Expert Systems with Applications 36 (2009) 9007–9014

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

An expert system to predict protein thermostability using decision tree

Li-Cheng Wu [a], Jian-Xin Lee [b], Hsien-Da Huang [c], Baw-Juine Liu [d], Jorng-Tzong Horng [a,b,e,*]

[a] Institute of System Biology and Bioinformatics, National Central University, Taiwan
[b] Department of Computer Science and Information Engineering, National Central University, Taiwan
[c] Institute of Bioinformatics, National Chiao-Tung University, Taiwan
[d] Computer Science and Information Engineering, Yuan Ze University, Taiwan
[e] Department of Bioinformatics, Asia University

ARTICLE INFO

Keywords:
Expert system
Machine learning
Bioinformatics
Protein thermostability
Decision Tree

ABSTRACT

Protein thermostability information is closely linked to commercial production of many biomaterials. Recent developments have shown that amino acid composition, special sequence patterns and hydrogen bonds, disulfide bonds, salt bridges and so on are of considerable importance to thermostability. In this study, we present a system to integrate these various factors that predict protein thermostability. In this study, the features of proteins in the PGTdb are analyzed. We consider both structure and sequence features and correlation coefficients are incorporated into the feature selection algorithm. Machine learning algorithms are then used to develop identification systems and performances between the different algorithms are compared. In this research, two features, (E + F + M + R)/residue and charged/non-charged, are found to be critical to the thermostability of proteins. Although the sequence and structural models achieve a higher accuracy, sequence-only models provides sufficient accuracy for sequence-only thermostability prediction.

1. Introduction

Protein stability is intimately connected with protein folding and all proteins have to be folded into their final active state to be active and stable. In biotechnology, chemical reactions need to be performed at high temperatures to decrease reaction time. However, many proteins are not very stable when heated. Research is needed that helps proteins to remain active and stable at high temperatures, which will overcome many limitations to their industrial applications (Huang et al., 2004). There are four thermostability classes of protein, psychrophilic, mesophilic, thermophilic and hyperthermophilic, separated by denaturation temperature of 20 °C, 50 °C and 80 °C (Vieille & Zeikus, 2001). Current research has shown that the properties of proteins such as number of ion pairs, and salt bridges are considerably related to protein thermostability (Vieille & Zeikus, 2001). Research has shown that single point mutation, when related to thermostability, may allow thermostability prediction (Capriotti, Fariselli, & Casadio, 2004; Capriotti, Fariselli, & Casadio, 2005; Saraboji, Gromiha, & Ponnuswamy, 2006). Instead of analyzing mutant proteins, research is needed that investigates the global relationships between thermostability, protein structure and sequence properties.

Recent investigations have shown that features such as amino acid composition and ion pairs are related to the thermostability. However, a general model of how these features are related to thermostability remains a scientific challenge. Our previous work (Huang et al., 2004) has identified a general model relating thermostability and ion pairs. Since structured folding is a complex question, practically a model of thermostability based on sequence information would be more useful than one based on other types of analysis.

Our goal is to develop a predictive model for thermostability based on sequence and structural features. Given a protein sequence/structure, the system aims to classify proteins into three different thermostability classes: mesophilic, thermophilic and hyperthermophilic. We aim to achieve high accuracy, sensitivity, specificity and precision. Higher accuracy may be available if both specific protein structure and sequence information are available, but a general model that will accept sequence-only information is also constructed.

2. Related works

The optimal growth temperature of an organism indicates the temperature at which the organism's growth is most rapid. The prokaryotic Growth Temperature Database (PGTdb[1]) is a database of prokaryotic thermostability information. We adapted the thermostability information from the PGTdb for this work. The PDB[2] is the

* Corresponding author. Address: Department of Computer Science, National Central University, Taiwan.
E-mail address: horng@db.csie.ncu.edu.tw (J.-T. Horng).

[1] http://pgtdb.csie.ncu.edu.tw/.
[2] http://www.rcsb.org/pdb/.

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.12.020

single worldwide repository for the processing and distribution of 3-D structure data of large molecules including proteins and nucleic acids (Berman et al., 2000). Protein sequence and structure information are collected through the PDB database.

Prediction of protein thermostability based on single mutation points has shown that prediction of thermostability is possible (Capriotti et al., 2004, 2005; Farias, van der Linden, Rego, Araujo, & Bonato, 2004; Saraboji et al., 2006). Previous research (Haney, Stees, & Konisky, 1999; Vieille & Zeikus, 2001) has shown that many special properties of proteins are related to protein thermostability. Usually, protein stability is intimately connected with the protein three-dimensional structure. It is believed that sequence determines the final folding of protein although three-dimensional free folding of an arbitrary protein sequence is still a scientific challenge.

The thermostability of proteins is directly connected to its structural stability. Thus, the features maintaining the structure folding are often considered as significant factors in protein thermal stability. It has been demonstrated that there is a relationship between thermostability and sequence information. Amino acid composition and intrinsic propensity are thought to play important roles in thermal stability (Vieille & Zeikus, 2001). For instance, the protein's content of hydrophobic amino acids is reasonably related to its thermal stability. Many mutant rules, such as Trp → Tyr, Cys → Ile, have been proposed to enhance thermostability (Gromiha, Oobatake, & Sarai, 1999). In addition to the amino acid composition, research has also found that there is a relationship between sequence pattern and optimal growth temperature. The pattern [*EdH*] and [*EdT*] are favored by mesophilic proteins (Liang, Huang, Ko, & Hwang, 2005). Sometimes sequence analysis can focus on a special pattern and its corresponding structure. The sequence pattern is transformed into a favored secondary structure by the database records giving a good linear relationship between local structure and melting temperature (Chan et al., 2004).

