EXCLI Journal 2009;8:218-233 – ISSN 1611-2156
Received: August 22, 2009, accepted: August 29, 2009, published: September 01, 2009

ORIGINAL ARTICLE:

SEARCHING FOR PATTERNS OF THERMOSTABILITY IN PROTEINS AND DEFINING THE MAIN FEATURES CONTRIBUTING TO ENZYME THERMOSTABILITY THROUGH SCREENING, CLUSTERING, AND DECISION TREE ALGORITHMS

M. Ebrahimi[1§], E. Ebrahimie[2], M. Ebrahimi[3]

[1] Bioinformatics Research Group, Green Research Center, Qom University, Qom, IRAN
[2] Department of Crop Production & Plant Breeding, College of Agriculture, Shiraz University, Shiraz, IRAN; Email: ebrahimie@shirazu.ac.ir
[3] Department of Information Technology, International University in Germany, 76646 Bruchsal, Germany; Email: mahdi.ebrahimi@i-u.de

[§] Corresponding author: M. Ebrahimi, email: mebrahimi14@gmail.com

ABSTRACT

Finding or making thermostable enzymes has been identified as an important goal in a number of different industries. Therefore, understanding the features involved in enzyme thermostability is crucial, and different approaches have been used to extract or manufacture thermostable enzymes. Herein we examined features that contribute to the thermostability of 2,946 proteins. We used various screening techniques (anomaly detection, feature selection), clustering methods (K-Means, TwoStep cluster), decision tree models (Classification and Regression Tree, CHAID, Exhaustive CHAID, QUEST, C5.0), and generalized rule induction (association) (GRI) models to search for patterns of thermostability and to find features that contribute to enzyme thermal stability. We found that Arg as the N-terminal amino acid was found solely in proteins working at temperatures higher than 70 ºC. Fifty-four protein features were shown to be important in feature selection modeling, and the number of peer groups with an anomaly index of 2.12 declined from 7 to 2 after being run using only important selected features; however, no changes were found in the numbers of groups when K-Means and TwoStep clustering modeling was performed on datasets with/without feature selection filtering. The depth of the trees generated by various decision tree models varied from 14 (in the C5.0 model with 10-fold cross-validation and with feature selection of the dataset) to 4 (in CHAID models) branches. The performance evaluation of the decision tree models tested here showed that C5.0 was the best and the Quest model was the worst. We did not find any significant difference in the percent of correctness, performance evaluation, and mean correctness of various decision tree models when feature selected datasets were used, but the number of peer groups in clustering models was reduced significantly ($p<0.05$) compared to datasets without feature selection. In all decision tree models, the frequency of Gln was the most important feature for decision tree rule sets; moreover, in all GRI association rules (100 rules), the frequency of Gln was used in antecedent to support the rules. The importance of Gln in protein thermostability is discussed in this paper.

Keywords: Bioinformatics, modeling, protein, thermostability

Received: August 22, 2009, accepted: August 29, 2009, published: September 01, 2009

INTRODUCTION

One of the most important tasks in protein engineering is to understand the factors responsible for the extreme stability of thermophilic proteins and to discriminate these proteins from mesophilic ones (Kongsted et al., 2007). Several methods based on amino acid substitutions have been proposed for predicting the stability of proteins (Gromiha et al., 2002). These methods are mainly based on distance and torsion potentials (Gromiha, 2007), multiple regression techniques (Gromiha, 2003), energy functions (Hamelryck, 2009), contact potentials (Gromiha & Selvaraj, 2004; Lisewski, 2008), neural networks (Hayashi et al., 2005), support vector machines (Garg & Raghava, 2008; Kumar et al., 2007), average assignment (Gromiha, 2007), classification and regression tools (Huang et al., 2007a, c), and backbone flexibility (Davis & Baker, 2009). In many of these cases, the discrimination of stabilizing and destabilizing mutants was reported to be more important than the actual magnitude of stability (Gromiha et al., 2002; Huang et al., 2007b). Most of these methods use information about the three-dimensional structures of proteins for discrimination/prediction. Prediction accuracy using amino acid sequences is significantly lower than that using structural data (Parthiban et al., 2007). However, several attempts have been made to understand the role of amino acid sequences on thermophilic protein stability. For example, an increased number of salt bridges and side chain-side chain interactions (Natesh et al., 2003; Yang et al., 2005), counterbalance between packing and solubility (Gromiha et al., 1999b), aromatic clusters (Cicortas Gunnarsson et al., 2007), contacts between the residues of hydrogen bonds (Kongsted et al., 2007), ion pairs (Ihsanawati et al., 2005), cation-π interactions (Kiarie et al., 2007), non-canonical interactions (Chakkaravarthi et al., 2006), electrostatic interactions of charged residues and the dielectric response (De Lemos Esteves et al., 2005), amino acid coupling patterns (Schubert et al., 2007), main-chain hydrophobic free energy, and hydrophobic residues (Miyazaki et al., 2006) have been reported to enhance stability. In addition, the amino acid sequences of genomes have been used to study the stability of thermophilic proteins (Ralph et al., 2008). Intra-helical salt bridges reportedly are prevalent in thermophiles, and the amino acid composition on the protein surface might be an important factor in stability (Umemoto et al., 2007). Moreover, the proteomes of thermophilic proteins are enriched in hydrophobic and charged amino acids at the expense of polar ones (Yang et al., 2005). Although numerous studies have focused on studying the stability of thermophilic proteins, a system that derives stability rules for any input data and converts them into a prediction is still lacking.

Data mining problems often involve hundreds, or even thousands, of variables (Ye et al., 2009). Fitting a neural network or a decision tree to a set of variables this large may require more time than is practical (Gromiha & Yakubi, 2008). Usually, many attributes determine the different characteristics of a protein molecule. As a result, the majority of time and effort spent in the model-building process involves determining which variables to include in the model. Feature selection allows the variable set to be reduced in size, creating a more manageable set of attributes for modeling (Thai & Ecker, 2009).

The decision tree algorithm (Dancey et al., 2007) 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, determining the features to add at the next level of the tree using an entropy calculation, and then choosing the feature that minimizes the entropy impurity (Gromiha, 2007). Several well-known decision tree algorithms are available. To better understand the features that contribute to an enzyme's thermal stability, it is necessary to identify the main features responsible for this valuable characteristic. Herein we used various clustering, screening, and decision tree models to determine the most important features responsible for thermostability.

Received: August 22, 2009, accepted: August 29, 2009, published: September 01, 2009

MATERIALS & METHODS

From the UniProt Knowledgebase (Swiss-Prot and TrEMBL) database, sequences from 2,946 proteins with different optimum temperature activities were retrieved and categorized into F (optimum temperature < 70 ºC) and T (optimum temperature $\geq$ 70 ºC) groups. Seventy-four protein attributes or features from all of those proteins were extracted. All features were classified as continuous variables, except for the N-terminal amino acid, which was classified as categorical. A dataset of these protein features was imported into Clementine software (Clementine_NLV-11.1.0.95; Integral Solution, Ltd.), null data for optimum temperature were discarded, and optimum temperature was set as the output variable and the other variables were set as input variables.

To identify the most important features and find possible patterns that contribute to protein thermostability, various decision tree algorithms were applied to the datasets. These models allowed the development of classification systems that automatically included in their rules only the attributes that really matter in making a decision. Attributes that did not contribute to the accuracy of the tree were ignored. This process yielded very useful information about the data and could be used to reduce the data to relevant fields only before training another learning technique, such as a neural network. Various algorithms are available for performing classification and segmentation analysis, and herein we used different decision tree and cluster analysis models. To investigate the effects of the feature selection algorithm on other models behaviour, all models were run both with and without feature selection criteria.

