Available online at www.sciencedirect.com
ScienceDirect

Computational Biology and Chemistry

Computational Biology and Chemistry 30 (2006) 408–415

www.elsevier.com/locate/compbiolchem

Knowledge acquisition and development of accurate rules for predicting protein stability changes

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

[a] *Institute of Information Engineering and Computer Science, Feng Chia University, Taichung 407, Taiwan*
[b] *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*
[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*

Received 30 March 2006; received in revised form 19 June 2006; accepted 19 June 2006

Abstract

Knowing the mechanisms by which protein stability change is one of the most important and valuable tasks in molecular biology. The conventional methods of predicting protein stability changes mainly focus on improving prediction accuracy. However, it is desirable to extract domain knowledge from large databases that is beneficial to accurate prediction of the protein stability change. This paper presents an interpretable prediction tree method (named iPTREE) that produces explanatory rules to explore hidden knowledge accompanied with high prediction accuracy and consequently analyzes the factors influencing the protein stability changes. To evaluate iPTREE and the knowledge upon protein stability changes, a thermodynamic dataset consisting of 1615 mutants led by single point mutation from ProTherm is adopted. Being as a predictor for protein stability changes, the rule-based approach can achieve a prediction accuracy of 87%, which is better than other methods based on artificial neural networks (ANN) and support vector machines (SVM). Besides, these methods lack the ability in biological knowledge discovery. The human-interpretable rules produced by iPTREE reveal that temperature is a factor of concern in predicting protein stability changes. For example, one of interpretable rules with high support is as follows: *if the introduced residue type is Alanine and temperature is between* 4 °C *and* 40 °C, *then the stability change will be negative (destabilizing)*. The present study demonstrates that iPTREE can easily be used in the application of protein stability changes where one requires more understandable knowledge.

Keywords: Protein stability; Prediction; Data mining; Decision trees; Bioinformatics

1. Introduction

Understanding the relationship between structure, function, and property of proteins is helpful to protein design that produces novel protein sequences. For this purpose, interpreting stability is a precursor and also a goal to the ability to successfully design stable proteins (Daggett and Fersht, 2003). Up to now, various methods have been proposed to predict stability changes ($\Delta\Delta G$) upon protein mutation, including energy-based methods and machine learning approaches. Energy-based methods base on force fields, which can be categorized into three major classes depending on the energy functions (Guerois et al., 2002): (a) those using physically effective energy functions (Prevost et al., 1991); (b) those based on statistical potentials for which energies are derived from the frequencies of residue contacts (Gilis and Rooman, 1997); and (c) those using empirically effective energy functions obtained from experimental data (Funahashi et al., 2001). Recently, machine learning approaches based on artificial neural network (ANN) (Capriotti et al., 2004) and support vector machines (Capriotti et al., 2005) have been proposed.

All the above-mentioned methods are concentrated on raising prediction accuracy but not accompanied with knowledge acquisition. However, only predicting protein stability is not satisfactory for the goal of understanding the relationship between structure, function, and property of proteins. Besides, because the sizes of datasets used to design predictors are often insufficiently large, the overfitting problem may be occurred resulting in a wrong model and incorrect inference. Therefore, the validation for the model and inference is necessary and

* Corresponding author. Tel.: +886 3 5131405; fax: +886 3 5729288.
E-mail address: syho@mail.nctu.edu.tw (S.-Y. Ho).

1476-9271/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2006.06.004

crucial. If the prediction model was established accompanied with human-interpretable knowledge generated, it would be more credible after confirmation. Thus, it is better to design an interpretable predictor that takes both prediction accuracy and knowledge acquisition into account simultaneously. In this study, the proposed interpretable prediction tree method (named iPTREE) aims to simultaneously achieve the following three objectives, described below.

1.1. *High prediction accuracy*

The ANN predictor (Capriotti et al., 2004) reached the accuracy as high as 81% in predicting the stability (stabilizing/destabilizing) of protein mutants (sign of $\Delta\Delta G$ values), and performs better than the existing energy-based methods in terms of prediction accuracy. However, the ANN predictor lacks the ability in biological knowledge discovery. The rule-based approach generated from iPTREE is able to successfully predict the sign of the $\Delta\Delta G$ value with accuracy 87% using a 10-fold cross-validation test, which is significantly better than the ANN-predictor using the same features and dataset. The high accuracy of prediction model will provide more confidence to the knowledge discovery derived from this model.

1.2. *Interpretable rules for knowledge acquisition*

The mechanism of systematically and actively capturing knowledge from experiment results is valuable to understanding an unknown concept. iPTREE can reveal the important factors and decision rules about protein stability changes upon mutation from a large and confused database.