There are various structure features that show a relationship with thermostability. Secondary structure, such as α-helices, and intrahelical interactions within the protein have been analyzed and it has been concluded that high α-helical stability is necessary for protein thermostability (Petukhov, Kil, Kuramitsu, & Lanzov, 1997). Hydrogen bonds, ion pairs, hydrophobic interactions and disulfide bonds are thought to be the major forces affecting protein tertiary structure. Hydrogen bonds are non-covalent bonds between donor and acceptor atoms and have been widely discussed in protein structure research. Some studies have indicated that an increase in hydrogen bonds contributes to thermostability (Vogt, Woell, & Argos, 1997), and that side-chain–side-chain hydrogen bonding plays the major role in this (Ragone, 2001). Ion pairs are another feature that has been widely discussed with respect to thermostability. Researchers have found that −1.1 ions pairs are lost for every 10 °C fall in thermostability per subunit (Gianese, Bossa, & Pascarella, 2002). When the distance between the ion pair is subdivided into <4 Å, 4–6 Å and 6–8 Å, it was found that the 6–8 Å group seem to play a more significant role in thermostability than the other two (Szilagyi & Zavodszky, 2000). Other possible properties, such as a hydrophobic core (Haney et al., 1999), electrostatic interactions and dielectric response (Dominy, Minoux, & Brooks, 2004), aromatic clusters on protein surface (Kannan & Vishveshwara, 2000) and *B* values reported in high-resolution X-ray crystal structure (Parthasarathy & Murthy, 2000), are all considered to be related to thermostability. An important restriction is that high-resolution and a high-quality structure are needed to calculate such tertiary structure properties. Researchers usually utilize homology tools to produce theoretical protein structures when the real structure is unknown, and then compare those properties across the different thermostability classes.

In addition to structural and sequence features, combination of features may also act as a key to thermostability (Farias et al., 2004). Research (Farias et al., 2004) has shown that the ratio between glutamic acid plus lysine and glutamine plus histidine gives a higher thermostability. This means that combination of amino acid (E + K)/(Q + H) may be important to protein thermostability.

The above research shows that almost all properties of a protein are related to the thermostability based on observations that compare homologous protein pairs or several protein mutants. A global view of how these features are related to thermostability is needed. In this study, we incorporate several machine learning approach such as Naïve Bayes (Huang et al., 2004), SVM and neural network together with k-NN (Baumgartner et al., 2004) into an investigation of the relationship between protein features and thermostability.

3. Materials and methods

3.1. System flow

We have developed a data-mining system to analysis the relationship between protein features and thermostability. The system builds a model based on given protein features and thermostability information. The system then takes the inputted protein features and tries to predict the thermostability of a given protein as one of three classes, mesophilic, thermophilic and hyperthermophilic. The system incorporates feature selection to eliminate low-contributing features and incorporates several machine learning approaches. Fig. 1 shows the system flow.

3.2. Materials

3.2.1. Sampling data

The optimal growth temperature of an organism indicates the temperature at which the organism's growth is most rapid. Research on the source organism's optimal growth temperature can be used in place of experimental thermostability data when comparing proteins. Most proteins of an organism are reasonably active and quite stable at the optimal growth temperature of the organism.

Two databases, PGTdb and PDB, are used in our research. We identified 5487 proteins for which there are both optimal growth temperature information in PGTdb and structural information in PDB. Fig. 2 show the data distribution for these proteins with respect to the temperature.

In total, there are 41 different optimal temperatures for the proteins. Fig. 2 shows that the number of proteins with an optimal temperature of 37 °C is far more than others. Most studied prokaryotic organisms are mesophilic and therefore over-sampling at this temperature in the PGTdb is not unexpected. As a result, the distribution is unbalanced and therefore any prediction will focus on proteins with an optimal temperature of 37 °C, In other words, a prediction model that predicts every protein to have an optimum temperature of 37 °C will achieve high accuracy, but will fail to reach the goal of linking features and thermostability. In order to balance the temperature difference sample size, we carried out a *Z*-test to pick the temperatures that contain significantly more proteins then other temperatures. There were seven temperatures that exceeded the *Z*-test limit, which indicated a barrier of 125 proteins. We select 125 proteins randomly by computer from each of these seven temperatures to balance the data. As a result, 1810 proteins from different temperatures form our sample dataset. The dataset consists of 878 mesophilic, 580 thermophilic and 352 hyperthermophilic proteins.