1. Screening Models

a. Anomaly detection model

This model was used to identify outliers or unusual cases in the data. Unlike other modeling methods that store rules about unusual cases, anomaly detection models store information on what normal behavior looks like. This makes it possible to identify outliers even if they do not conform to any known pattern. While traditional methods of identifying outliers generally examine one or two variables at a time, anomaly detection can examine large numbers of fields to identify clusters or peer groups into which similar records fall. Each record then can be compared to others in its peer group to identify possible anomalies. The further away a case is from the normal center, the more likely it is to be unusual.

b. Feature selection algorithm

The feature selection algorithm was applied to identify the attributes that have a strong correlation with the thermostability of enzymes. The algorithm considers one attribute at a time to determine how well each predictor alone predicts the target variable. The important value for each variable is then calculated as (1–p), where p is the p value of the appropriate test of association between the candidate predictor and the target variable. The association test for categorized output variables differs from the test for continuous variables. In our study, when the target value was categorical (as in our datasets), p values based on the F statistic were used. The idea was to perform a one-way ANOVA F test for each predictor; otherwise, the p value was based on the asymptotic t distribution of a transformation of the Pearson correlation coefficient. Other models, such as likelihood-ratio chi-square (also tests for target-predictor independence), Cramer's V (a measure of association based on Pearson's chi-square statistic), and Lambda (a measure of association that reflects the proportional reduction in error when the variable is used to predict the target value) were conducted to check for possible effects of calculation on feature selection criteria. The predictors were then labeled as important, marginal, and unimportant, with values > 0.95, between 0.95 and 0.90, and < 0.90, respectively.

2. Clustering Models

a. K-Means

The K-Means model can be used to cluster data into distinct groups when clustering groups are unknown. Unlike most learning methods in Clementine, K-Means models do not use a target field. This type of learning, with no target field, is called unsupervised learning. Instead of trying to predict an outcome, K-Means tries to uncover patterns in the set of input fields. Records are grouped so that records within a group or cluster tend to be similar to each other, whereas records in different groups are dissimilar. K-Means works by defining a set of starting cluster centers derived from the data. It then assigns each record to the cluster to which it is most similar based on the record's input field values. After all cases have been assigned, the cluster centers are updated to reflect the new set of records assigned to each cluster. The records are then checked again to see whether they should be reassigned to a different cluster, and the record assignment/cluster iteration process continues until either the maximum number of iterations is reached or the change between one iteration and the next fails to exceed a specified threshold.

b. TwoStep cluster

The TwoStep cluster model is a two-step clustering method. The first step makes a single pass through the data, during which it compresses the raw input data into a manageable set of subclusters. The second step uses a hierarchical clustering method to progressively merge the subclusters into larger and larger clusters, without requiring another pass through the data. Hierarchical clustering has the advantage of not requiring the number of clusters to be selected ahead of time. Many hierarchical clustering methods start with individual records as starting clusters and merge them recursively to produce ever-larger clusters.

3. Decision Tree Models

a. Classification and regression tree (C&RT)

This model uses recursive partitioning to split the training records into segments by minimizing the impurity at each step. A node is considered pure if 100 % of cases in the node fall into a specific category of the target field.

b. CHAID

This method generates decision trees using chi-square statistics to identify optimal splits. Unlike the C&RT and QUEST models, CHAID can generate non-binary trees, meaning that some splits can have more than two branches.

c. Exhaustive CHAID

This model is a modification of CHAID that does a more thorough job of examining all possible splits, but it takes longer to compute.

d. QUEST

The QUEST model provides a binary classification method for building decision trees. It is designed to reduce the processing time required for large C&RT analyses while also reducing the tendency found in classification tree methods to favor predictors that allow more splits.

e. C5.0

The C5.0 model builds either a decision tree or a rule set. The model works by splitting the sample based on the field that provides the maximum information gain at each level. The target field must be categorical. Multiple splits into more than two subgroups are allowed.

4. Association Model

The generalized rule induction (GRI) model discovers association rules in the data. GRI extracts a set of rules from the data, pulling out the rules with the highest information content. Information content is measured using an index that takes both the generality (support) and accuracy (confidence) of rules into account.

IMPLEMENTATION

To compare the effects of feature selection algorithms on various models used in this study, all models were applied to the data sets with or without the use of feature selection. This means that we reapplied the models on all protein features and on selected features suggested by feature selection. As a result, C5.0, C5.0 with 10-fold cross validation, C&RT,

QUEST, CHAID, Exhaustive CHAID, K-Means, TwoStep cluster, Anomaly and GRI were created both for datasets with all and with only important protein features (10 models) on each dataset. The percentage of correct and wrong, performance evaluation, range, mean correct, and mean incorrect variables for all models were calculated and are presented here.

RESULTS

The average length, weight, isoelectric point, and aliphatic indices of proteins studied here were 322.4 ± 209.9, 36.2 ± 24.9, 7.2.4 ± 1.7, and 97.9 ± 15.2 (mean ± SD), respectively. The average counts of sulphur, carbon, nitrogen, oxygen, and hydrogen were 11.01, 201.86, 368.57, 383.65, and 89.55, respectively, and the average counts of hydrophobic, hydrophilic, and other residues were 217.1, 137.2, and 102.3, respectively. The frequencies of hydrogen, carbon, oxygen, nitrogen, and sulphur in all enzymes were 0.504 ± 0.006, 0.316 ± 0.006, 0.092 ± 0.005, 0.86 ± 0.005 and 0.002 ± 0.001, respectively, and the frequencies of hydrophobic, hydrophilic, other, negatively, and positively residues were 0.521 ± 0.067, 0.217 ± 0.45, 0.263 ± 0.065, 3.83 ± 14.25, and 3.38 ± 12.02, respectively. The frequencies of amino acids ranged from a low of 0.01 ± 0.001 for Cys to a high of 0.102 ± 0.031 for Leu.

In 97.89 % of proteins the N-terminal amino acid was Met; in 0.85 %, 0.48 %, 0.31 %, and 0.17 % of proteins the same position was occupied by Ala, Ser, Thr, and Pro, respectively. In 0.07 % the last amino acid was either Isl, Gly, or Asp, and in 0.03 % the N-terminal amino acid was Lys, Cys, or Arg. The average non-reduced Cys extinction coefficient at 280 nm was 60.51, non-reduced Cys absorption was 0.91, the reduced Cys extinction coefficient was 39.07, and the reduced Cys absorption was 0.90.

Figure 1 is a web graph that illustrates the strength of the relationship between N-terminal amino acids and optimum temperature of proteins. Met exhibits a strong relationship with the proteins' temperature character (a thicker line shows a stronger relationship). Arg was the only N-terminal amino acid found in proteins with optimum temperatures > 70 ºC, whereas Ala, Pro, Gly, Ser, Cys, Thr, Lys, Asp, and Ile were found at the N-terminal in proteins with optimum temperatures < 70 ºC.

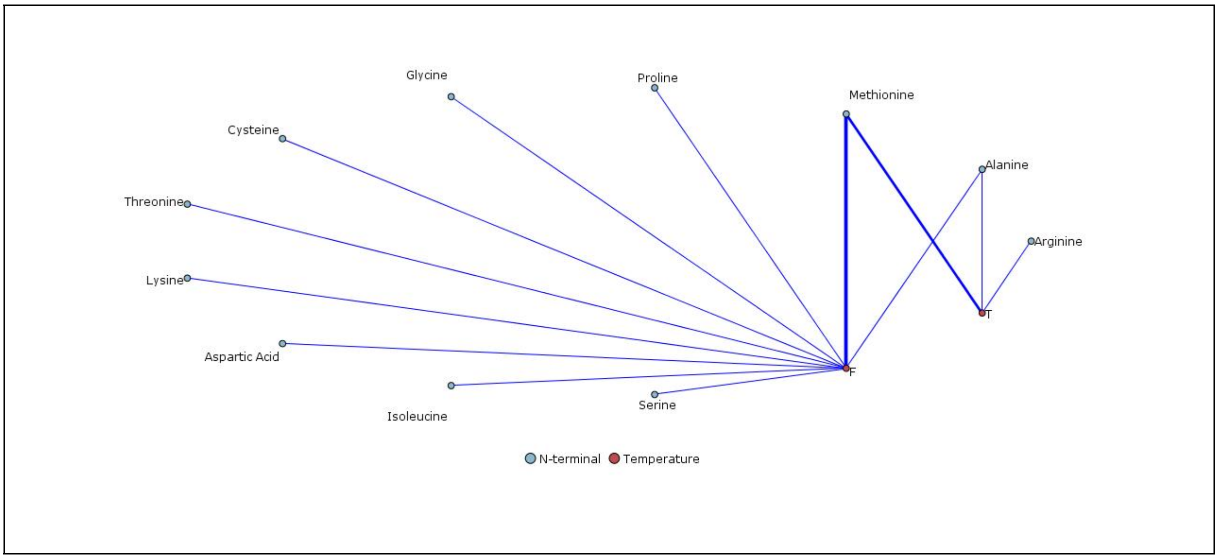


Figure 1: Web graph of N-terminal amino acids in both F and T protein groups, thicker lines showing higher incidences of amino acids.

