delta\Delta G$ are reported with same T and pH, and different conditions (buffers/ions). We have used five variables for implementing the discrimination/prediction algorithm: (i) Md, mutated (deleted) residue, (ii) Mi, mutant (introduced) residue, (iii) pH, (iv) T (°C) at which the stability of the mutated protein was measured explicitly and (v) three neighboring residues of the central residue. These attributes have been selected with the balance between experimental conditions and sequence information.

Zhang and Fang [33] used 4895 mesophilic and 3522 thermophilic proteins for discriminating them using dipeptide composition. The proteins in each set contain many redundant sequences and we removed the redundancy using CD-HIT algorithm, [34] as implemented by Holm and Sander [35]. The final dataset contains 3075 mesophilic proteins and 1609 thermophilic proteins. Further, we have used a test set of 325 mesophilic and 382 thermophilic proteins belonging to Xylella fastidosa and Aquifex aeolicus families, respectively. These datasets have the proteins with less than 40% sequence identity.

2.2 Computation of Amino Acid Composition

The amino acid composition for each protein has been computed using the number of amino acids of each type and the total number of residues:

$$\text{Comp}(i) = \Sigma\, n_i/N, \tag{1}$$

where i stands for the 20 amino acid residues; n_i is the number of residues of each type and N is the total number of residues. The summation is through all the residues in the particular protein.

2.3 Methods for Discrimination and Prediction

We have used decision tree [36] along with adaptive boosting algorithm [37] for discriminating the stability of protein mutants, and classification and regression tree (CART) [38] for predicting the stability changes of proteins upon mutations. The decision tree algorithms can efficiently construct interpretable prediction models by measuring input variables directly from training data, which is suitable for large datasets and unknown data distribution. The decision tree has been selected with two

steps: in the first step, a recursive split procedure builds a tree, named maximum tree, which closely describes the training dataset and in the second step, the maximum tree is cut off for finding optimal sub tree. The adaptive boosting algorithm generates a set of classifiers from the data, each optimized to classify the correct ones that were misclassified in previous pass. Considering the exploitation of sets of hypotheses with independent errors it can improve the classification accuracy and reduce the variance as well as the bias.

We have analyzed several machine learning techniques implemented in WEKA program [39] for discriminating mesophilic and thermophilic proteins. This program includes several methods based on Bayes functions, neural networks, logistic functions, support vector machines, regression analysis, nearest neighbor methods, meta learning, decision trees and rules. The details of these methods have been explained in our earlier article [40]. We have analyzed different classifiers and datasets to discriminate mesophilic and thermophilic proteins.

2.4 Assessment of Predictive Ability

We have used different measures to assess the accuracy of discriminating mesophilic and thermophilic proteins, and stabilizing and destabilizing mutants. The term, sensitivity shows the correct prediction of thermophiles (stabilizing mutants), specificity about the mesophilies (destabilizing mutants) and accuracy indicates the overall assessment. The agreement between experimental and predicted stability changes has been assessed with correlation coefficient. These terms are defined as follows:

$$\text{Sensitivity} = TP/(TP+FN) \tag{2}$$

$$\text{Specificity} = TN/(TN+FP) \tag{3}$$

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \tag{4}$$

$$r = [N \Sigma XY - (\Sigma X\ \Sigma Y)]/\{[N \Sigma X^2 - (\Sigma X)^2]\ [N \Sigma Y^2 - (\Sigma Y)^2]\}^{1/2} \tag{5}$$

where, TP, FP, TN and FN refer to the number of true positives, false positives, true negatives, and false negatives respectively; r is the correlation coefficient, N, X, and Y are the number of data, experimental and predicted stability, respectively.

We have performed n-fold cross-validation test for assessing the validity of the present work. In this method, the data set is divided into n groups, n-1 of them are used for training and the rest is used for testing the method. The same procedure is repeated for n times and the average is computed for obtaining the accuracy of the method. We have carried out 2-fold, 3-fold, 4-fold, 5-fold and 10-fold cross validation tests.

