BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 10 2007, pages 1292–1293
doi:10.1093/bioinformatics/btm100

Structural bioinformatics

iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations

Liang-Tsung Huang[1,2], M. Michael Gromiha[3,*] and Shinn-Ying Ho[4]

[1]Department of Computer Science and Information Engineering, Ming-Dao University, Changhua 523, [2]Institute of Information Engineering and Computer Science, Feng-Chia University, Taichung 407, Taiwan, [3]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 and [4]Department of Biological Science and Technology, and Institute of Bioinformatics, National Chiao-Tung University, Hsinchu 300, Taiwan

Received on January 9, 2007; revised on February 18, 2007; accepted on March 8, 2007

Advance Access publication March 22, 2007

Associate Editor: Burkhard Rost

ABSTRACT

Summary: We have developed a web server, iPTREE-STAB for discriminating the stability of proteins (stabilizing or destabilizing) and predicting their stability changes ($\Delta\Delta G$) upon single amino acid substitutions from amino acid sequence. The discrimination and prediction are mainly based on decision tree coupled with adaptive boosting algorithm, and classification and regression tree, respectively, using three neighboring residues of the mutant site along N- and C-terminals. Our method showed an accuracy of 82% for discriminating the stabilizing and destabilizing mutants, and a correlation of 0.70 for predicting protein stability changes upon mutations.

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

Contact: michael-gromiha@aist.go.jp

Supplementary information: Dataset and other details are given.

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. 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 (Gilis and Rooman, 1996; Parthiban *et al.*, 2006), multiple regression techniques (Gromiha *et al.*, 1999b), energy functions (Guerois *et al.*, 2002), contact potentials (Khatun *et al.*, 2004), neural networks (Capriotti *et al.*, 2004), support vector machines (SVMs) (Cheng *et al.*, 2006), average assignment (Saraboji *et al.*, 2006), classification and regression tool (Huang *et al.*, 2007), etc. Further, it has been reported that the discrimination of stabilizing and destabilizing mutants is more important than its magnitude in many cases (Capriotti *et al.*, 2004). Most of these methods used the information from the 3D structures of proteins for discrimination/prediction. On the other hand, prediction accuracy using amino acid sequence is significantly lower than that with structural data.

In this work, we have developed a server (iPTREE-STAB) 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 (predicted) the stability changes with an accuracy (correlation) of 82% (0.70).

2 METHODS

In the present study, 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 (Bava *et al.*, 2004; Gromiha *et al.*, 1999a). 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.

We have implemented the server iPTREE-STAB, using decision tree (Quinlan, 1993) along with adaptive boosting algorithm (Freund and Schapire, 1997) for discriminating the stability of protein mutants, and classification and regression tree (CART) (Breiman, 1984) 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 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.

The reliability of prediction has been tested with sensitivity (TP/(TP + FN), specificity (TN/(TN + FP), accuracy and correlation coefficient obtained with *n*-fold cross-validation technique. True positives (TP) and true negatives (TN) are, respectively, the

*To whom correspondence should be addressed.

correctly identified stabilizing and destabilizing mutants. False positive (FP) and false negatives (FN) are destabilizing mutants identified as stabilizing ones and vice versa.

3 RESULTS

The accuracy, sensitivity and specificity of our method have been tested with 4-, 10- and 20-fold cross-validation procedures. The 4- 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. 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: (a) if the deleted residue is Ala and the neighboring residues contain Gln, then the predicted stability change will be negative (accuracy = 97.1%), (b) if the deleted residue is Glu and its second neighbor at N-terminal is met, the mutation stabilizes the protein (accuracy = 100%) and (c) 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). Additional rules are provided on the web.

4 SERVER DESCRIPTION

The input options for discrimination/prediction are shown in Figure 1. 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 (Fig. 2). In the case 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 (Fig. 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.

In addition, we have provided the datasets used in the present work along with the references and links to related web servers. A help page is also provided for the details to be given in the input.

Conflict of Interest: none declared.

Fig. 1. Snapshot showing the necessary items to be given as input for discrimination and prediction.

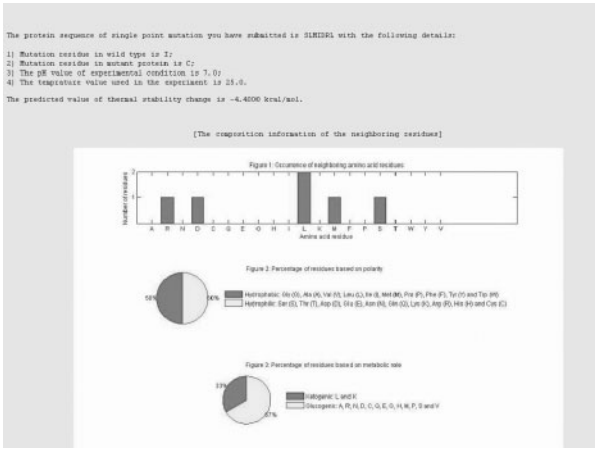

Fig. 2. The results obtained for predicting the stability change along with the related information of neighboring residues.

REFERENCES

Bava,K.A. *et al.* (2004) ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **32**, D120–D121.

Breiman,L. (1984) *Classification and regression trees*. Wadsworth International Group, Belmont, California.

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.

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

Freund,Y. and Schapire,R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci*, **55**, 119–139.

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. *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.

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.

Huang,L.-T. *et al.* (2007) Prediction of protein mutant stability using classification and regression tool. *Biophys. Chem.*, **125**, 462–470.

Khatun,J. *et al.* (2004) Can contact potentials reliably predict stability of proteins? *J. Mol. Biol.*, **336**, 1223–1238.

Parthiban,V. *et al.* (2006) CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.*, **34**, W239–W242.

Quinlan,J.R. (1993) C4.5: *Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, California.

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