1. Screening Models

a. Anomaly detection model

When the anomaly detection model was used, the records divided into seven peer groups with an anomaly index cutoff of 2.12 (Figure 2). In the first peer group of 814 records, 6 records found to be anomalies. In peer groups 2 to 7 there were 520, 262, 321, 441, 246, and 342 records with 2, 4, 4, 5, 6, and 3 anomalous records, respectively. The highest anomaly index was 5 (for two records: in peer groups 3 (half life of *E. Coli*) and 1 (frequency of His)) followed by 4.35 in peer group 6 (frequency of Phe). When the models were applied using feature selection criteria, just two peer groups with an anomaly index cutoff of 2.65 were found. In the first peer group of 1402 records, 21 anomalous records were found, and 8 anomalies were found in the 1544 records of the second peer group. Again, the highest anomaly index was 5 (for five records, one in the first peer group and the others in the second peer group, for count of Asn, frequency of Phe, frequency of Trp, count of hydrogen, and frequency of His fields).

b. Feature selection

Fifty-four out of 75 features ranked as important (p > 0.95) in contributing to protein thermostability (Table 1), and just one feature (weight p = 0.94) was found to be marginal. A node generated with just important features and used whenever it was necessary to run all other models on feature selection dataset (as mentioned in Materials and Methods).

2. Clustering Models

a. K-Means

In this clustering model, more than 55 % of the records (1630) were put into the first cluster and 58, 191, 296, and 771 records were put into the second, third, fourth, and fifth clusters, respectively. When the K-Means model was applied on the dataset with feature selection filtering, again five clusters were generated, with 1420, 427, 1044, 54, and 1 records in each cluster, respectively.

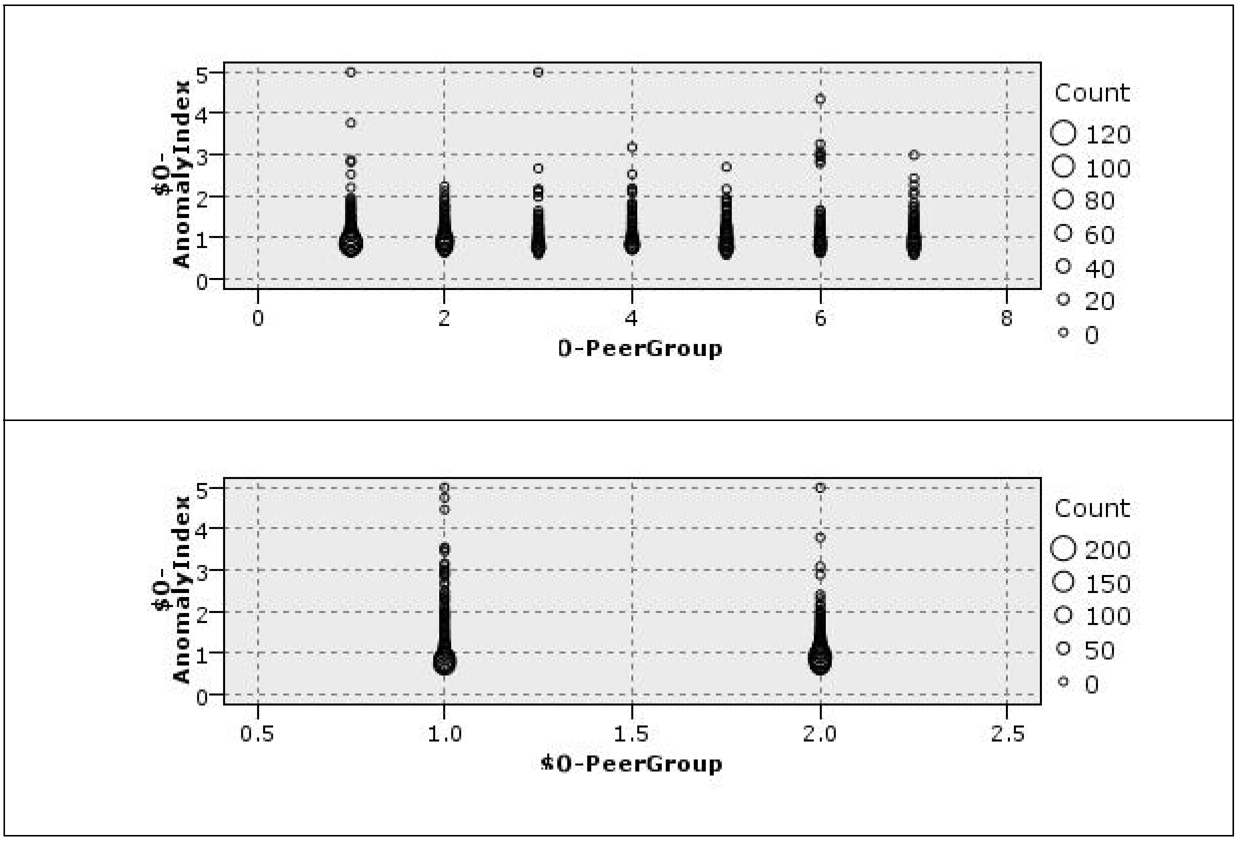


Figure 2: Anomaly peer groups in all records without feature selection (top) and with feature selection filtering (bottom).

Table 1: Results of feature selection on important (and one marginal) features contributing to the optimum temperature of proteins

No	Field	Value	Rank	No	Field	Value	Rank
1	Freq. of Gln	1	Important	**29**	Freq. of Ser	0.99	Important
2	Freq. of Glu	1	Important	**30**	Count of Tyr	0.99	Important
3	Freq. of Hydrophil (C.N.Q.S.T.Y)	1	Important	**31**	Count of His	0.99	Important
4	Freq. of Other	1	Important	**32**	Freq. of Hydrophobic	0.99	Important
5	Count of Gln	1	Important	**33**	Count of Ala	0.99	Important
6	Freq. of Lys	1	Important	**34**	Count of Hydrophilic (C.N.Q.S.T.Y)	0.99	Important
7	Freq. of Thr	1	Important	**35**	Count of Arg	0.99	Important
8	Freq. of Val	1	Important	**36**	Freq. of Ile	0.99	Important
9	Freq. of Ala	1	Important	**37**	Freq. of Met	0.99	Important
10	Count of Glu	1	Important	**38**	Count of Thr	0.99	Important
11	Freq. of Other	1	Important	**39**	Count of Cys	0.99	Important
12	Freq. of Leu (L)	1	Important	**40**	Half-life mammals	0.99	Important
13	Freq. of Trp	1	Important	**41**	Freq. of Cys	0.99	Important
14	Count of Lys	1	Important	**42**	Freq. of Gly	0.99	Important
15	Freq. of Tyr	1	Important	**43**	Count of Leu (L)	0.99	Important
16	Freq. of Positively Charged (R & K)	1	Important	**44**	Count of Asn	0.99	Important
17	Freq. of Negatively Charged (D & E)	1	Important	**45**	Count of sulphur (S)	0.99	Important
18	Freq. of His	1	Important	**46**	Freq. of Asp	0.99	Important
19	Positively Charged (R & K)	1	Important	**47**	Count of Ile	0.99	Important
20	Freq. of Arg	1	Important	**48**	Freq. of Phe	0.99	Important
21	Freq. of Negatively Charged (D & E)	1	Important	**49**	Count of hydrogen (H)	0.99	Important
22	Freq. of Other	1	Important	**50**	Count of Asp	0.99	Important
23	Freq. of Asn	1	Important	**51**	Count of Ser	0.99	Important
24	Non-reduced Cys Absorption at 280 nm	1	Important	**52**	Count of Gly	0.98	Important
25	Reduced Cys Absorption at 280 nm	1	Important	**53**	Count of Met	0.98	Important
26	Freq. of sulphur (S)	1	Important	**54**	Count of Other charged residues	0.96	Important
27	Count of Val	0.99	Important	**55**	Weight	0.94	Marginal
28	Count of Trp	0.99	Important				