At the same time, the rule base also demonstrates interpretable decision rules. One of those rules with high support is as follows:

> If the introduced residue type is Alanine and temperature is between 4 °C and 40 °C, then the stability change will be negative.

Those interpretable rules may agree with previous researches or belong to new discovery that still requires a confirmation. However, according to those interpretable conditions (temperature, introduced/deleted residue type and the environment information of the mutation position), rules can more easily be validated to be usable knowledge. In this study, although iPTREE was applied to predict protein stability changes, it can be extended to other applications and has been successfully used in prediction and analysis of DNA-binding sites of proteins (Ho et al., 2005).

1.3. *Analysis of influence factors*

From various viewpoints, several studies have similarly revealed that the positional parameters play an important role in understanding the folding and stability of protein mutants (Gilis and Rooman, 1997; Gromiha et al., 1999; Gromiha and Selvaraj, 2002; Capriotti et al., 2004). However, the comparison of relative importance between secondary structure and solvent accessibility of mutant residues from the viewpoint of predicting the stability of protein mutants has not yet been completely explored. In the recent discussion about the relative importance, the secondary structure carries similar or more information than solvent accessibility for understanding the stability of protein mutants (Saraboji et al., 2005). Through this topic, iPTREE performed the factor analysis and made a discussion.

Instead of the conventional investigation using linear correlation between one individual factor and real experiment value, iPTREE further considers interaction between the concerned factor and other pre-existing factors, namely the surrounding effect, in the factor analysis using prediction accuracy as a measurement of importance. That is to say, relationship between one feature set and real experiment value is considered. Based on the one-factor-at-once strategy for analysis of the two influence factors, secondary structure and solvent accessibility, iPTREE used the three feature sets: (1) including solvent accessibility, (2) including secondary structure, and (3) including both two, with the same surrounding effect. The statistic result $F = 0.92$ of one-way analysis of variance (ANOVA) for difference in means indicates the hypothesis: three conditions have equal means. It may result from that the environment information of the mutation position is enough to cover those from the secondary structure and solvent accessibility.

2. Materials and methods

2.1. *Protein and mutant datasets*

For comparisons, the same dataset used by (Capriotti et al., 2004) is conducted, which is obtained from the thermodynamic database for proteins and mutants (ProTherm, Gromiha et al., 2000). The dataset (S1615) consists of 1615 single point mutations obtained from 42 protein sequences. Each record of S1615 contains the following seven features:

(1) Md: deleted-residue mutation type;
(2) Mi: introduced-residue mutation type;
(3) pH: the pH value of the experimental condition;
(4) Temp: the temperature (°C) used in the experiment at which the stability of the mutated protein was measured explicitly;
(5) ASA: accessible surface area of the mutated reside computed by the DSSP program (Kabsch and Sander, 1983);
(6) N_X: the number of the encoded residue type X, which is found inside a sphere with a center on the mutated residue. Where the local spatial environment is computed using a radius 9 Å and X is an abbreviation of 1 of 20 residues;
(7) the secondary structure information centered on the mutated residue.

The main character of this dataset includes: (1) the $\Delta\Delta G$ value is experimentally detected, (2) the protein structure is known with atomic resolution, and (3) the data is based upon single mutations. All demonstrations of iPTREE are based on S1615.

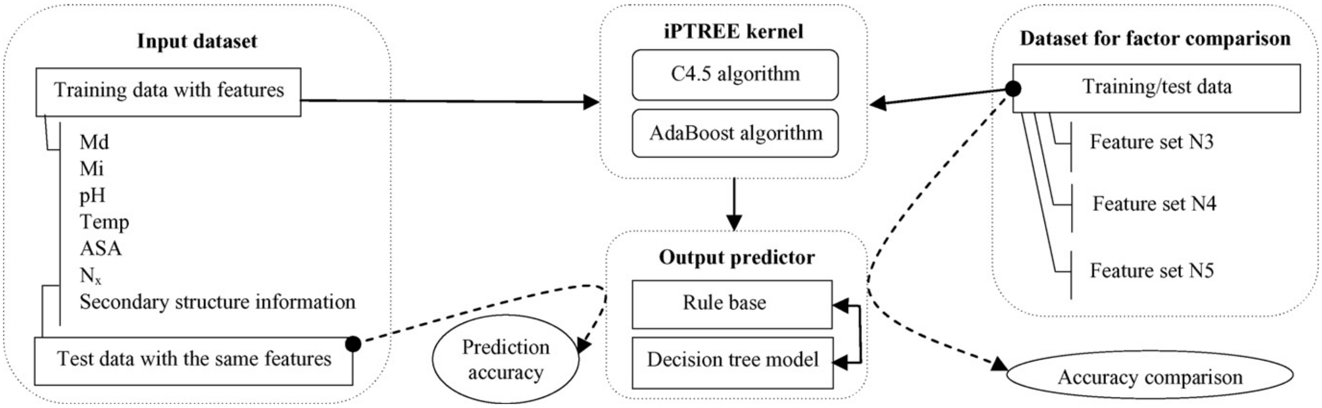

Fig. 1. The framework of iPTREE for factor analysis and protein stability change prediction by establishing an accurate rule base.

