Structural bioinformatics

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

Liang-Tsung Huang\textsuperscript{1,2}, M. Michael Gromiha\textsuperscript{3,*} and Shinn-Ying Ho\textsuperscript{4}

\textsuperscript{1}Department of Computer Science and Information Engineering, Ming-Dao University, Changhua 523, \textsuperscript{2}Institute of Information Engineering and Computer Science, Feng-Chia University, Taichung 407, Taiwan, \textsuperscript{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 \textsuperscript{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.

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) $M_d$, mutated (deleted) residue, (ii) $M_i$, mutant (introduced) residue, (iii) pH, (iv) $T$ ($^\circ$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
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 $-2.05\text{ kcal/mol}$ (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.

REFERENCES