3 Results and Discussion

3.1 Development of Classification Rules

We have developed a system composed of three components, which can sequentially develop protein sequence information to classification rules along with related analysis (Figure 1).

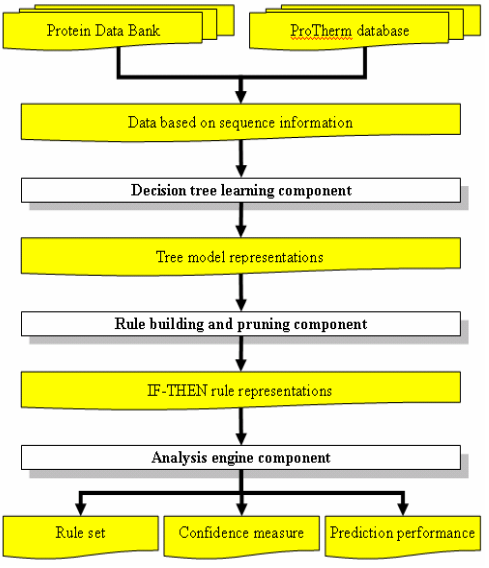

Fig. 1. Flowsheet of the learning process for depicting the relationship between components and data

The first component constructs a decision tree from the information about the mutated residue with three neighboring residues and the mutant residue. The mutation and neighboring residues information have been obtained from ProTherm database [31,32] and Protein Data Bank [41], respectively. Then the second one converts the learned tree into an equivalent set of rules, which may discriminate the stabilizing and destabilizing mutants as well as to explore the underlying concept of experimental data. The third provides further analyses from different viewpoints to clarify the characteristics of generated rules.

From the dataset of 1859 mutants, a total of 104 rules were generated. The rule size of the rule set being about 2 indicates the antecedent of these rules consist of about two statements on average. Generally, a shorter rule may make the rule easier to understand and to be examined. We further observed that 1535 samples of the dataset can match the antecedent of these rules with 175 errors, which showed the accuracy of 88.6%. It reveals that most samples in the dataset can be correctly inferred by using the rule set. In Table 1, we have given few examples of rules and their details: (i) if the mutated residue is Asp, its third neighbor at N-terminal is Glu, and its second neighbor at C-terminal is Leu, then the predicted stability change will be positive (stabilizing); we obtained an accuracy 96% in a set of 25 data; (ii) if the deleted residue is Ser and its first neighbor at N-terminal is Pro, then the predicted stability change will be negative (destabilizing), which correctly predicted all the 29 data with an accuracy of 100%; (iii) if the deleted residue is Leu, then the protein will be destabilizing; this rule is applied to 122 mutants and 115 are predicted correctly (accuracy 94%).

Table 1. Confidence measure for 5 rules with high accuracy and sufficient number of data from a dataset of 1859 non-redundant single mutants

Rule	Rule size	Number of data	Percentage of data (%)	Correctly predicted	Accuracy (%)	Predicted class
Mutated residue=D, N3=E, C2=L	3	25	1.34	24	96	Stabilizing
Mutated residue=T, C1=V	2	29	1.56	24	83	Stabilizing
Mutated residue=S, N1=P	2	29	1.56	29	100	Destabilizing
Mutated residue=L	1	122	6.56	115	94	Destabilizing
Mutant residue=G	1	66	3.55	62	94	Destabilizing

We have developed a web interface for generating rules for any set of stability data using wild type, mutant and three neighboring residue information. We have also provided the related dataset for different tests along with the generated rules on the web server.

3.2 Prediction of Protein Stability

We have utilized the rules for discriminating the stabilizing and destabilizing mutants and predicting the stability change upon mutation along with the information about pH and T. The validity of our approach has been assessed with 4-fold, 10-fold and 20-fold cross-validation procedures. The 4-fold and 20-fold cross-validation tests yielded the accuracy of 81.4% and 82.1% for discriminating the stability of protein mutants. The sensitivity and specificity are 75.3% and 84.5%, respectively [42]. Further, our method could predict the stability of protein mutants with the correlation coefficient of 0.70.