Received: August 22, 2009, accepted: August 29, 2009, published: September 01, 2009

b. TwoStep Cluster

This method clustered records into six groups with 281, 44, 890, 1046, 181, and 495 records in each cluster, respectively. Only two clusters (with 418 and 2528 records in each cluster) were created for the dataset filtered using feature selection criteria.

3. Decision Tree Models

The C5.0 model generated a decision tree with a depth of 12 and cross-validation of 86.9 ± 0.8 was created. The most important feature used to build the tree was the frequency of Gln. If the value of this feature was equal to or less than 0.031, the optimum temperature of proteins fell into the T category ($\geq$ 70 ºC); otherwise they were put into the F category (< 70 ºC). In the T subgroup, the frequency of positively charged residues was used to create the next tree branches, with $\leq$ 3 as T mode and > 3 as F mode. In the F subgroup, if the value for the frequency of hydrophilic residues was equal to or greater than 0.172, they were placed in the T subgroup; otherwise they were put into the F subgroup. When 10-fold cross-validation was applied to the same dataset, again a tree with a depth of 12 and cross-validation of 86.5 ± 0.4 was created. The same protein features and values were used to create tree branches. When the same models were applied to datasets using feature selection filtering, a tree with a depth of 14 and cross-validation of 87.3 ± 0.7 and 86.6 ± 0.8 were generated for C5.0 and C5.0 with 10-fold cross-validation, respectively. The frequency of Gln, the frequency of positively charged residues, and the frequency of hydrophilic residues, again with the same values, were used to create the first and second subgroups.

In C&RT node, a tree with a depth of 5 was created, and the most important feature used to build the tree was the frequency of Gln (value $\leq$ 0.028 for T and > 0.028 for F). The frequency of other charged residues was used to create the second level for both subgroups (0.822 for T and 0.732 for F). The same results were obtained when feature selection was used.

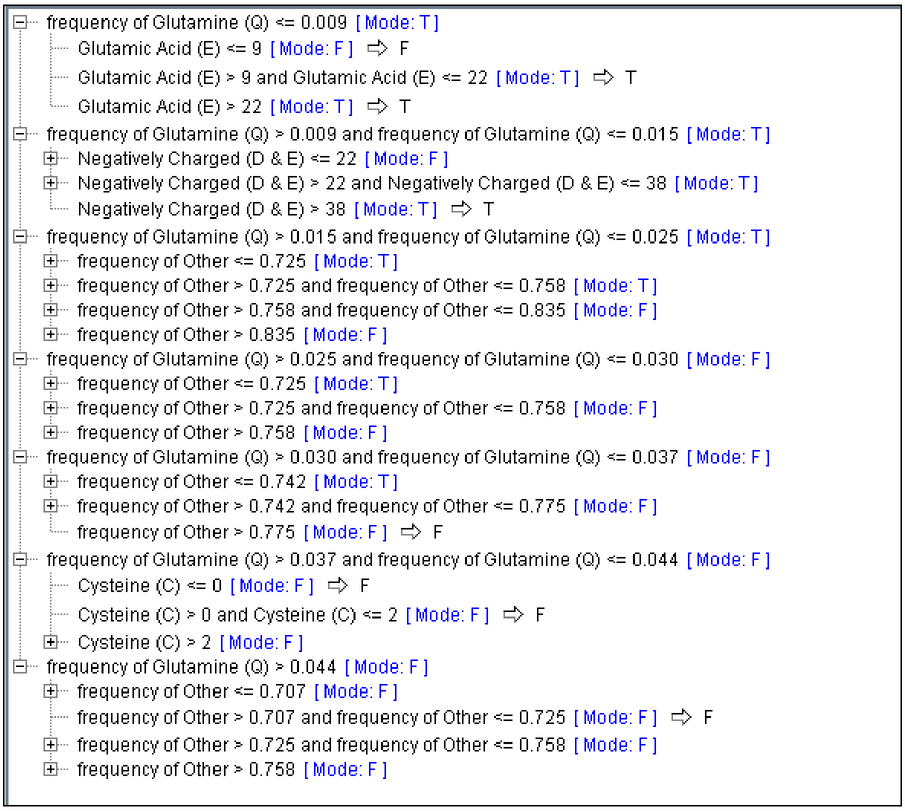
In Quest modeling, a tree with a depth of 5 was generated, and again the frequency of Gln (with a value of 0.028) was used to create the first tree branches. In the T subgroup, frequency of Glu (0.053) was used to generate the next subgroup; in the F subgroup, the frequency of Lys (0.113) was used. The same results occurred when feature selection filtering was applied.

When the CHAID model was applied to the data with and without feature selection, a tree with a depth of 4 was generated. If the frequency of Gln was > 0.044, the mode was F; if it was $\leq$ 0.009 the mode was T. If the frequency of Gln was between 0.009 and 0.015 and the count of negatively charged residues was greater than 22, the mode was T; otherwise it was F. When the frequency of Gln was > 0.015 and < 0.025 it formed the next branch, and three other branches were created when the same feature was between 0.025 and 0.030, 0.030 and 0.037, and 0.037 and 0.044 (Figure 3.). The same trees with the same features and values were generated when exhaustive CHAID models were applied on datasets with and without feature selection.

The best percentage of correctness, performance evaluation, and mean correctness in the decision tree models were observed in the C5.0 model, followed by the CR&T, CHAID, and finally the Quest models (Table 2).