2.2. *The proposed iPTREE*

The hybrid method iPTREE is mainly based on a decision tree algorithm C4.5 and an adaptive boosting algorithm named AdaBoost. Since no backtracking strategy is used, iPTREE can efficiently operate on large-scale datasets. It is worthy to mention that it can establish a rule-based model for prediction, which is helpful to explore the hidden information in the datasets. Fig. 1 illustrates the framework of establishing a predictor by iPTREE, and of comparing factors between different feature sets for a certain database using prediction accuracy as an index.

2.2.1. *iPTREE kernel-decision tree algorithm C4.5 and AdaBoost algorithm*

The decision tree algorithm (Quinlan, 1986) predicts the value of a discrete dependent variable with a finite set from the values of a set of independent variables. A decision tree is constructed by looking for regularities in data. It examines the features to add at the next level of the tree by using an entropy calculation, and then chooses the feature which minimizes the entropy impurity (or occasionally information impurity). Once a decision tree has been constructed, it is a simple matter to convert it into an equivalent set of rules that increase the readability and understanding of data (Geurts et al., 2005). Several well-known decision tree algorithms are available. Here, the decision tree algorithm used is the well-known C4.5 algorithm (Quinlan, 1993), and gain ratio criterion is adopted as entropy calculation which is the ratio of information gain over potential information. C4.5 is based on a nonparametric type of regression fitting approach, which is suitable for unknown data distribution. Another advantage is that C4.5 deals effectively with large datasets and the issues of higher dimensionality, such as protein mutant data.

Overfitting is a significant practical difficulty for many learning methods. One approach to avoiding overfitting in decision tree learning is tree pruning. The parameter CL of confidence level used to prune the decision tree affects both tree size and accuracy rate, which can be adaptively tuned to avoid overfitting. The appropriate value of CL is problem-dependent (see Section 3.1). Generally, the smaller value of CL results in a smaller tree but lower training accuracy.

The idea of the adaptive boosting algorithm (Freund and Schapire, 1997), or AdaBoost algorithm, is to improve the classification process by generating a number of classifiers from the data, each optimized to classify correctly the cases most obviously misclassified on the previous pass. Due to exploitation of groups of hypotheses with independent errors, the main advantage of AdaBoost is to increases the overall accuracy of the classification and to reduce both the variance and the bias of the classification. The parameter TR of trail in AdaBoost algorithm controls the total number of classifiers where the proper value of TR is also problem-dependent (see Section 3.2). Naturally, constructing multiple classifiers requires more computation than building a single classifier.

In this study, iPTREE uses both C4.5 and AdaBoost algorithm which are fully cooperated to be the kernel of iPTREE. And parameters CL and TR of iPTREE can be adaptively tuned for advancing prediction performance.

2.2.2. *Input feature sets and output predictor*

Based on the dataset S1615, iPTREE can adopt all seven features as input feature set by default. However, in order to demonstrate iPTREE in several different aspects, five various combinations of features were considered. For comparison with previous ANN predictor (see Section 3.3), the same feature sets are required. Thus, three feature sets named F1, F2 and F3 were used:

(1) F1 consists of Md, Mi, pH and Temp;
(2) F2 adds ASA to F1;
(3) F3 adds additional N_X to F2.

On the other hand, for factor comparison of solvent accessibility and secondary structure (see Section 3.5), the following three feature sets named F3, F4 and F5 were used:

(1) F3 contains the solvent accessibility but not the secondary structure;
(2) F4 replaces the solvent accessibility of F3 with the secondary structure;
(3) F5 adds the secondary structure to F3.

There are five common features considered in F3, F4 and F5. Namely, the surrounding effect is taken into account.

Both types of symbolic and numeric features can be handled by iPTREE. iPTREE is able to directly deal with the symbol features such as residue type (e.g. glycine, alanine, etc.), but some other predictors such as the neural network must transform the symbol to a numerical value such as 11001001. This indirect process may change the degree of freedom which leads to another problem. This ability of processing different feature types gives the system more flexibility to varied applications.

iPTREE produces a decision tree model for prediction of the stability change direction (increasing or decreasing) as well as an interpretable rule base from the tree model for the purpose of data mining. If necessary, the predictor can be evaluated by test data assigned. iPTREE relies on a greedy search which iteratively selects the candidate that maximizes a heuristic splitting criterion from the feature set. The selection order will expose the contribution of features to predict stability changes. Besides, decision rules can be constructed from a decision tree straightforward by traversing any given path from the root to any leaf. Those interpretable type of knowledge can be validated by biochemistry experts. By contrast, investigators, who want to get more information and analyze the meaning from neural network would meet a setback, since the analysis of interaction relationship between neurons of ANN is rather difficult. Section 3.4 shows the knowledge acquisition of protein stability changes in which important factors and decision rules are extracted and discussed.