The main features of the present method are: (i) it is based on the neighboring residues of short window length, (ii) it can predict the stability from amino acid sequence alone, (iii) developed different servers for discrimination and prediction, and integrated them together, (iv) utilized the information about experimental conditions, pH and T, and (v) implemented several rules for discrimination and prediction from the knowledge of experimental stability and input conditions: (i) if the deleted residue is Ala and the neighboring residues contain Gln, then the predicted stability change will be negative (accuracy = 97.1%), (ii) if the deleted residue is Glu and its second neighbor at N-terminal is Met, the mutation stabilizes the protein (accuracy = 100%) and (iii) if the deleted-residue belongs to Y, W, V, R, P, M, L, I, G, F or C, and the introduced-residue belongs to T, S, P, K, H, G or A, then the predicted stability change will be -2.05 kcal/mol (mean absolute error = 1.57 kcal/mol).

We have developed a web server for discriminating the stabilizing and destabilizing mutants and predicting the stability of proteins upon mutations. The program takes the information about the mutant and mutated residues, three neighboring residues on both sides of the mutant residue along with pH and T. In the output, we display the predicted protein stability change upon mutation along with input conditions (Figure 2). In the case

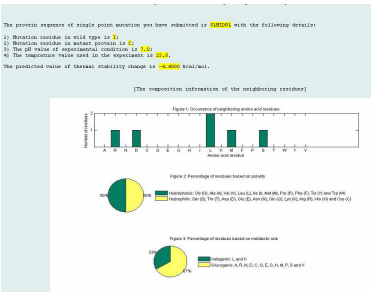

Fig. 2. The results obtained for predicting the stability change along with the related information of neighboring residues

of discrimination, we show the effect of the mutation to protein stability, whether stabilizing or destabilizing. Both discrimination and prediction services offer an option for additional sequence composition information of neighboring residues (Figure 2). The bar chart shows the number of amino acids of each type. The two pie charts below represent the percentage of residues according to polarity and the metabolic role of amino acids. The prediction/discrimination results are available at http://bioinformatics.myweb.hinet.net/iptree.htm.

In our method, we have used the balance between experimental conditions and sequence information as features for prediction. These features are different from other methods, which mainly used contact potentials, 40 different combinations of mutations, solvent accessibility, secondary structure, average stability value for each mutation, experimental conditions etc. for predicting the stability. In addition, we have used different features including the variation of window length along the sequence and we observed the best performance with the information about mutant and mutated residues as well as three neighboring residues along the sequence.

We have compared the performance of CART with neural networks (NN) and support vector machines (SVM) using same features. The ROC curve obtained for the three methods with 20-fold cross-validation test is shown in Figure 3. We observed

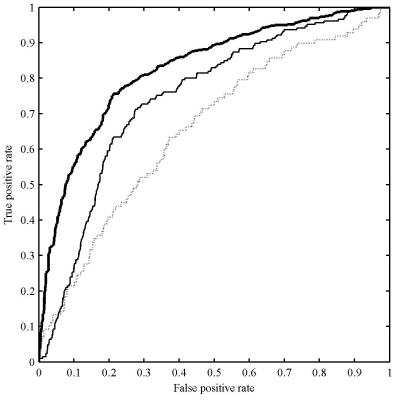

Fig. 3. ROC curves for CART (thick line), SVM (thin line) and NN (broken line)

that the performance of CART is the best among all the three methods. The areas under the curve (AUC) for CART, SVM and NN are 0.83, 0.75 and 0.66, respectively.

3.3 Discrimination of Mesophilic and Thermophilic Proteins

We have computed the amino acid composition of mesophilic and thermophilic proteins and the results are shown in Figure 4. From this figure, we observed that the composition of Ala, Leu, Gln and Thr are higher in mesophiles than thermophiles an opposite trend is observed for Glu, Lys, Arg and Val [43]. These preferences and the higher occurrence of other amino acids in thermophilic proteins reveal the implications for protein stability.