4. Association Model

GRI node analysis created 100 rules with 2947 valid transactions with minimum and maximum support of 15.82 % and 27.12 %, respectively. Maximum confidence reached 97.42 % and minimum confidence decreased to 85.86 %. When feature selection was used, minimum support, maximum support, maximum confidence, and minimum confidence changed to 15.17 %, 27.12 %, 97.42 %, and 84.81 %, respectively. The highest confidence (97.42 %) in both methods (with/without feature selection filtering) occurred when the frequency of Gln was lower than 0.028, the count of Val was greater than 14.5, and the frequency of Glu was greater than 0.086 (Table 3).



```
⊟ frequency of Glutamine (Q) <= 0.009 [Mode: T]
    Glutamic Acid (E) <= 9 [Mode: F] ⇨ F
    Glutamic Acid (E) > 9 and Glutamic Acid (E) <= 22 [Mode: T] ⇨ T
    Glutamic Acid (E) > 22 [Mode: T] ⇨ T
⊟ frequency of Glutamine (Q) > 0.009 and frequency of Glutamine (Q) <= 0.015 [Mode: T]
  ⊞ Negatively Charged (D & E) <= 22 [Mode: F]
  ⊞ Negatively Charged (D & E) > 22 and Negatively Charged (D & E) <= 38 [Mode: T]
    Negatively Charged (D & E) > 38 [Mode: T] ⇨ T
⊟ frequency of Glutamine (Q) > 0.015 and frequency of Glutamine (Q) <= 0.025 [Mode: T]
  ⊞ frequency of Other <= 0.725 [Mode: T]
  ⊞ frequency of Other > 0.725 and frequency of Other <= 0.758 [Mode: T]
  ⊞ frequency of Other > 0.758 and frequency of Other <= 0.835 [Mode: F]
  ⊞ frequency of Other > 0.835 [Mode: F]
⊟ frequency of Glutamine (Q) > 0.025 and frequency of Glutamine (Q) <= 0.030 [Mode: F]
  ⊞ frequency of Other <= 0.725 [Mode: T]
  ⊞ frequency of Other > 0.725 and frequency of Other <= 0.758 [Mode: F]
  ⊞ frequency of Other > 0.758 [Mode: F]
⊟ frequency of Glutamine (Q) > 0.030 and frequency of Glutamine (Q) <= 0.037 [Mode: F]
  ⊞ frequency of Other <= 0.742 [Mode: T]
  ⊞ frequency of Other > 0.742 and frequency of Other <= 0.775 [Mode: F]
    frequency of Other > 0.775 [Mode: F] ⇨ F
⊟ frequency of Glutamine (Q) > 0.037 and frequency of Glutamine (Q) <= 0.044 [Mode: F]
    Cysteine (C) <= 0 [Mode: F] ⇨ F
    Cysteine (C) > 0 and Cysteine (C) <= 2 [Mode: F] ⇨ F
  ⊞ Cysteine (C) > 2 [Mode: F]
⊟ frequency of Glutamine (Q) > 0.044 [Mode: F]
  ⊞ frequency of Other <= 0.707 [Mode: F]
    frequency of Other > 0.707 and frequency of Other <= 0.725 [Mode: F] ⇨ F
  ⊞ frequency of Other > 0.725 and frequency of Other <= 0.758 [Mode: F]
  ⊞ frequency of Other > 0.758 [Mode: F]
```

Figure 3: A decision tree generated by the CHAID modeling method without feature selection filtering

Table 2: Percentage of correctness, wrongness, performance evaluation (T & F), and mean correct and incorrect in various decision tree models, in datasets without feature selection (a) and with feature selection (b)

(a)						
	% Correct	**% Wrong**	**Performance evaluation (T)**	**Performance evaluation (F)**	**Mean correct**	**Mean incorrect**
C5.0	95.52 %	4.48 %	0.915	0.436	0.942	0.814
C5.0 with 10-fold validation	95.52 %	4.48 %	0.915	0.436	0.942	0.814
CR&T	88.56 %	11.44 %	0.798	0.386	0.894	0.799
QUEST	85.85 %	14.15 %	0.806	0.331	0.869	0.785
CHAID	88.46 %	11.54 %	0.836	0.36	0.898	0.702
Exhaustive CHAID	88.46 %	11.54 %	0.836	0.36	0.898	0.702
(b)						
C5.0	95.52 %	4.48 %	0.884	0.457	0.942	0.838
C5.0 with 10-fold validation	95.52 %	4.48 %	0.884	0.457	0.942	0.838
CR&T	88.56 %	11.44 %	0.798	0.386	0.894	0.799
QUEST	85.85 %	14.15 %	0.806	0.331	0.869	0.785
CHAID	88.46 %	11.54 %	0.836	0.36	0.898	0.704
Exhaustive CHAID	88.05 %	11.95 %	0.8	0.375	0.897	0.69

Table 3: The association rules found in the data by the generalized rule induction (GRI) method

Antecedent	Confidence %
Fre. of Glu > 0.086 and Val > 14.500 and Fre. of Gln < 0.028	97.42
Fre. of Glu > 0.086 and Gly > 9.500 and Fre. of Gln < 0.028	96.39
Fre. of Glu > 0.086 and Hydrophobic > 81.500 and Fre. of Gln < 0.028	96.33
Fre. of Glu > 0.086 and Other > 46.500 and Fre. of Gln < 0.028	96.32
Fre. of Glu > 0.086 and Glu > 15.500 and Fre. of Gln < 0.028	96.30
Fre. of Glu > 0.086 and Negatively Charged > 23.500 and Fre. of Gln < 0.028	96.26
Fre. of Gln < 0.022 and Val > 14.500 and Fre. of Glu > 0.062	96.02
Fre. of Glu > 0.086 and Positively Charged > 20.500 and Fre. of Gln < 0.028	95.91
Fre. of Gln < 0.022 and Fre. of Glu > 0.062 and Val > 13.500	95.64
Fre. of Gln < 0.022 and Negatively Charged > 23.500 and Fre. of Glu > 0.072	95.36
Fre. of Gln < 0.022 and Other > 46.500 and Fre. of Glu > 0.070	94.80
Fre. of Gln < 0.022 and Fre. of Glu > 0.062 and Glu > 15.500	94.68
Fre. of Gln < 0.022 and Gly > 11.500 and Fre. of Glu > 0.062	94.67
Fre. of Gln < 0.022 and Positively Charged > 23.500 and Fre. of Glu > 0.068	94.17
Fre. of Gln < 0.022 and length > 153.500 and Fre. of Glu > 0.062	94.11
Fre. of Gln < 0.022 and carbon < 7.996 and Fre. of Glu > 0.060	94.01
Fre. of Gln < 0.022 and Other > 111.500 and Fre. of Glu > 0.062	93.98
Fre. of Gln < 0.022 and Hydrophobic > 82.500 and Fre. of Glu > 0.062	93.98
Fre. of Gln < 0.022 and Weight > 16.310 and Fre. of Glu > 0.062	93.44
Fre. of Gln < 0.022 and nitrogen > 193.500 and Fre. of Glu > 0.062	93.38
Fre. of Gln < 0.022 and Other > 46.500 and Fre. of Other < 0.756	93.38
Fre. of Gln < 0.022 and Fre. of Hydrophilic residues < 0.222 and Glu > 12.500	93.13
Fre. of Gln < 0.022 and Fre. of Other < 0.786 and Glu > 13.500	93.12
Fre. of Gln < 0.022 and length > 153.500 and Fre. of Other < 0.766	93.12
Fre. of Gln < 0.022 and Fre. of Negatively Charged > 0.108 and Glu > 14.500	93.09
Fre. of Negatively Charged > 0.132 and Fre. of Gln < 0.026 and Glu > 15.500	92.99
Fre. of Glu > 0.086 and Fre. of Gln < 0.028 and N-terminal = Met	92.88
Fre. of Glu > 0.086 and N-terminal = Met and Fre. of Gln < 0.028	92.88
N-terminal = Met and Fre. of Glu > 0.086 and Fre. of Gln < 0.028	92.88
Fre. of Gln < 0.022 and oxygen > 207.500 and Fre. of Glu > 0.062	92.86
Fre. of Gln < 0.022 and Fre. of oxygen > 0.088 and Glu > 15.500	92.81
Fre. of Gln < 0.022 and Hydrophilic > 28.500 and Fre. of Glu > 0.062	92.81
Fre. of Gln < 0.022 and Fre. of oxygen > 0.088 and Val > 13.500	92.70
Fre. of Gln < 0.022 and Fre. of Other < 0.786 and Other > 49.500	92.65
Fre. of Gln < 0.022 and Fre. of Negatively Charged > 0.108 and Positively Charged > 23.500	92.62
Fre. of Other < 0.736 and Fre. of Gln < 0.032 and Other > 53.500	92.60
Fre. of Other < 0.736 and Fre. of Gln < 0.032 and Negatively Charged > 23.500	92.56
Fre. of Gln < 0.022 and Fre. of Other > 0.240 and Glu > 13.500	92.45
Fre. of Glu > 0.086 and Fre. of Gln < 0.028	92.28
Fre. of Other < 0.736 and Val > 10.500 and Fre. of Gln < 0.030	92.21
Fre. of Gln < 0.022 and Fre. of Other < 0.786 and length > 160.500	92.20
Fre. of Gln < 0.022 and Hydrophobic > 82.500 and Fre. of Other < 0.786	92.17
Fre. of Gln < 0.022 and Fre. of Negatively Charged > 0.108 and Other > 50.500	92.11
Fre. of Gln < 0.022 and Fre. of Other > 0.240 and Other > 49.500	92.09
Fre. of Gln < 0.022 and Fre. of Hydrophobic < 0.580 and Glu > 13.500	92.06
Fre. of Gln < 0.022 and Weight > 16.310 and Fre. of Other < 0.772	91.99
Fre. of Gln < 0.022 and Fre. of oxygen > 0.088 and Negatively Charged > 25.500	91.98
Fre. of Gln < 0.022 and Other charge > 111.500 and Fre. Of Other < 0.786	91.92
Fre. of Gln < 0.022 and Reduced Cys Absorption at 280nm 0.1% (=1 g/l) < 1.212 and Glu > 12.500	91.83
Fre. of Gln < 0.022 and Non-reduced Cys Absorption at 280nm 0.1% (=1 g/l) < 1.218 and Glu > 12.500	91.83
Fre. of Other < 0.736 and Fre. of Gln < 0.032 and Hydrophobic > 72.500	91.75