The present method was validated by both self-consistency and 10-fold cross-validation tests. The latter was adopted when comparing prediction performance; and the former was applied to what focuses on the analysis of the existing dataset. Self-consistency includes all the stability data for training the decision tree model and prediction has been made for all the mutants. The 10-fold cross-validation partitions samples into 10 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.

2.2.3. Factor comparison based on iPTREE

iPTREE can establish a predictor for the direction of protein stability change by generating a rule base. Moreover, by taking all factors in a feature set into account, the accuracy by the way of iPTREE can be regarded as a comprehensive index for importance evaluation of factors using one-factor-at-once strategy. Therefore, by observing prediction accuracy based on different feature sets, relative influence of factors can be comparison. On this basis, Section 3.5 centers on the discussion about relative importance of secondary structure and solvent accessibility to the stability of protein mutants. In order to compare the two factors, three designed feature sets F3, F4 and F5 (see Section 2.2.2) were adopted with the same surrounding effects. Then the effects of two factors can be discussed by comparing the accuracy based on three feature sets (also see Fig. 1).

The prediction of the stability change (stabilizing/destabilizing) can be regarded as one of binary classification problems in which several scoring functions are usually used. The prediction accuracy AC for the two-class classification, positive and negative of the $\Delta\Delta G$ value, is as follows:

$$\mathrm{AC} = \left(\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}\right) \times 100\%, \tag{1}$$

where TP and TN are the total numbers of correctly predicted samples for positive and negative classes, respectively; FP and FN are the total numbers of incorrectly predicted samples for positive and negative classes, respectively. The sensitivity SE is

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100\%; \tag{2}$$

the specificity SP is

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \times 100\%. \tag{3}$$

Those scoring functions mean whether the values between prediction and experiment value can fit well. Namely, the higher accuracy indicates the dataset with certain of feature set can make a larger contribution to prediction model.

3. Results and discussions

3.1. Confidence level effects of C4.5 algorithm

In order to avoid overfitting in decision tree learning, we manipulated confidence level that effects tree pruning of C4.5. Whereas the appropriate confidence level is problem-dependent, a preliminary analysis was applied using the following simulations. iPTREE was applied to S1615 with feature set F3 and based on 10-fold cross-validation test. The computing platform is Intel Celeron processor 2.4 GHz with 768 MB RAM running Microsoft Windows XP.

Table 1 shows the prediction results using various values of confidence level. The insignificant improvement between the highest and lowest prediction accuracy (AC = 80.6% and 78.1%) reveals that the effects of confidence level are moderate. We noticed that all the specificity values are higher than the sensitivity values. This may be due to unequal sample sizes between positive and negative class (449 and 1166) which affects the

Table 1
Prediction performance of various CL values

CL	AC (%)	SE (%)	SP (%)	Number of rules	Time (s)
10	78.5	67.0	81.0	17.9	0.6
15	78.5	66.8	81.2	27.6	0.6
20	78.1	67.1	80.4	36.0	0.6
25	79.9	69.8	82.4	46.2	0.6
30	80.6	68.2	84.2	79.9	0.6
40	79.1	64.4	83.8	124.9	0.7
50	78.6	62.5	84.1	138.7	0.6
60	79.6	64.5	84.7	166.3	0.7
70	80.4	65.1	86.0	197.0	0.7
80	79.9	64.0	86.0	237.8	0.6
90	80.1	64.1	86.4	256.0	0.7
100	80.1	64.0	86.5	275.5	0.7
Mean	79.5	65.6	83.9	133.7	0.6