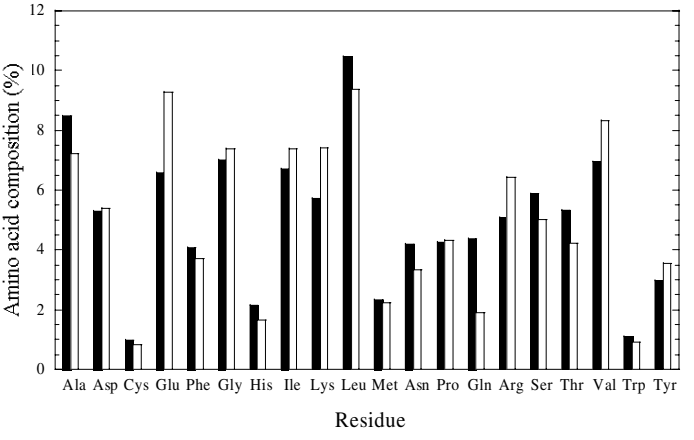

Fig. 4. Amino acid composition in mesophilic (■) and thermophilic (□) proteins

The comparative analysis on the occurrence of Cys, Ile and Val in the structural homologues of 23 mesophilic and thermophilic proteins [25] showed that the occurrence of Cys is less in thermophiles than mesophiles. On the other hand, the occurrence of Val/Ile is higher in thermophiles than mesophiles. In addition, it has been reported that Cys can be replaced by Val/Ile to enhance the stability [14]. Interestingly, these trends were reflected in the analysis of amino acid composition. Further, the charged residues, Lys, Arg and Glu have significantly higher occurrence in thermophilic proteins than mesophilic ones and the composition of Asp showed a moderate difference (Figure 4). We have analyzed the composition of charged residues in the structural homologues of thermophilic and mesophilic proteins and observed that the thermophiles have more number of charged residues than mesophiles. This result supports our observation obtained with amino acid sequence analysis.

We have analyzed the performance of different machine learning techniques for discriminating mesophilic and thermophilic proteins. In this discrimination, we have used the amino acid composition as the main attributes. We observed that most of the machine learning methods discriminated the mesophilic and thermophilic proteins with the accuracy in the range of 84-89% in a set of 4684 proteins. This analysis showed that there is no significant difference in performance between different

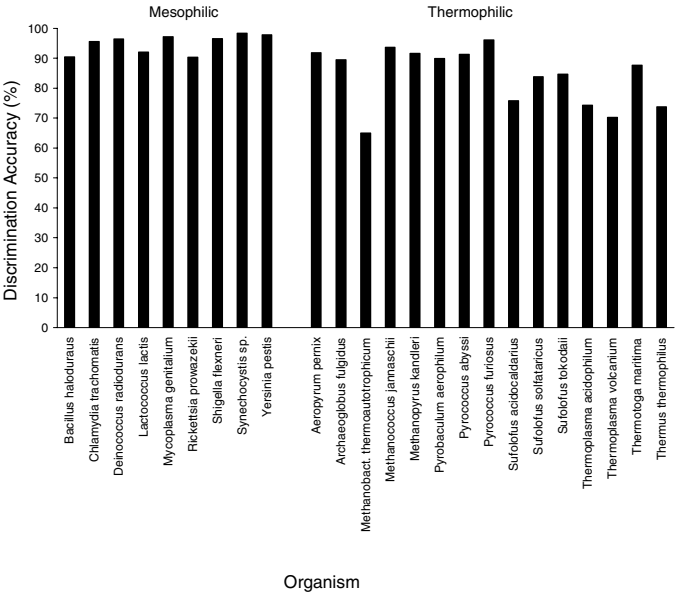

Fig. 5. Discrimination accuracy in different mesophilic and thermophilic organisms

machine learning methods. Interestingly, the methods neural networks, support vector machines and logistic functions discriminated mesophilic and thermophilic proteins at similar accuracy of 89%. The accuracy of identifying thermophilic proteins is 87% where as that of excluding mesophilic proteins is 96%. The overall accuracy is 89.4% for distinguishing mesophilic and thermophilic proteins.