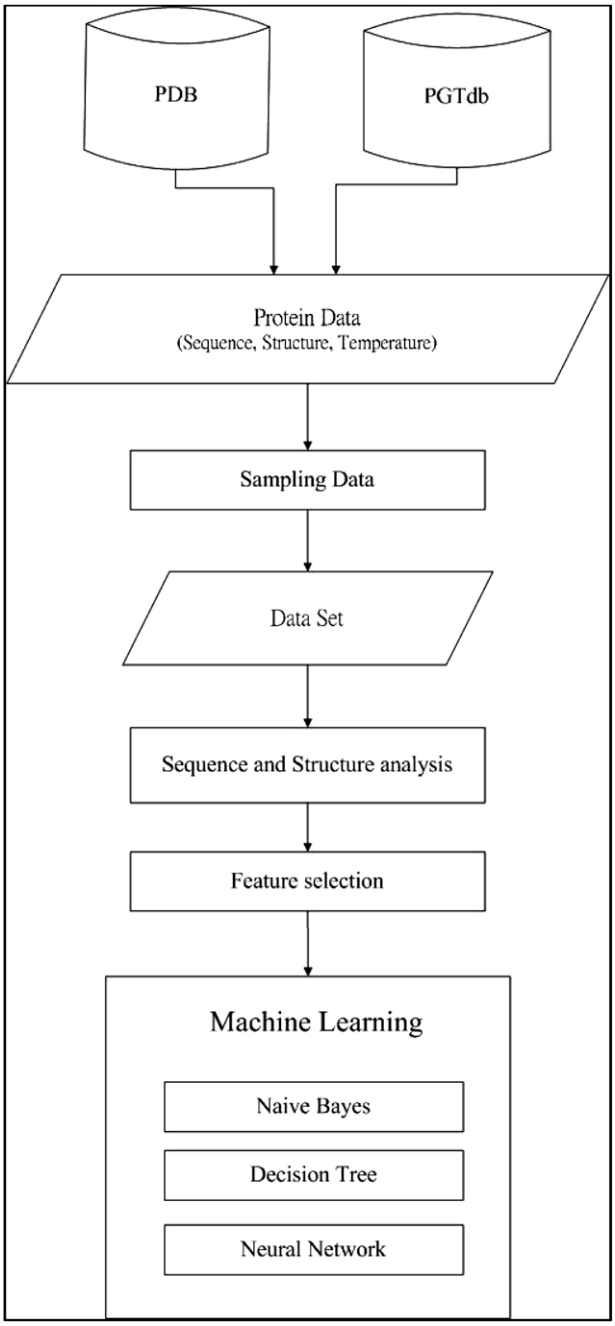

Fig. 1. System flow.

3.2.2. Candidate features

Several candidate features are considered to be globally related to the thermostability of proteins. We categorize these features into four categories as follows.

(A) *Primary structure:* amino acid composition of the protein sequence.
(B) *Secondary structure:* the amount of helix structure found within the protein and the number of atoms making up the helices.
(C) *Tertiary structure:* Ion pairs, hydrogen bonds, disulfide bonds and accessible surface area (ASA).
(D) *Extended properties:* Normalize the value obtained from (A), (B) and (C) above and the ratio of various other features such as polar/nonpolar.

3.2.3. Data generation

Several tools were used to calculate the secondary and tertiary structural features of each protein as needed. Helix packing pair (Dalton, Michalopoulos, & Westhead, 2003) was used to calculate the helix properties of the proteins. HBPLUS v3.0 (McDonald & Thornton, 1994) was used to calculate the hydrogen bonds in the proteins using the default parameters and with the distance between donor and accept atoms set at <3 Å for strong hydrogen bonds. EDPDB v03c (Matthews, 1995) was used to calculate the disulfide bonds and the accessible surface areas (ASA) of the proteins using the default parameters. In total, 111 features were created for each protein.

3.3. Method

3.3.1. Feature selection

The total number of features was 111. Some features will contribute in only a minor way to the thermostability, but others will have a major effect. Clearly, not all of these candidate features are critical related to protein thermostability. Inputting the whole dataset will therefore result in the data analysis process screening a lot of noisy information and this will give rise to weak results. Therefore, we adapted the correlation coefficient method in order to carry out feature selection. The use of correlation coefficients is similar to the way they are normally used in statistical (Wackerly & Scheaffer, 1996) and genetic analysis (Baumgartner et al., 2004). In statistics, correlation coefficients are used to measurement the linear relation between two random variables. Let the two random variables, X and Y, have n pair elements each. Under these circumstances, the correlation coefficient (r) between X and Y is defined as follow:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where (x_i, y_i) is the ith pair and $\bar{x}\bar{y}$ is the mean of X *and* Y, respectively.

Here we use correlation coefficients to measure the significance of features. We calculate the correlation between optimal growth temperature and each feature. After calculating all absolute values for the correlation coefficients, we extracted the features with $|r| \geqq 0.3$.

3.3.2. Naïve Bayes

Naïve Bayes is a data-mining and machine learning tool base on the Bayes theorem and assumes that each variable (property) is independence of each other. Compared with other machine learning approach, Naïve Bayes usually shows great speed and high accuracy when analyzing a large dataset. Let c denote a class, x denote a piece of evidence available to the machine learning algorithm. The conditional probability $P(c|x)$ represents the probability of class c when evidence x occurs. By Bayes' rule:

$$P(c|x) = \frac{p(c \cap x)}{p(x)} = \frac{p(c)p(x|c)}{p(x)}$$

where $p(x)$ is the prior probability of evidence x, $p(c)$ is the prior probability of class c, $p(x|c)$ is the probability of occurence that the evidence x is found under class c. For each test data with an unknown class label, we can assign a class label c_i which has maximum $P(c_i|x)$. We used the implementation of the Naïve Bayes classifier (NBC) by Tim Menzies.[3]

[3] NBC (http://www.cs.pdx.edu/~timm/dm/nbc.html).