Table 2
Performance comparison of various TR values using CL = 30 and 70

CL	TR	AC (%)	SE (%)	SP (%)	Number of rules	Time (s)
30	10	84.0	78.4	85.4	65.3	7.5
	20	84.0	78.4	85.4	30.5	15.0
	30	86.3	82.0	87.5	76.1	22.6
	40	85.4	82.8	86.0	94.3	28.4
	50	85.6	82.5	86.4	90.0	36.0
	100	86.0	82.5	86.9	90.3	70.3
	200	85.9	82.1	87.0	71.6	134.0
	300	85.6	82.0	86.6	78.9	190.3
	400	87.1	83.6	88.1	98.8	295.1
	500	86.3	83.9	86.9	67.4	308.0
	1000	85.3	85.5	85.5	88.2	660.8
Mean		85.6	82.2	86.5	77.4	160.7
70	10	85.5	79.6	87.2	122.4	7.5
	20	87.0	82.2	88.4	155.3	15.0
	30	86.0	81.1	87.4	123.8	23.5
	40	87.0	81.2	88.8	151.7	32.3
	50	86.1	80.6	87.7	136.5	38.2
	100	87.0	82.0	88.5	121.9	80.7
	200	86.7	81.8	88.1	125.0	159.9
	300	86.5	80.8	88.2	141.2	234.0
	400	86.4	81.1	88.0	147.5	318.9
	500	86.1	81.5	87.4	136.5	387.9
	1000	86.5	81.1	88.1	153.4	782.9
Mean		86.4	81.2	88.0	137.7	189.2

entropy calculation. The confidence level and number of rules are in direct proportion, which confirms to the mechanism of C4.5 algorithm. Although the number of rules varied from 17.9 to 275.5, the execution time taking 0.7 second at most shows iPTREE can effectively be applied.

3.2. Trail effects of AdaBoost algorithm

iPTREE combines C4.5 with AdaBoost algorithm in which the trail concerns with the number of classifiers. To observe the trail effects, various TR values are used at the same confidence level (CL = 30) with the highest accuracy as above. In Table 2, the mean sensitivity value which increases from 65.6% up to 82.2% reveals the prediction ability for positive class has advanced greatly after introducing AdaBoost algorithm. Consequently, the mean accuracy value is also improved from 79.5% up to 85.6%. The execution time which ranges from 7.5 s to 660.8 s is proportional to TR since AdaBoost needs more time to construct decision trees for voting.

Besides the confidence level with the highest accuracy, we used the confidence level (CL = 70) with the next highest accuracy in Table 1. The results shows that the prediction accuracy values are 85.6% and 86.4% and the standard deviation are 0.93% and 0.48% for CL = 30 and 70, respectively (also see Table 2). It seems that the difference of accuracy between two different confidence levels is insignificant. To make sure this observation, a paired *t*-test was performed. According *t*-test for paired difference in means at an α of 0.05, *p*-value = 0.011 indicates that the hypothesis: the means of accuracy rate are equal between two sets with confidence level value of 70 and 30, respectively, cannot be rejected for single-tailed test.

Apparently, by considering the effects of confidence level and trail, the best accuracy (AC = 87.1%) occurs when CL = 30 and TR = 400. Meanwhile, results also show that the parameter TR is more effective than CL in prediction accuracy of S1615 with F3.

3.3. Comparison between iPTREE and ANN predictors

S1615 dataset was previously used by a neural-network-based method which introduces three different input feature sets to generate corresponding prediction models. For understanding the performance difference between two predictors, the same feature sets F1, F2 and F3 (including 4, 5 and 6 features, respectively) were used in iPTREE with the parameter found in Section 3.2.

In Table 3, we observe that all the prediction performance was improved by the addition of new feature. It indicates that the added feature contains valuable information about stability prediction. Based on F3, the accuracy value 87.1% of iPTREE is better than 81.0% of ANN. A major cause is the improvement in prediction performance of positive class, where the sensitivity value increase from 71.0% to 83.6%. The comparison shows the high prediction performance of iPTREE.

For analysis of prediction ability in different ranges of solvent accessibility, Fig. 2 shows prediction accuracy as a function of solvent accessibility value of the mutated residue. The group ranges of solvent accessibility value begin with ASA = 0, then 0 < ASA ≤ 10 and so forth. Darker bars represent the accuracy values and lighter bars are the number of data in each category. We observed that the accuracy values of each group are almost equal. It reveals that the prediction ability of iPTREE can work for a comprehensive range of solvent accessibility. In other words, it is unrestricted to the previous observation (Zhou and Zhou, 2004), showing the prediction ability of system may be affect by high exposed residue mutation (high solvent accessibility value).

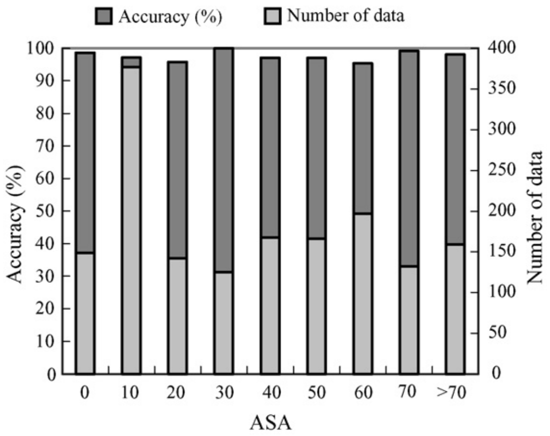

Fig. 2. The prediction accuracy as a function of the relative accessibility value of the mutated residue.