The accuracy of discriminating mesophilic and thermophilic proteins in different families has been analyzed and the results are depicted in Figure 5.

We observed that the proteins in most of the mesophilic families are discriminated with the accuracy of more than 90%. On the other hand, the accuracy of discriminating thermophilic proteins showed a vide variation of 65 to 96%. Further analysis on this family of proteins revealed that the number of proteins in this family is significantly less (20 proteins) and most of the proteins are showing high sequence identity with mesophilic proteins. In addition, we have analyzed the discrimination accuracy of thermophilic (moderate) and hyper (extreme) thermophilic proteins from mesophilic proteins. Interestingly, we observed that hyper-thermophilic proteins are discriminated with higher accuracy than moderate thermophilic proteins. The accuracies of discriminating hyper-thermophilic and thermophilic proteins from mesophilic ones are, 90% and 73%, respectively.

We have assessed the reliability of the present method by discriminating mesophilic and thermophilic proteins from different families that are not considered in the work for training/ testing. We have collected the data of 325 mesophilic and 382 thermophilic proteins from Xylella fastidosa and Aquifex aeolicus families, respectively. We observed that the present method based on neural networks correctly identified the thermophilic proteins with the sensitivity of 87.6%. Further, the

mesophilic proteins are excluded with the specificity of 95.7% and the overall accuracy is 91.3%. These results demonstrated that our method is performing extremely well in distinguishing mesophilic and thermophilic proteins.

4 Conclusions

We have developed a rule generator for classifying the stabilizing and destabilizing protein mutants based on wild type, mutant and three neighboring residue information. These rules have been effectively used to discriminate the stabilizing and destabilizing mutants, and predicting the stability of a protein upon point mutation. Our method could achieve the accuracy of 82% and a correlation of 0.70 for discrimination and prediction, respectively, just from amino acid sequence. Further, different machine learning techniques have been analyzed for discriminating the mesophilic and thermophilic proteins and showed that these proteins are discriminated with the accuracy of 89%. Our method used simple features and achieved high accuracy and hence it is suitable for prediction. We suggest that our method could be effectively used in protein design.

References

1. Gilis, D., Rooman, M.: Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials. J. Mol. Biol. 257, 1112–1126 (1996)
2. Parthiban, V., Gromiha, M.M., Hoppe, C., Schomburg, D.: Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. Proteins 66, 41–52 (2007)
3. Gromiha, M.M., Oobatake, M., Kono, H., Uedaira, H., Sarai, A.: 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 (1999)
4. Guerois, R., Nielsen, J.E., Serrano, L.: Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J. Mol. Biol. 320, 369–387 (2002)
5. Khatun, J., Khare, S.D., Dokholyan, N.V.: Can contact potentials reliably predict stability of proteins? J. Mol. Biol. 336, 1223–1238 (2004)
6. Capriotti, E., Fariselli, P., Casadio, R.: A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics 20(1), i63–68 (2004)
7. Capriotti, E., Fariselli, P., Casadio, R.: I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 33, w306–310 (2005)
8. Cheng, J., Randall, A., Baldi, P.: Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62, 1125–1132 (2006)
9. Saraboji, K., Gromiha, M.M., Ponnuswamy, M.N.: Average assignment method for predicting the stability of protein mutants. Biopolymers 82, 80–92 (2006)
10. Huang, L.T., Saraboji, K., Ho, S.Y., Hwang, S.F., Ponnuswamy, M.N., Gromiha, M.M.: Prediction of protein mutant stability using classification and regression tool. Biophys. Chem. 125, 462–470 (2007)
11. Yin, S., Ding, F., Dokholyan, N.V.: Modeling backbone flexibility improves protein stability estimation. Structure 15, 1567–1576 (2007)