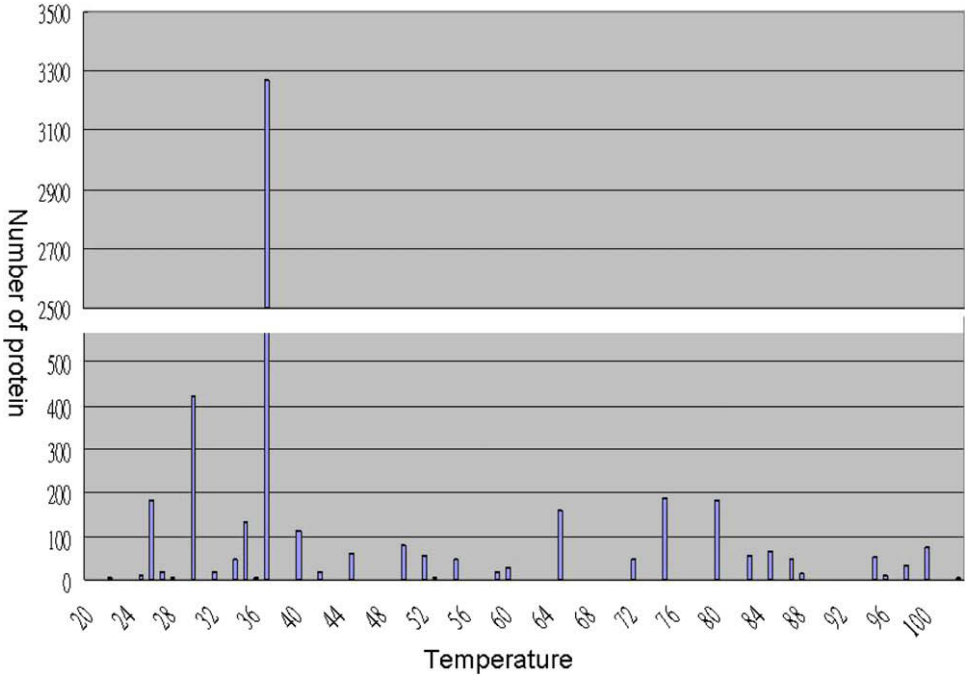

Fig. 2. Distribution of data.

3.3.3. Decision Tree

Decision Tree is a flow-chart-like tree structure (Han and Kamber, 2005). The topmost node is the root node, each internal node denotes an attribute test, each branch represents an outcome of the test, and each leaf node represents classes. In order to classify an unknown dataset, the attribute values are tested against the Decision Tree. A path is traced from the root to a leaf node that holds the class prediction for the test data, Different to Naïve Bayes and Neural Network, the output model built by Decision Tree is interpretable and can easily be converted to classification rules. We use Decision Tree tool C4.5[4] to training our dataset.

3.3.4. Neural Network

Neural Network is a machine learning approach that was inspired by biological nervous systems. Usually, the network consists of three major layers, input layer, hidden layer, output layer. Each layer consists of several perceptions unit (nerves), and sometimes the hidden layer consists of more than one layer when solving a complex problem.

We use a neural network tool (SNNS[5]) to design a network using the standard backpropagation algorithm, 10,000 cycles, learning rate 0.2, weight range $[-1,1]$ and one hidden layer .The amount of units in input layer is equal to the dimension of feature vector after feature selection. The number of units in hidden layer is double that of the input layer.

3.3.5. Model evaluation

An evaluation process is needed to measure the performance of models created by machine learning approach. We used 10-fold cross-validation to evaluate the model. Thus, 1810 proteins are separated into 10 subsets randomly and then each subset is taken as test data in turn. Four performance indices, accuracy, specificity, sensitivity and precision, were used during our evaluation. Accuracy is the percentage of all correct decisions made by classification algorithm. Sensitivity is the proportion of the data in class A that is also classified into class C. Specificity represent the proportion of data not in class C that are not classified into class C either. Precision is the proportion of data predicted in class C that is really in class A.

For example, each protein in data set has two thermostability class labels after the *K*-fold process. One is the protein's real thermostability class (C_{real}), and another is the thermostability class ($C_{predict}$) assigned by the model built by machine learning. If we focus on thermostability class mesophilic, then there is a subset labeled REAL that is made up of proteins whose C_{real} = mesophilic. After the *K*-fold process, another subset os produced labeled PREDICT, which consists of protein whose $C_{predict}$ = mesophilic. Let the four subsets TP, FP, TN and FN refer to True Positive, False Positive, True Negative, and False Negative, respectively, then

TP is consists of proteins whose both C_{real} and $C_{predict}$ are mesophilic.
FP is consists of proteins whose $C_{predict}$ = mesophilic, but C_{real} is not mesophilic.
TN is consists of proteins whose neither C_{real} nor $C_{predict}$ are mesophilic.
FN is consists of proteins whose C_{real} = mesophilic, but $C_{predict}$ is not mesophilic.
REAL is the union of TP and FN.
PREDICT is the union of FP and TP

Then the four performance indices are as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificivity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

[4] C4.5 (http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html).

[5] Stuttgart Neural Network Simulator (http://www-ra.informatik.uni-tuebingen.de/SNNS/).

To minimize the bias within such a random process, we repeated the above procedure one hundred times and used the average of the 100 results to measure our system.

Fig. 3 show the training and predicting processes. A model was created using a machine learning approach. The model contains several interpretable or non-interpretable rules depend on the training algorithm. From a given feature set for a protein, the model will calculate the thermostability class of protein following these rules.