Table 3
Performance comparison between the two predictors of iPTREE and ANN using various feature sets

Feature set	Number of features	iPTREE			ANN		
		AC (%)	SE (%)	SP (%)	AC (%)	SE (%)	SP (%)
F1	4	77.3	66.3	79.4	74.0	59.0	76.0
F2	5	81.2	67.8	85.7	75.0	57.0	80.0
F3	6	87.1	83.6	88.1	81.0	71.0	83.0
Mean		81.9	72.6	84.4	76.7	62.3	79.7

Table 4
The factors with high gain ratio and corresponding splitting values

Tree level	Factors	Splitting value	Potential information	Information gain	Gain ratio
1	Temp	42.5	1.000	0.048	0.048
2	N_{Y}	2.5	0.792	0.019	0.024
2	N_{R}	2.5	0.892	0.058	0.065
3	Mi	–	3.710	0.075	0.020
3	Md	–	3.675	0.121	0.033
3	N_{L}	3.5	0.918	0.049	0.054
3	N_{H}	0.5	0.697	0.046	0.065

3.4. Mining important factors and decision rules

We have applied iPTREE to observe the selected candidate factors from S1615 dataset with F3 and the extracted decision rules with high accuracy. Table 4 shows the selected factors with high gain ratio at the top 3 tree levels and corresponding split values. The first important factor obtained is the temperature with splitting value 42.5 and gain ratio 0.048. Namely, the temperature factor made a major contribution to the distinct ability of predicting protein stability changes. In previous related researches, the free energy (ΔG) can be regarded as a function of temperature (Robertson and Murphy, 1997), which shows the temperature is one of important factors for the free energy. From molecular dynamics simulations, the mutant trajectory was observed to be much less stable than for the wild-type protein trajectory at normal and elevated temperature (el-Bastawissy et al., 2001). However, this contention still needs more evidences to support it. On the other, the Mi and Md are also considered as important factors appearing at level 3. Being with discrete (−) values means that more than two splits may be generated.

Tables 5 and 6 list the best seven antecedents of decision rules generated from iPTREE for negative and positive sign of stability change, respectively. For convenience, all factors are substituted for symbols described in Section 2.1. The rule size means the length of antecedent sentence and the support of one decision rule refers to the number of samples to which the rule applies in the dataset. For example, the first rule antecedent in Table 5: If Temp > 42 and $N_{\mathrm{M}} > 1$ and $N_{\mathrm{P}} > 0$, explores the information:

> If temperature is larger than 42 °C, and Methionine appears above two times, and Proline appears, then the predicted stability change will be negative.

The accuracy of 100% with support of 52 means that total 52 samples fit this rule and their stability changes are predicted exactly in the whole dataset. For the first rule antecedent in Table 6: If Mi = M and ASA > 15.92 and N_{V} < =1, explores the information:

> If introduced residue is Methionine and tend toward exposed mutation, and the Valine appear one time or not, then the predicted stability change will be positive.

The accuracy of 100% with support of 10 means that total 10 samples fit this rule and their stability changes are predicted exactly in the whole dataset.

Table 5
Antecedents of decision rules with high accuracy and corresponding details for negative sign of stability change (destabilizing)

Antecedent	Rule size	AC (%)	Number of data
If Temp > 42 & $N_{\mathrm{M}} > 1$ & $N_{\mathrm{P}} > 0$	3.0	100.0	52.0
If Temp > 42 & $N_{\mathrm{G}} > 2$ & $N_{\mathrm{G}} \leq 3$ & $N_{\mathrm{K}} \leq 0$	4.0	100.0	17.0
If Mi = F & $N_{\mathrm{I}} \leq 3$ & $N_{\mathrm{K}} \leq 0$	3.0	100.0	9.0
If Mi = T & Temp ≤ 42	2.0	97.4	39.0
If Temp > 42 & $N_{\mathrm{A}} > 0$ & $\mathrm{N_F} > 1$ & $N_{\mathrm{G}} \leq 3$ & $\mathrm{N_L} \leq 3$ & $\mathrm{N_Q} \leq 1$ & V ≤ 3	7.0	96.9	32.0
If Temp ≤ 42 & $N_{\mathrm{Y}} > 2$	2.0	94.4	195.0
If Mi = A & Temp > 4 & Temp ≤ 40	3.0	89.7	232.0
Mean	3.4	96.9	82.3

Table 6
Antecedents of decision rules with high accuracy and corresponding details for positive sign of stability change (stabilizing)