Antecedent	Confidence %
Fre. of Gln < 0.022 and length > 153.500 and Fre. of Other > 0.244	91.69
Fre. of Gln < 0.022 and Hydrophobic > 82.500 and Fre. of Other > 0.240	91.59
Fre. of Gln < 0.022 and oxygen > 207.500 and Fre. of Other < 0.772	91.57
Fre. of Gln < 0.022 and Fre. of Other > 0.240 and length > 156.500	91.54
Fre. of Other < 0.736 and Negatively Charged > 20.500 and Fre. of Gln < 0.032	91.50
Fre. of Gln < 0.022 and Fre. of oxygen > 0.088 and Other > 52.500	91.46
Fre. of Gln < 0.022 and Fre. of oxygen > 0.088 and Hydrophobic > 89.500	91.43
Fre. of Other < 0.736 and Other > 46.500 and Fre. of Gln < 0.032	91.41
Fre. of Other < 0.736 and Fre. of Gln < 0.032 and Weight > 16.621	91.39
Fre. of Other < 0.736 and Weight > 15.882 and Fre. of Gln < 0.032	91.38
Fre. of Gln < 0.022 and Fre. of oxygen > 0.088 and length > 182.500	91.38
Fre. of Other < 0.736 and Glu > 12.500 and Fre. of Gln < 0.032	91.34
Fre. of Gln < 0.022 and Weight > 16.310 and Fre. of Other > 0.248	91.33
Fre. of Gln < 0.022 and Hydrophobic (A.F.G.I.L.M.P.V.W) > 82.500 and Fre. of oxygen > 0.088	91.32
Fre. of Other < 0.736 and Gly > 8.500 and Fre. of Gln < 0.030	91.16
Fre. of Gln < 0.022 and nitrogen > 193.500 and Fre. of Other < 0.776	91.08
Fre. of Other < 0.736 and Fre. of Gln < 0.032 and length > 146.500	91.08
Fre. of Gln < 0.022 and Aliphatic index < 111.473 and Glu > 12.500	91.02
Fre. of Other < 0.736 and Hydrophobic (A.F.G.I.L.M.P.V.W) > 63.500 and Fre. of Gln < 0.032	90.68
Fre. of Other < 0.736 and Fre. of Asn < 0.048 and Fre. of Gln < 0.036	90.68
Fre. of Gln < 0.022 and Fre. of hydrogen (H) < 0.510 and Glu > 13.500	90.64
Fre. of Gln < 0.022 and nitrogen > 193.500 and Fre. of Other > 0.242	90.62
Fre. of Gln < 0.022 and Hydrophilic > 28.500 and Fre. of Other < 0.782	90.59
Fre. of Other < 0.736 and Fre. of Negatively Charged > 0.134 and Fre. of Gln < 0.032	90.56
Fre. of Other < 0.736 and length > 131.500 and Fre. of Gln < 0.032	90.55
Fre. of Other < 0.736 and Other charge > 91.500 and Fre. of Gln < 0.032	90.47
Fre. of Gln < 0.022 and length > 153.500 and Fre. of Hydrophobic (A.F.G.I.L.M.P.V.W) < 0.578	90.45
Fre. of Gln < 0.022 and Glu > 12.500 and N-terminal = Met	90.41
Fre. of Gln < 0.022 and N-terminal = Met and Glu > 12.500	90.41
N-terminal = Met and Fre. of Gln < 0.022 and Glu > 12.500	90.41
Fre. of Negatively Charged > 0.132 and Fre. of Other < 0.736 and Fre. of Gln < 0.032	90.36
Fre. of Gln < 0.022 and Aliphatic index < 111.473 and Other > 47.500	90.35
Fre. of Other < 0.736 and Leu > 11.500 and Fre. of Gln < 0.032	90.26
Fre. of Gln < 0.022 and Negatively Charged > 23.500 and N-terminal = Met	89.97
Fre. of Gln < 0.022 and Glu > 12.500	89.93
Fre. of Other < 0.736 and Asp > 6.500 and Fre. of Gln < 0.032	89.85
Fre. of Gln < 0.022 and Negatively Charged > 23.500	89.72
Fre. of Gln < 0.022 and N-terminal = Met and Negatively Charged > 20.500	88.25
N-terminal = Met and Fre. of Gln < 0.022 and Negatively Charged > 20.500	88.25
Fre. of Gln < 0.022 and Other > 46.500 and N-terminal = Met	88.20
Fre. of Gln < 0.022 and N-terminal = Met and Other > 46.500	88.20
N-terminal = Met and Fre. of Gln < 0.022 and Other > 46.500	88.20
Fre. of Gln < 0.022 and Other > 46.500	87.96
Fre. of Negatively Charged > 0.132 and Fre. of Positively Charged < 0.259 and Fre. of Gln < 0.028	87.33
Fre. of Negatively Charged > 0.132 and Fre. of Negatively Charged < 1.634 and Fre. of Gln < 0.028	87.33
Fre. of Other < 0.736 and N-terminal = Met and Fre. of Gln < 0.032	86.24
N-terminal = Met and Fre. of Other < 0.736 and Fre. of Gln < 0.032	86.24
Fre. of Gln < 0.022 and N-terminal = Met and Fre. of oxygen > 0.088	85.86
N-terminal = Met and Fre. of Gln < 0.022 and Fre. of oxygen > 0.088	85.86

Received: August 22, 2009, accepted: August 29, 2009, published: September 01, 2009

DISCUSSION

Thermostable enzymes are of wide industrial and biotechnical interest because they are more stable and thus generally better suited for harsh processing conditions (Wakarchuk et al., 1994). The concept of thermostability is, however, not very clear, and thermostability is a relative term. Enzymatic activity is known to increase with increasing temperature up to the temperature at which inactivation starts to occur (Paloheimo et al., 2007). Thermostability is usually defined as the retention of activity after heating at a chosen temperature for a prolonged period. The most appropriate way to express thermostability is to measure the half-life of enzyme activity at elevated temperatures (Yang et al., 2007). Thermostable enzymes are produced both by thermophilic and mesophilic organisms. Although thermophilic microorganisms are a potential source for thermostable enzymes, the majority of industrial thermostable enzymes originate from mesophilic organisms (Yang et al., 2005). The successful discrimination of thermophilic proteins from mesophilic ones is an important problem, and it would help greatly in designing stable proteins. Several investigations have been conducted in an effort to understand the features that influence the stability of thermophilic proteins (Bergquist et al., 2002; Ihsanawati et al., 2005; Jaenicke & Bohm, 2001; Jiang et al., 2006; Ladenstein & Antranikian, 1998; Lo Leggio et al., 1999; Szilagyi & Zavodszky, 1995; Wu et al., 2006). An increase in the Gibbs free energy change of hydration (Gromiha et al., 1999a), and increases in the number of salt bridges and side chain-side chain interactions (Kumar et al., 2000b), aromatic clusters (Saelensminde et al., 2009), contacts between the residues of hydrogen bonds (Kumar et al., 2000a; Saraboji et al., 2005), ion pairs (Maugini et al., 2009), electrostatic interaction of charged residues (Yao et al., 2002), amino acid coupling patterns, main-chain hydrophobic free energy, and hydrophobic residues in thermophilic proteins have been show to enhance protein stability (Saraboji et al., 2005). The amino acid sequences of genomes also have been used to help understand the stability of thermophilic proteins (Liang et al., 2005). The amino acid composition on the protein surface might be an important factor that affects stability, as a specific trend was seen in the amino acid compositions in response to the requirement of stability at elevated environmental temperature (Dominy et al., 2002; Suzuki et al., 1991). The proteomes of thermophilic proteins are enriched in hydrophobic and charged amino acids at the expense of polar ones (Brouns et al., 2005).