4. Results

4.1. Correlation coefficient

We considered the 111 proteins features and the correlation coefficients showed that some feature only contributed in a minor way to thermostability. Table 1 shows the top 20 sequence features and their correlation coefficients. A full table of all correlation coefficients is given in the Appendix.

For this study, we selected a correlation coefficient threshold of 0.3, which gave a subset of eleven sequence features and three structural features. The sequence features were made up of glutamate (E)/residue, charged/non-charged, charged/residue, basic/residue, (E + F + M + R)/residue, lysine (K)/residue, glutamine (Q)/residue, threonine (T)/residue, acidic/residue, isoleucine (I)/residue and serine (S)/residue. The structural features consisted of the three strong ion pair formed by glutamate/arginine, glutamate/histidine and glutamate/lysine.

Table 1
Top 20 features with the highest correlation coefficients.

Feature	Correlation coefficient
Glutamate (E)/residue	0.561
Charged/non-charged	0.496
Non-charged/residue	0.496
Charged/residue	0.489
Basic/residue	0.458
(E + F + M + R)/residue	0.437
Lysine (K)/residue	0.421
Glutamine (Q)/residue	0.404
Threonine (T)/residue	0.388
Acidic/residue	0.363
Isoleucine (I)/residue	0.350
Glutamate (E)	0.328
Serine (S)/residue	0.325
Asparagine (N)/residue	0.286
Lysine (K)	0.285
Alanine (A)/residue	0.282
Glutamine (Q)	0.280
Ion pair (GLU_ARG_24)	0.330
Ion pair (GLU_HIS_24)	0.370
Ion pair (GLU_LYS_24)	0.380

4.2. Cross validation result

4.2.1. Three test cases

For this study, we designed three test cases to examine the models built by machine learning. We constructed three datasets that divided proteins into different temperature types, MT, TH and MTH. The MT dataset divided the protein thermostability by 50 °C. Using this division, we tried to identify the difference between mesophilic proteins and thermophilic proteins in the MT dataset. Thus, hyperthermophilic proteins are treated as thermophilic proteins in the MT dataset. Similarly, the TH dataset divided the protein thermostability by 80 °C and mesophilic proteins are treated as thermophilic proteins in the TH dataset. In the MTH dataset, the data is divided into three classes by 50 °C and by 80 °C and we attempted to identify all three different thermostability classes during the same process.

4.2.2. Comparing the three machine learning approaches

Table 2 show the accuracy results for mesophilicity based on the eleven features. This table reveals that the Decision Tree and Neural Network approaches gave better performance than Naïve Bayes. Overall, Decision Tree gave the highest performance overall because Neural Network gave a significantly poorer performance with class TH.

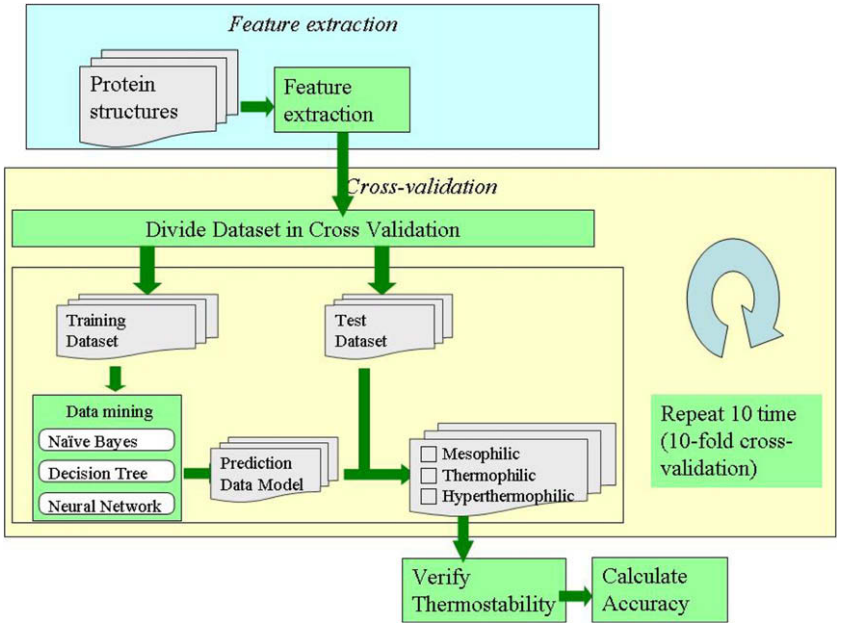

Fig. 3. Training and predicting system flow.

Table 2
Accuracy results based on the 11 sequence-only features.

Classifier	MT	TH	MTH
Naïve Bayes	0.76	0.78	0.77
Decision Tree	0.87	0.84	0.85
Neural Network	0.88	0.75	0.81

Table 3
Accuracy results based on all 14 features.

Classifier	MT	TH	MTH
Naïve Bayes	0.76	0.78	0.75
Decision Tree	0.87	0.84	0.86
Neural Network	0.85	0.82	0.74

Table 3 show the accuracy result for mesophilicity based on all 14 sequence and structural features. Again, like the results shown in Table 2, Decision Tree gave the best result overall.

4.3. Decision Tree detailed results

In the above section, Decision Tree gave good performance and the accuracy was generally better than 84%. In order to analyze the stability of the model, we repeat the cross-validation process 100 times and calculate the mean value and standard deviation of these 100 results.