Antecedent	Rule size	AC (%)	Number of data
If Mi = M & ASA > 15.92 & $N_V \leq 1$	3.0	100.0	10.0
If Temp > 42 & $N_I \leq 3$ & $N_K \leq 1$ & $N_L \leq 3$ & $N_P \leq 2$ & $N_R \leq 2$ & $N_V > 3$	7.0	100.0	9.0
If Temp > 42 & $N_G \leq 1$ & $N_L > 3$ & $N_M > 1$ & $N_P \leq 0$ & $N_R \leq 2$	6.0	100.0	8.0
If Mi = D & $N_A \leq 0$ & $N_Y \leq 2$	3.0	100.0	7.0
If Mi = A & pH ≤ 6 & Temp ≤ 4	3.0	100.0	6.0
If Temp > 42 & $N_E > 3$ & $N_L > 3$ & $N_M > 1$ & $N_R \leq 2$	5.0	100.0	6.0
If Temp > 42 & $N_G > 3$ & $N_K \leq 1$ & $N_L \leq 3$ & $N_R \leq 2$	5.0	91.9	37.0
Mean	4.6	98.8	11.9

Comparing Tables 5 and 6, the mean rule size values are 3.4 and 4.6 for negative and positive sign, respectively; the mean support values are 82.3 and 11.9, respectively. It seems that the rules for negative sign are simple and dominate more cases due to the smaller size and higher support. On the other hand, the prediction accuracy values are 96.9% and 98.8% for negative and positive sign, respectively. The 14 listed rules apply to total 659 samples (out of 1615), where negative and positive ones are 576 and 83, respectively. It shows that few rules can predict most of cases with high accuracy, namely those rules are worthy to pay attention.

Whereas the highest support may serve a general phenomenon, we focus on the seventh rule (with support of 232) in Table 5:

> If introduced residue is Alanine and temperature is between 4 °C and 40 °C, then the predicted relative stability change will be negative.

Several previous researches have revealed some properties about Alanine in protein stability. Early, Val to Ala mutations within the 50-residue major coat (gene VM) protein of bacteriophage M13 has been studied (Deber et al., 1993). It emphasized that those Val to Ala mutations enhance protein dimer stability in the M13 system. Recently, The proline-free triple mutant P7A/P9A/P50A was investigated using Fourier-transform infrared (FTIR) spectroscopy (Zscherp et al., 2003). The thermal stability of the proline-free mutant is reduced by 15 °C as compared to the wild type. Also, the impact of single cysteine residue mutations on the replication terminator protein (RTP) was reported (Vivian et al., 2003). The thermal unfolding temperatures (T_m) were calculated from thermal unfolding curves derived for the wild-type and mutant RTP. The RTP.C110A mutant with 55.8 °C possesses the lowest stability of the RTP molecules.

3.5. Factor comparison of solvent accessibility and secondary structure

To exhibit the factor analysis based on iPTREE method, this section focuses on the discussion about relative importance between two factors, solvent accessibility and secondary structure, to the stability of protein mutants. Through dataset S1615, we designed three feature sets which contain five common features, namely the surrounding effect was included. F3 and F4 can, respectively, be regarded as the solvent accessibility and the secondary structure factors working under the surrounding effect, and F5 indicates both factors exist with the same surrounding effect.

Table 7 shows the results based on three feature sets with various TR values. It reveals the performance among three sets is great but not significantly different. To make sure this observation, a one-way ANOVA was applied. In this case, the null hypothesis is that the average accuracy values of three feature sets are equal. And the alternative hypothesis is the average

Table 7
Performance comparison among three feature sets using various TR values

TR	F3			F4			F5		
	AC (%)	SE (%)	SP (%)	AC (%)	SE (%)	SP (%)	AC (%)	SE (%)	SP (%)
10	85.5	79.6	87.2	85.8	79.2	87.7	85.2	77.9	87.4
20	87.0	82.2	88.4	86.4	80.1	88.3	85.4	79.7	87.0
30	86.0	81.1	87.4	86.9	83.2	88.0	86.5	81.5	88.0
40	87.0	81.2	88.8	86.4	81.1	88.0	86.6	80.4	88.5
50	86.1	80.6	87.7	87.0	82.7	88.2	87.0	82.2	88.4
100	87.0	82.0	88.5	87.2	83.3	88.4	86.7	81.8	88.2
200	86.7	81.8	88.1	87.0	82.6	88.3	86.4	81.0	88.0
300	86.5	80.8	88.2	87.1	81.6	88.7	87.7	84.3	88.6
400	86.4	81.1	88.0	87.2	82.9	88.5	86.9	81.6	88.5
500	86.1	81.5	87.4	86.6	81.5	88.1	86.4	81.8	87.8
1000	86.5	81.1	88.1	86.6	82.2	87.8	86.9	81.8	88.4
Mean	86.4	81.2	88.0	86.7	81.9	88.2	86.5	81.3	88.1

accuracy of these is unequal. According to one-way ANOVA for difference in means at $\alpha = 0.05$, statistics $F = 0.92$ (<3.12) means that the hypothesis: three conditions have equal means, cannot be rejected for two-tailed test. Namely, the solvent accessibility or the secondary structure information provided similar prediction accuracy. A reasonable inference is that the information of local spatial environment (feature N_X) dominates both features under the same circumstance.