To date, various models have been employed to determine the most important features that contribute to protein thermostability (see Introduction); here we applied different modeling techniques to study more than 70 features of some meso- and thermostable enzymes in an attempt to understand their ability to withstand higher temperatures. We used different screening, clustering, and decision tree modeling on two datasets: one with and one without feature selection filtering.

Although the results of feature selection modeling showed that 47 features had a value equal to 1, the frequency of Gln ranked as the most important feature (Table 1), and it was used in decision tree models to create the main subgroups and branches. The number of peer groups with anomalies decreased from seven (without feature selection) to two (with feature selection) groups, showing the positive effects of feature selection filtering on removing outliers. The number of clusters generated by K-Means modeling did not change between the models with and without feature selection, although the number of records in the clusters changed. In the TwoStep model, the number of clusters decreased from six (without feature selection) to just two (with feature selection) groups.

The depth of trees generated by the various decision tree models varied from 14 (in the C5.0 model with 10-fold cross-validation and with the feature selection dataset) to 4 (in the CHAID models) branches. The best performance evaluation in the decision tree models tested here was found in the C5.0 model and the worst was found in the Quest model. No significant differences in the percent of correctness, performance evaluation, and mean correctness of various decision tree models were found when feature selected datasets were used, but when feature selection datasets used the number of peer-groups in clustering models reduced significantly.

In all decision tree models, the frequency of Gln was the most important feature for decision tree rule sets, and in all GRI association rules (100 rules) the frequency of Gln was used as an antecedent to support the rules. A consistent difference exists in the pattern of synonymous codon usage between thermophilic and mesophilic prokaryotes (Farias & Bonato, 2003; Haruki et al., 2007; Van der Linden & de Farias, 2006), and there is strong evidence that this difference is the result of selection linked to thermophily (Singer & Hickey, 2003). Thermophiles and mesophiles also can be distinguished based on the amino acid composition of their proteomes, and several authors have tried to relate these differences to functional adaptation (Gromiha & Suresh, 2008; Liang et al., 2005; Singer & Hickey, 2003). Significant changes in the frequencies of some amino acids and increases in the their proportions in thermophilic organisms (with a two-fold change in the frequency of Gln) have been documented (Singer & Hickey, 2003). In another study, the residues of some amino acids (as well as Gln) showed significant difference ($p < 0.01$) between mesophilic and thermophilic proteins (Gromiha & Suresh, 2008).

We analyzed the performance of different screening, clustering, and decision tree algorithms for discriminating mesophilic and thermophilic proteins. Our results showed that amino acid composition can be used to discriminate between protein groups. We found that most of the mentioned algorithms can be used to discriminate between mesophilic and thermophilic proteins with accuracy in the range of 88–96 % in a set of 2946 proteins. Our analysis detected no significant difference in performance between different methods used in this paper. Interestingly, the CR&T, QUEST, CHAID, and exhaustive CHAID methods had a similar accuracy (~88 %), and no differences were observed between analysis with and without feature selection. The best performance and results were obtained with C5.0 algorithms. Thus, we suggest that the C5.0 decision tree model can be used as an effective tool to discriminate mesophilic and thermophilic proteins.

Acknowledgment

The authors greatly appreciate and acknowledge the support of Bioinformatics Research Groups, Green Research Center, Qom University, and the School of Agriculture at Shiraz University for supporting the project.

REFERENCES

Bergquist P, Te'o V, Gibbs M, Cziferszky A, de Faria FP, Azevedo M, Nevalainen H. Expression of xylanase enzymes from thermophilic microorganisms in fungal hosts. Extremophiles 2002;6:177-84.

Brouns SJ, Wu H, Akerboom J, Turnbull AP, de Vos WM, van der Oost J. Engineering a selectable marker for hyperthermophiles. J Biol Chem 2005;280:11422-31.

Chakkaravarthi S, Babu MM, Gromiha MM, Jayaraman G, Sethumadhavan R. Exploring the environmental preference of weak interactions in (alpha/beta)8 barrel proteins. Proteins 2006;65:75-86.

Cicortas Gunnarsson L, Montanier C, Tunnicliffe RB, Williamson MP, Gilbert HJ, Nordberg Karlsson E, Ohlin M. Novel xylan-binding properties of an engineered family 4 carbohydrate-binding module. Biochem J 2007;406:209-14.

Dancey D, Bandar ZA, McLean D. Logistic model tree extraction from artificial neural networks. IEEE Trans Syst Man Cybern B Cybern 2007;37:794-802.

Davis IW, Baker D. RosettaLigand docking with full ligand and receptor flexibility. J Mol Biol 2009;385:381-92.

De Lemos Esteves F, Gouders T, Lamotte-Brasseur J, Rigali S, Frere JM. Improving the alkalophilic performances of the Xyl1 xylanase from Streptomyces sp. S38: structural comparison and mutational analysis. Protein Sci 2005;14:292-302.

Dominy BN, Perl D, Schmid FX, Brooks CL, 3rd. The effects of ionic strength on protein stability: the cold shock protein family. J Mol Biol 2002;319:541-54.

Farias ST, Bonato MC. Preferred amino acids and thermostability. Genet Mol Res 2003;2: 383-93.

Garg A, Raghava GP. A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. In Silico Biol 2008;8:129-40.

Gromiha MM. Importance of native-state topology for determining the folding rate of two-state proteins. J Chem Inf Comput Sci 2003; 43:1481-5.

Gromiha MM. Prediction of protein stability upon point mutations. Biochem Soc Trans 2007;35:1569-73.

Gromiha MM, Selvaraj S. Inter-residue interactions in protein folding and stability. Prog Biophys Mol Biol 2004;86:235-77.

Gromiha MM, Suresh MX. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins 2008; 70:1274-9.

Gromiha MM, Yabuki Y. Functional discrimination of membrane proteins using machine learning techniques. BMC Bioinformatics 2008;9:135.

Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A. Relationship between amino acid properties and protein stability: buried mutations. J Protein Chem 1999a;18:565-78.

Gromiha MM, Oobatake M, Sarai A. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys Chem 1999b;82:51-67.

Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A. Importance of mutant position in Ramachandran plot for predicting protein stability of surface mutations. Biopolymers 2002; 64:210-20.

Hamelryck T. Probabilistic models and machine learning in structural bioinformatics. Stat Methods Med Res 2009; online first; DOI 10.1177/0962280208099492

Haruki M, Tanaka M, Motegi T, Tadokoro T, Koga Y, Takano K, Kanaya S. Structural and thermodynamic analyses of Escherichia coli RNase HI variant with quintuple thermostabilizing mutations. Febs J 2007;274:5815-25.

Hayashi H, Abe T, Sakamoto M, Ohara H, Ikemura T, Sakka K, Benno Y. Direct cloning of genes encoding novel xylanases from the human gut. Can J Microbiol 2005;51:251-9.