Using the MT database, we can see from Table 4 that the mean values (AVG.) and standard deviations (STD.) for the four performance indices based on 100 results using 11 features. Firstly, accuracy, sensitivity, specificity and precision for mesophilicity and thermophilicity are all higher than 86%. Secondly, the standard deviations of all indices are 0.01 or lower. This means that our prediction system built by Decision Tree is both stable and gives good performance.

The features were then increased to 14 and Table 5 shows the mean values and standard deviations for the four performance indices based on 100 results. Compared to Table 4, the result still show good performance, which is all better than 85%.

For the TH dataset, the same format results are show in Tables 6 and 7. For 11 features system, all these indices are higher than 81% and the standard deviations are all lower than 0.024; thus the performance and stability are still good. For the 14 features system, all the indices are higher than 80% and the standard deviations are all lower than 0.021.

The same approach was used for the MTH dataset and Table 8 shows the results for the 11 features system, while Table 9 shows the results for the 14 features system. For 11 features, only precision for thermophilicity (65.0%), sensitivity for hyperthermophilicity (63.0%) and precision for hyperthermophilicity (69.4%) are lower than 70%. For mesophilicity, all four indices are better than 83%, while specificity for hyperthermophilicity is even higher than 93%. The standard deviations for the indices for three of the thermostability classes are all lower than 0.03. For 14 features, only precision in thermophilicity (68.9%) and sensitivity in hyperthermophilicity (69.7%) are lower than 70%, and all indices have improved compared with 11 features. Thus, based on standard deviation and error classification, we can say that our model shows stable performance under cross-validation.

In addition to these four indices, we also carried out a special test to evaluate the model's consistency comparing the model created by the MT dataset and that created by the TH dataset. For each protein, we use both models to predict the thermostability. Inconsistencies between the models may occur if the MT model specifies a protein is mesophilic while the TH model specifies it is hyperthermophilic. We calculate the percentage of proteins that show such inconsistency. Using two random models, the result should be an average of inconsistency of 25%. The output inconsistencies based on 11 features and 14 features were 1.8% and 1.6%, respectively. This meant that the models created by MT and TH datasets showed a low level of inconsistency.

Table 4
Detailed results for the four indices based on 11 features using the MT dataset.

	Mesophilicity		Thermophilicity	
	AVG. (%)	STD.	AVG. (%)	STD.
Accuracy	87.3	0.005	87.3	0.005
Sensitivity	86.1	0.008	88.5	0.010
Specificity	88.5	0.010	86.1	0.008
Precision	88.2	0.009	86.4	0.007

Table 5
Detailed results for the four indices based on all 14 features using the MT dataset.

	Mesophilic		Thermophilic	
	AVG. (%)	STD.	AVG. (%)	STD.
Accuracy	86.8	0.005	86.8	0.005
Sensitivity	85.8	0.007	87.7	0.010
Specificity	87.7	0.010	85.8	0.007
Precision	87.5	0.009	86.1	0.006

Table 6
Detailed results for the four indices based on 11 features using the TH dataset.

	Thermophilic		Hyperthermophilic	
	AVG. (%)	STD.	AVG. (%)	STD.
Accuracy	84.3	0.016	84.3	0.016
Sensitivity	81.1	0.024	87.6	0.018
Specificity	87.6	0.018	81.1	0.024
Precision	86.8	0.018	82.3	0.019

Table 7
Detailed results for the four indices based on all 14 features using the TH dataset.

	Thermophilic		Hyperthermophilic	
	AVG. (%)	STD.	AVG. (%)	STD.
Accuracy	83.5	0.016	83.5	0.016
Sensitivity	80.5	0.021	86.5	0.019
Specificity	86.5	0.019	80.5	0.021
Precision	85.6	0.019	81.7	0.017

Table 8
Detailed results for the four indices based on 11 features using the MTH dataset.

	Mesophilic		Thermophilic		Hyperthermophilic	
	AVG. (%)	STD.	AVG. (%)	STD.	AVG. (%)	STD.
Accuracy	87.0	0.0066	83.6	0.0075	89.3	0.0056
Sensitivity	87.3	0.0090	73.1	0.0144	73.2	0.0181
Specificity	86.8	0.0086	88.5	0.0084	93.2	0.0057
Precision	86.2	0.0080	75.0	0.0143	72.2	0.0171

Table 9
Detailed results for the four indices based on all 14 features for the MTH dataset.

	Mesophilic		Thermophilic		Hyperthermophilic	
	AVG. (%)	STD.	AVG. (%)	STD.	AVG. (%)	STD.
Accuracy	87.7	0.006	81.2	0.007	89.6	0.006
Sensitivity	85.2	0.009	75.8	0.015	69.7	0.023
Specificity	90.0	0.007	83.7	0.010	94.3	0.006
Precision	88.9	0.007	68.9	0.013	74.5	0.020

Table 10
New protein sequence dataset prediction results.

Index	MT	TH	MTH
Accuracy	0.67	0.74	0.68
Sensitivity	0.56	0.82	0.54
Specificity	0.77	0.51	0.80
Precision	0.68	0.82	0.70

4.4. Prediction results

Here we applied new protein datasets to our model. First, 11415 proteins were downloaded from SWISS-Pro via PGTdb. Among these, there were 9959 proteins with an optimal growth temperature lower than 50 °C, 869 proteins with one between 50 °C and 80 °C and 990 proteins with one higher than 80 °C. In order to make this test fair, we also used sampling of this data set. After sampling, the new dataset consisted of a total of 2943 proteins made up of 1377 mesophilic, 752 thermophilic and 814 hyperthermophilic proteins.