4. Conclusions

In this paper, the proposed iPTREE was effectively applied to establish an accurate rule base from the thermodynamic database of proteins and mutants. On the framework of iPTREE, potential knowledge of protein stability prediction can be extracted and transform to interpretable rules which can help the further validation by biochemistry experts. Meanwhile, the importance of factors effecting protein stability changes can be compared by the prediction accuracy served as a comprehensive index.

In addition, since the expression of one rule sentence is composed of several key words of features and relation operators. In other words, a set of significant features can directly help to establish more interpretable rules. Besides, a set of effective features also can improve the prediction performance, namely, the correctness of the rule base. For this reason, selecting appropriate features will be a worthy issue in future researches.

Even though there is something to work on, we have showed that the method is relatively available, readable and fast to explore the knowledge of predicting protein stability changes from a large database. And the knowledge can provide us more understanding about the protein stability change.

References

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

Capriotti, E., Fariselli, P., Casadio, R., 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 33 (Web Server issue), W306–W310.

Daggett, V., Fersht, A.R., 2003. Is there a unifying mechanism for protein folding? Trends Biochem. Sci. 28, 18–25.

Deber, C.M., Khan, A.R., Li, Z., Joensson, C., Glibowicka, M., Wang, J., 1993. Val → Ala mutations selectively alter helix–helix packing in the transmembrane segment of phage M13 coat protein. Proc. Natl. Acad. Sci. U.S.A. 90 (24), 11648–11652.

el-Bastawissy, E., Knaggs, M.H., Gilbert, I.H., 2001. Molecular dynamics simulations of wild-type and point mutation human prion protein at normal and elevated temperature. J. Mol. Graph. Model. 20 (2), 145–154.

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

Funahashi, J., Takano, K., Yutani, K., 2001. Are the parameters of various stabilization factors estimated from mutant human lysozymes compatible with other proteins? Protein Eng. 14, 127–134.

Geurts, P., Fillet, M., de Seny, D., Meuwis, M.A., Malaise, M., Merville, M.P., Wehenkel, L., 2005. Proteomic mass spectra classification using decision tree based ensemble methods. Bioinformatics 21 (14), 3138–3145.

Gilis, D., Rooman, M., 1997. 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, 276–290.

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

Gromiha, M.M., An, J., Kono, H., Oobatake, M., Uedaira, H., Prabakaran, P., Sarai, A., 2000. ProTherm, version 2.0: thermodynamic database for proteins and mutants. Nucleic Acids Res. 28, 283–285.

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

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

Ho, S.-Y., Chang, C.-Y., Huang, L.-T., Huang, W.-L., Hwang, S.-F., 2005. Accurate prediction and analysis of DNA-binding proteins using a rule-based decision tree system. In: Proceedings of 12th International Conference on BioMedical Engineering: ICBME 2005, Singapore, p. 108.

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

Prevost, M., Wodak, S.J., Tidor, B., Karplus, M., 1991. Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the Ile-96-Ala mutation in barnase. Proc. Natl. Acad. Sci. U.S.A. 88, 10880–10884.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learn. 1 (1), 81–106.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco.

Robertson, A.D., Murphy, K.P., 1997. Protein structure and the energetics of protein stability. Chem. Rev. 97, 1251–1267.

Saraboji, K., Gromiha, M.M., Ponnuswamy, M.N., 2005. 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 (1), 25–35.

Vivian, J.P., Hastings, A.F., Duggin, I.G., Wake, R.G., Wilce, M.C., Wilce, J.A., 2003. The impact of single cysteine residue mutations on the replication terminator protein. Biochem. Biophys. Res. Commun. 310 (4), 1096–1103.

Zhou, H., Zhou, Y., 2004. Quantifying the effect of burial of amino acid residues on protein stability. Proteins 54, 315–322.

Zscherp, C., Aygun, H., Engels, J.W., Mantele, W., 2003. Effect of proline to alanine mutation on the thermal stability of the all-beta-sheet protein tendamistat. Biochim. Biophys. Acta 1651 (1–2), 139–145.