Huang LT, Gromiha MM, Ho SY. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. Bioinformatics 2007a;23: 1292-3.

Huang LT, Gromiha MM, Ho SY. Sequence analysis and rule development of predicting protein stability change upon mutation using decision tree model. J Mol Model 2007b;13: 879-90.

Huang LT, Saraboji K, Ho SY, Hwang SF, Ponnuswamy MN, Gromiha MM. Prediction of protein mutant stability using classification and regression tool. Biophys Chem 2007c;125: 462-70.

Ihsanawati, Kumasaka T, Kaneko T, Morokuma C, Yatsunami R, Sato T, Nakamura S, Tanaka N. Structural basis of the substrate subsite and the highly thermal stability of xylanase 10B from Thermotoga maritima MSB8. Proteins 2005;61:999-1009.

Jaenicke R, Bohm G. Thermostability of proteins from Thermotoga maritima. Methods Enzymol 2001;334:438-69.

Jiang ZQ, Li XT, Yang SQ, Li LT, Li Y, Feng WY. Biobleach boosting effect of recombinant xylanase B from the hyperthermophilic Thermotoga maritima on wheat straw pulp. Appl Microbiol Biotechnol 2006;70:65-71.

Kiarie E, Nyachoti CM, Slominski BA, Blank G. Growth performance, gastrointestinal microbial activity, and nutrient digestibility in early-weaned pigs fed diets containing flaxseed and carbohydrase enzyme. J Anim Sci 2007;85:2982-93.

Kongsted J, Ryde U, Wydra J, Jensen JH. Prediction and rationalization of the pH dependence of the activity and stability of family 11 xylanases. Biochemistry 2007;46:13581-92.

Kumar M, Gromiha MM, Raghava GP. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics 2007;8:463.

Kumar PR, Eswaramoorthy S, Vithayathil PJ, Viswamitra MA. The tertiary structure at 1.59 A resolution and the proposed amino acid sequence of a family-11 xylanase from the thermophilic fungus Paecilomyces varioti bainier. J Mol Biol 2000a;295:581-93.

Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. Protein Eng 2000b;13:179-91.

Ladenstein R, Antranikian G. Proteins from hyperthermophiles: stability and enzymatic catalysis close to the boiling point of water. Adv Biochem Eng Biotechnol 1998;61:37-85.

Liang HK, Huang CM, Ko MT, Hwang JK. Amino acid coupling patterns in thermophilic proteins. Proteins 2005;59:58-63.

Lisewski AM. Random amino acid mutations and protein misfolding lead to Shannon limit in sequence-structure communication. PLoS ONE 2008;3:e3110.

Lo Leggio L, Kalogiannis S, Bhat MK, Pickersgill RW. High resolution structure and sequence of T. aurantiacus xylanase I: implications for the evolution of thermostability in family 10 xylanases and enzymes with (beta)alpha-barrel architecture. Proteins 1999; 36:295-306.

Maugini E, Tronelli D, Bossa F, Pascarella S. Structural adaptation of the subunit interface of oligomeric thermophilic and hyperthermophilic enzymes. Comput Biol Chem 2009;33: 137-48.

Miyazaki K, Takenouchi M, Kondo H, Noro N, Suzuki M, Tsuda S. Thermal stabilization of Bacillus subtilis family-11 xylanase by directed evolution. J Biol Chem 2006;281: 10236-42.

Natesh R, Manikandan K, Bhanumoorthy P, Viswamitra MA, Ramakumar S. Thermostable xylanase from Thermoascus aurantiacus at ultrahigh resolution (0.89 A) at 100 K and atomic resolution (1.11 A) at 293 K refined anisotropically to small-molecule accuracy. Acta Crystallogr D Biol Crystallogr 2003;59: 105-17.

Paloheimo M, Mantyla A, Kallio J, Puranen T, Suominen P. Increased production of xylanase by expression of a truncated version of the xyn11A gene from Nonomuraea flexuosa in Trichoderma reesei. Appl Environ Microbiol 2007;73:3215-24.

Parthiban V, Gromiha MM, Abhinandan M, Schomburg D. Computational modeling of protein mutant stability: analysis and optimization of statistical potentials and structural features reveal insights into prediction model development. BMC Struct Biol 2007;7:54.

Received: August 22, 2009, accepted: August 29, 2009, published: September 01, 2009

Ralph SG, Chun HJ, Kolosova N, Cooper D, Oddy C, Ritland CE, Kirkpatrick R, Moore R, Barber S, Holt RA, Jones SJ, Marra MA, Douglas CJ, Ritland K, Bohlmann J. A conifer genomics resource of 200,000 spruce (Picea spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (Picea sitchensis). BMC Genomics 2008;9: 484.

Saelensminde G, Halskau O, Jr., Jonassen I. Amino acid contacts in proteins adapted to different temperatures: hydrophobic interactions and surface charges play a key role. Extremophiles 2009;13:11-20.

Saraboji K, Gromiha MM, Ponnuswamy MN. Importance of main-chain hydrophobic free energy to the stability of thermophilic proteins. Int J Biol Macromol 2005;35:211-20.

Schubert M, Poon DK, Wicki J, Tarling CA, Kwan EM, Nielsen JE, Withers SG, McIntosh LP. Probing electrostatic interactions along the reaction pathway of a glycoside hydrolase: histidine characterization by NMR spectroscopy. Biochemistry 2007;46:7383-95.

Singer GA, Hickey DA. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene 2003;317:39-47.

Suzuki Y, Hatagaki K, Oda H. A hyperthermostable pullulanase produced by an extreme thermophile, Bacillus flavocaldarius KP 1228, and evidence for the proline theory of increasing protein thermostability. Appl Microbiol Biotechnol 1991;34:707-14.

Szilagyi A, Zavodszky P. Structural basis for the extreme thermostability of D-glyceraldehyde-3-phosphate dehydrogenase from Thermotoga maritima: analysis based on homology modelling. Protein Eng 1995;8:779-89.

Thai KM, Ecker GF. Similarity-based SIBAR descriptors for classification of chemically diverse hERG blockers. Mol Divers 2009;13: 321-36.

Umemoto H, Ihsanawati, Inami M, Yatsunami R, Fukui T, Kumasaka T, Tanaka N, Nakamura S. Contribution of salt bridges to alkaliphily of Bacillus alkaline xylanase. Nucleic Acids Symp Ser (Oxf) 2007;461-2.

Van der Linden MG, de Farias ST. Correlation between codon usage and thermostability. Extremophiles 2006;10:479-81.

Wakarchuk WW, Sung WL, Campbell RL, Cunningham A, Watson DC, Yaguchi M. Thermostabilization of the Bacillus circulans xylanase by the introduction of disulfide bonds. Protein Eng 1994;7:1379-86.

Wu S, Liu B, Zhang X. Characterization of a recombinant thermostable xylanase from deep-sea thermophilic Geobacillus sp. MT-1 in East Pacific. Appl Microbiol Biotechnol 2006;72: 1210-6.

Yang HM, Yao B, Fan YL. Recent advances in structures and relative enzyme properties of xylanase. Sheng Wu Gong Cheng Xue Bao 2005;21:6-11.

Yang HM, Yao B, Meng K, Wang YR, Bai Y G, Wu NF. Introduction of a disulfide bridge enhances the thermostability of a Streptomyces olivaceoviridis xylanase mutant. J Ind Microbiol Biotechnol 2007;34:213-8.

Yao X, Nguyen V, Wriggers W, Rubenstein PA. Regulation of yeast actin behavior by interaction of charged residues across the interdomain cleft. J Biol Chem 2002;277:22875-82.

Ye X, Fu Z, Wang H, Du W, Wang R, Sun Y, Gao Q, He J. A computerized system for signal detection in spontaneous reporting system of Shanghai China. Pharmacoepidemiol Drug Saf 2009;18:154-8.