Table 10 shows the performance result. These proteins had only sequence information available and therefore the structural features aspect of the analysis could not be applied. This dataset is made up of many difference sequences, and the results showed a lower performance index than with the previously used dataset, especially that including structural information. The model based on the TH dataset showed highest sensitivity, which suggests that the TH model is better at detecting hyperthermophilic proteins.

5. Discussion and conclusions

A large amount of useful thermostability data on protein was obtained from PGTdb (Huang et al., 2004) and thermostability prediction based on this data was carried out. Several important features need to be highlighted from this research and the comparison of the various different machine learning techniques. Important structural and sequence features are highlighted. Those features with a higher correlation coefficient absolute values are important not only to our prediction but will also be useful to protein engineering. After comparing the difference methods, Decision Tree showed the best performance. Four performance indices, accuracy, sensitivity, specificity and precision were obtained for each dataset and overall high accuracy was achieved. The prediction results from the test data made up of 2943 protein from SWISS-Prot showed that an accuracy of better then 67% could be obtained when doing real predictions.

The sequence and structure features together gave a better performance than sequence features alone, which confirms previous studies, which have shown that structural feature contribute to thermostability. The sequence and structural model increased not only the four indices of accuracy, sensitivity, specificity and precision, but also reduced their standard deviations. Thus the sequence and structural model gives higher stability than the sequence alone model.

Two non-standard features were used in our research. (E + F + M + R)/residue and charged/non-charged get high linear correlation coefficients with optimal growth temperature. In Farias's research (Farias et al., 2004), a special ratio (E + K)/(Q + H) was also considered a significant factor when distinguishing thermophilic and mesophilic proteins. In our research, charged amino acids consisted of aspartic acid, glutamic acid, lysine, arginine and histidine. Non-charged amino acids are the remaining 15 amino acids. Although the biological significance of these two features has not been mentioned in previous research, the features (E + F + M + R)/residue and charged/non-charged were important factors in this work and need to be consider as new key features when studying protein thermostability.

Our final prediction system is based on Decision Tree. Decision Tree is an interpretable machine learning approach, which means that the output from Decision Tree consists of many explicable rules. Thus, hundreds of such rules were made available by this research. Each rule shows a series of conditions for determining protein thermostability. Future works will be related to how to use these rules to alter protein thermostability and this has been initiated.

Appendix A

All features and their correlation coefficients.

Features	Correlation coefficient
Alanine (A)	−0.117
Arginine (R)	0.171
Asparagine (N)	−0.197
Aspartic acid (D)	−0.126
Cysteine (C)	−0.009
Glutamic acid (E)	0.328
Glutamine (Q)	−0.280
Glycine (G)	−0.082
Histidine (H)	−0.049
Isoleucine (I)	0.177
Leucine (L)	0.101
Lysine (K)	0.285
Methionine (M)	−0.042
Phenylalanine (F)	−0.013
Proline (P)	−0.002
Serine (S)	−0.181
Threonine (T)	−0.216
Tryptophan (W)	−0.195
Tyrosine (Y)	−0.049
Valine (V)	0.139
Length	0.014
Acidic	0.161
Basic	0.218
Polar	−0.001
Non-polar	0.032
Cyclic	−0.057
Acyclic	0.030
Aliphatic	0.037
Aromatic	−0.075
Hydrophobic	0.032
Hydrophobic (+G)	0.014
Hydrophilic	−0.001
Charged	0.193
Non-charged	−0.056
Alanine (A)/residue	−0.282
Arginine (R)/residue	0.246
Asparagine (N)/residue	−0.286
Aspartic acid (D)/residue	−0.233
Cysteine (C)/residue	0.003
Glutamic acid (E)/residue	0.561
Glutamine (Q)/residue	−0.404
Glycine (G)/residue	−0.183
Histidine (H)/residue	−0.092
Isoleucine (I)/residue	0.350
Leucine (L)/residue	0.158
Lysine (K)/residue	0.421
Methionine (M)/residue	−0.040
Phenylalanine (F)/residue	−0.022

(*continued on next page*)

Appendix A (*continued*)