12. Gromiha, M.M.: Prediction of protein stability upon point mutations. Biochem. Soc. Trans. 35, 1569–1573 (2007)
13. Kumar, S., Tsai, C.J., Nussinov, R.: Factors enhancing protein thermostability. Protein Eng. 13, 179–191 (2000)
14. Gromiha, M.M., Oobatake, M., Sarai, A.: Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys. Chem. 82, 51–67 (1999)
15. Kannan, N., Vishveshwara, S.: Aromatic clusters: a determinant of thermal stability of thermophilic proteins. Protein Eng. 13, 753–761 (2000)
16. Gromiha, M.M.: Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. Biophys. Chem. 91, 71–77 (2001)
17. Gromiha, M.M., Selvaraj, S.: Inter-residue interactions in protein folding and stability. Prog. Biophys. Mol. Biol. 86, 235–277 (2004)
18. Kumar, S., Tsai, C.J., Nussinov, R.: Thermodynamic differences among homologous thermophilic and mesophilic proteins. Biochemistry 40, 14152–14165 (2001)
19. Gromiha, M.M., Thomas, S., Santhosh, C.: Role of cation-pi interactions to the stability of thermophilic proteins. Prep. Biochem. Biotechnol. 32, 355–362 (2002)
20. Chakravarty, S., Varadarajan, R.: Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. Biochemistry 41, 8152–8161 (2002)
21. Ibrahim, B.S., Pattabhi, V.: Role of weak interactions in thermal stability of proteins. Biochem. Biophys. Res. Commun. 325, 1082–1089 (2004)
22. Xiao, L., Honig, B.: Electrostatic contributions to the stability of hyperthermophilic proteins. J. Mol. Biol. 289, 1435–1444 (1999)
23. Dominy, B.N., Minoux, H., Brooks, C.L.: 3rd: An electrostatic basis for the stability of thermophilic proteins. Proteins 57, 128–141 (2004)
24. Liang, H.K., Huang, C.M., Ko, M.T., Hwang, J.K.: Amino acid coupling patterns in thermophilic proteins. Proteins 59, 58–63 (2005)
25. Saraboji, K., Gromiha, M.M., Ponnuswamy, M.N.: Importance of main-chain hydrophobic free energy to the stability of thermophilic proteins. Int. J. Biol. Macromol. 35, 211–220 (2005)
26. Sadeghi, M., Naderi-Manesh, H., Zarrabi, M., Ranjbar, B.: Effective factors in thermostability of thermophilic proteins. Biophys. Chem. 119, 256–270 (2006)
27. Das, R., Gerstein, M.: The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct. Integr. Genomics 1, 76–88 (2000)
28. Fukuchi, S., Nishikawa, K.: Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J. Mol. Biol. 309, 835–843 (2001)
29. Ding, Y., Cai, Y., Zhang, G., Xu, W.: The influence of dipeptide composition on protein thermostability. FEBS Lett 569, 284–288 (2004)
30. Berezovsky, I.N., Zeldovich, K.B., Shakhnovich, E.I.: Positive and negative design in stability and thermal adaptation of natural proteins. PLoS Comput. Biol. 3, 52 (2007)
31. Gromiha, M.M., An, J., Kono, H., Oobatake, M., Uedaira, H., Sarai, A.: ProTherm: Thermodynamic Database for Proteins and Mutants. Nucleic Acids Res. 27, 286–288 (1999)
32. Bava, K.A., Gromiha, M.M., Uedaira, H., Kitajima, K., Sarai, A.: ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Res. 32, D120–D121 (2004)

33. Zhang, G., Fang, B.: Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins. Process biochemistry 41, 1792–1798 (2006)
34. Li, W., Jaroszewski, L., Godzik, A.: Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17, 282–283 (2001)
35. Holm, L., Sander, C.: Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 14, 423–429 (1998)
36. Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers, San Mateo (1993)
37. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55, 119–139 (1997)
38. Breiman, L.: Classification and regression trees. Wadsworth International Group, Belmont (1984)
39. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2005)
40. Gromiha, M.M., Suwa, M.: Discrimination of outer membrane proteins using machine learning algorithms. Proteins 63, 1031–1037 (2006)
41. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000)
42. Huang, L.T., Gromiha, M.M., Ho, S.Y.: iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. Bioinformatics 23, 1292–1293 (2007)
43. Gromiha, M.M., Suresh, M.X.: Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins 70, 1274–1279 (2008)