Features	Correlation coefficient
Proline (P)/residue	−0.027
Serine (S)/residue	−0.325
Threonine (T)/residue	−0.388
Tryptophan (W)/residue	−0.235
Tyrosine (Y)/residue	−0.121
Valine (V)/residue	0.249
Acidic/residue	0.363
Basic/residue	0.458
Polar/residue	−0.087
Non-polar/residue	0.087
Cyclic/residue	−0.185
Acyclic/residue	0.185
Aliphatic/residue	0.081
Aromatic/residue	−0.195
Hydrophobic/residue	0.087
Hydrophobic (+G)/residue	0.004
Hydrophilic/residue	−0.087
Charged/residue	0.489
Non-charged/residue	−0.489
Acidic/basic	−0.109
Basic/acidic	0.124
Polar/nonpolar	−0.121
Nonpolar/polar	0.052
Cyclic/acyclic	−0.171
Acyclic/cyclic	0.188
Charged/non-charged	0.496
Non-charged/charged	−0.393
Hydrophobic/hydrophilic	0.052
Hydrophilic/hydrophobic	−0.121
Hydrophobic (+G)/hydrophilic	0.012
Hydrophilic/hydrophobic (+G)	−0.074
(E + F + M + R)/residue	0.437
ASA	0.029
Disulfide bond	0.017
Hydrogen bond	−0.038
Strong hydrogen bond	−0.019
Hydrogen bond/residue	−0.034
Strong hydrogen bond/residue	−0.015
Helix	0.034
Atom in helix	0.063
Helix/residue	0.050
Atom in helix/residue	0.091
Ion pair (ASP_ARG_24)	0.013
Ion pair (ASP_HIS_24)	0.025
Ion pair (ASP_LYS_24)	−0.008
Ion pair (GLU_ARG_24)	0.330
Ion pair (GLU_HIS_24)	0.370
Ion pair (GLU_LYS_24)	0.380
All ion pair_24	0.224
Ion pair (ASP_ARG_46)	0.233
Ion pair (ASP_HIS_46)	0.245
Ion pair (ASP_LYS_46)	−0.067
Ion pair (GLU_ARG_46)	−0.201
Ion pair (GLU_HIS_46)	−0.174
Ion pair (GLU_LYS_46)	0.033
All ion pair_46	0.082
Ion pair (ASP_ARG_68)	0.064
Ion pair (ASP_HIS_68)	−0.042
Ion pair (ASP_LYS_68)	−0.057
Ion pair (GLU_ARG_68)	−0.035
Ion pair (GLU_HIS_68)	0.181
Ion pair (GLU_LYS_68)	0.194
All ion pair_68	0.180

References

Baumgartner, C., Bohm, C., Baumgartner, D., Marini, G., Weinberger, K., Olgemoller, B., et al. (2004). *Bioinformatics, 20*(17), 2985–2996.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). *Nucleic Acids Research, 28*(1), 235–242.

Capriotti, E., Fariselli, P., & Casadio, R. (2004). *Bioinformatics, 20*(Suppl. 1), I63–I68.

Capriotti, E., Fariselli, P., & Casadio, R. (2005). *Nucleic Acids Research, 33*(Web Server issue), W306–310.

Chan, C. H., Liang, H. K., Hsiao, N. W., Ko, M. T., Lyu, P. C., & Hwang, J. K. (2004). *Proteins, 57*(4), 684–691.

Dalton, J. A., Michalopoulos, I., & Westhead, D. R. (2003). *Bioinformatics, 19*(10), 1298–1299.

Dominy, B. N., Minoux, H., & Brooks, C. L. 3rd, (2004). *Proteins, 57*(1), 128–141.

Farias, S. T., van der Linden, M. G., Rego, T. G., Araujo, D. A., & Bonato, M. C. (2004). *In Silico Biology, 4*(3), 377–380.

Gianese, G., Bossa, F., & Pascarella, S. (2002). *Proteins, 47*(2), 236–249.

Gromiha, M. M., Oobatake, M., & Sarai, A. (1999). *Biophysical Chemistry, 82*(1), 51–67.

Haney, P. J., Stees, M., & Konisky, J. (1999). *Journal of Biological Chemistry, 274*(40), 28453–28458.

Huang, S. L., Wu, L. C., Huang, H. D., Liang, H. K., Ko, M. T., & Horng, J. T. (2004). *Applied Bioinformatics, 3*(1), 21–29.

Huang, Shir-Ly, Wu, Li-Cheng, Huang, Hsien-Da, Liang, Han-Kuen, Ko, Ming-Tat, & Horng, J.-T. (2004). *Applied Bioformatics, 3*(1), 21–29.

Huang, S. L., Wu, L. C., Liang, H. K., Pan, K. T., Horng, J. T., & Ko, M. T. (2004). *Bioinformatics, 20*(2), 276–278.

Han, J., & Kamber, M. (2005). *Data mining concepts and techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Kannan, N., & Vishveshwara, S. (2000). *Protein Engineering, 13*(11), 753–761.

Liang, H. K., Huang, C. M., Ko, M. T., & Hwang, J. K. (2005). *Proteins, 59*(1), 58–63.

Matthews, X. Z. a. B. W. (1995). *Journal of Applied Crystallography, 28*, 624–630.

McDonald, I. K., & Thornton, J. M. (1994). *Journal of Molecular Biology, 238*(5), 777–793.

Parthasarathy, S., & Murthy, M. R. (2000). *Protein Engineering, 13*(1), 9–13.

Petukhov, M., Kil, Y., Kuramitsu, S., & Lanzov, V. (1997). *Proteins, 29*(3), 309–320.

Ragone, R. (2001). *Protein Science, 10*(10), 2075–2082.

Saraboji, K., Gromiha, M. M., & Ponnuswamy, M. N. (2006). *Biopolymers*.

Szilagyi, A., & Zavodszky, P. (2000). *Structure with Folding and Design, 8*(5), 493–504.

Vieille, C., & Zeikus, G. J. (2001). *Microbiology and Molecular Biology Reviews, 65*(1), 1–43.

Vogt, G., Woell, S., & Argos, P. (1997). *Journal of Molecular Biology, 269*(4), 631–643.

Wackerly, D. D. W. M., III, & Scheaffer, R. L. (1996). *Mathematical statistics with applications*. New York: Duxbury Press.