Advanced Studies in Biology, Vol. 3, 2011, no. 2, 63 - 78

Engineering Thermostable Enzymes;

Application of Unsupervised Clustering Algorithms

[1]Amir Lakizadeh, [2]Parisa Agha-Golzadeh, [3]Mansour Ebrahimi,

[2]Esmaeil Ebrahimie and [4]Mahdi Ebrahimi

[1]Department of Computer Sciences & Bioinformatics Research Group
University of Qom, Qom, Iran

[2]Department of Crop Production & Plant Breeding
College of Agriculture, Shiraz University, Shiraz, Iran

[3]Department of Biology & Bioinformatics Research Group
University of Qom, Qom, Iran
mebrahimi14@gmail.com

[4]Department of Informatics, Saarland University
Saarbrucken, Germany

ABSTRACT

There is a high demand for engineering thermostable enzymes in some industries; especially in paper industries to use environmental friendly enzymes instead of toxic chlorine chemicals. Hence, understanding protein attributes involved in enzyme thermostability is important. Herein, the most important protein features contributing to enzyme thermostability was searched by using data mining algorithms. Combination of attribute weighting and unsupervised clustering algorithms were used to explore protein attributes which play major roles in thermostability. The results showed that expectation maximization clustering with uncertainly and correlation attribute weighting algorithms can effectively (100%) classify thermo- and meso-stable proteins. Gln content and frequency of hydrophilic residues were the most important protein features selected by 70% of weighing methods. The findings of this research provide the required knowledge for engineering thermostable enzymes in laboratory.

INTRODUCTION

Enzyme thermostability, an intrinsic property mainly determined by the primary structure of the protein, increases by external environmental factors including cations, substrates, co-enzymes and modulators. With some exceptions, enzymes present in thermophiles are more stable than their mesophilic counterparts. However, most thermostable enzymes on the market have been derived from mesophiles. Further research, especially on cloning enzymes from thermophiles into mesophilic hosts, will greatly increase the exploitation of thermophiles in biotechnology. Thermostable enzymes are receiving considerable attention. These enzymes have many industrial applications already, because they are more stable and generally better suited to harsh processing conditions.

Understanding the factors responsible for enzyme thermostability and discriminating them from mesophilic proteins is one of the most important duties in engineering new proteins. Various methods have been proposed for predicting the stability of proteins based on amino acid substitutions. In some studies, mutants were reported to be more important than the actual magnitude of stability [1]. It has been shown that Intra-helical salt bridges are more prevalent in thermostable enzymes, and the composition of amino acids might be an important factor in the stability. Moreover, hydrophobic and charged amino acids are more prevalent in thermophilic proteins [2]. Although the issue of thermostability in proteins has gained attentions in recent years due to high demand from industries, need for a system which derives stability rules for any input data and converts them into a prediction has been elaborated.

Data mining problems often involve hundreds, or even thousands, of variables [3], when the number of variables are huge, more time may be needed to apply a neural network or a decision tree to such a dataset [4]. Usually, many attributes determine the different characteristics of a protein molecule [5, 6]. Therefore, majority of time and process should be spent to determining which variables to include in the model. Attribute weighting (or feature selection) models reduce the size of attributes, creating a more manageable set of attributes for modelling [7].

On the other hand, the objective of clustering algorithms is data partitioning into groups or clusters according to various criteria in an attempt to organize data into a more meaningful form [8]. Clustering, often used as a synonym for unsupervised learning, may proceed according to some parametric model or by grouping points according to some distance or similarity measure as in hierarchical clustering algorithms [9]. It is anticipated that a suitable, unsupervised algorithm is capable of discovering structure on its own by exploring similarities or differences between individual data points in a data set under consideration [10]. K-Means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids,

one for each cluster [11]. K-Medoids method uses representative objects as reference points instead of taking the mean value of the objects in each cluster [12]. The support vector clustering (SVC) algorithm is a recently emerged unsupervised learning method inspired by support vector machines. One key step involved in the SVC algorithm is the cluster assignment of each data point [13]. The Expectation-Maximization Clustering (EMC) is an effective, popular technique for estimating mixture model parameters (cluster parameters and their mixture weights). The EM algorithm iteratively refines initial mixture model parameter estimates to better fit the data and terminates at a locally optimal solution [14].

To find important features contributing to enzyme thermostability, herein, we used various attribute weighting algorithms and four different unsupervised clustering models to determine the most important features responsible for thermostability.

MATERIALS & METHODS

Two thousands and ninety protein sequences were retrieved from the UniProt Knowledgebase (Swiss-Prot and Tremble) database. They were categorized into two groups: 1573 or 75% to T (optimum temperature < 70 ºC) and 517 or 25% to F (optimum temperature > 70ºC, thermostable enzymes). Eight-hundred and fifty two protein attributes or features such as length, weight, isoelectric point, count and frequency of each element (carbon, nitrogen, sulphur, oxygen, and hydrogen), count and frequency of each amino acid, count and frequency of negatively charged, positively charged, hydrophilic and hydrophobic residues, count and frequency of dipeptides, number of α-helix and β-strand, and other secondary protein features were extracted. All features were classified as continuous variables, except for optimum tempreature and N-terminal amino acids which were classified as categorical. A dataset of these protein features was imported into Rapid Miner (RapidMiner 5.0.001, Rapid-I GmbH, Stochumer Str. 475, 44227 Dortmund, Germany) and the optimum temperature (categorized as T and F) was set as target or label attribute.

Afterwards, the following steps applied on dataset:

1. DATA CLEANSING

Duplicate features were omitted by comparing all examples with each other on the basis of the specified selection of attributes (two examples were assumed equal if all values of all selected attributes were equal). Then, useless attributes were removed from the dataset. Nominal attributes were regarded as useless when most frequent values were contained in more or less than nominal useless above or below percent of all examples. Numerical attributes which standard deviations were less or equal to a given deviation threshold (0.1). Finally correlated features (with correlation greater than 0.9) removed by using Pearson correlation function. So the number of attributes decreased to 794.

2. ATTRIBUTE WEIGHTING

To identify the most important features and to find the possible patterns that contribute to the types of thermostable enzymes, the following attribute weightings were applied on cleansed dataset as follows:

- *Weight by Information gain:* this operator calculated the relevance of a feature by computing the information gain in class distribution.
- *Weight by Information Gain ratio:* this operator calculated the relevance of a feature by computing the information gain ratio for the class distribution.
- *Weight by Rule:* this operator calculated the relevance of a feature by computing the error rate of a OneR Model on the example set without this feature.
- *Weight Deviation:* the operator created weights from the standard deviations of all attributes. The values normalized by the average, the minimum, or the maximum of the attribute
- *Weight by Chi squared statistic:* This operator calculated the relevance of a feature by computing for each attribute of the input example set the value of the chi-squared statistic with respect to the class attribute.
- *Weight by Gini index:* This operator calculated the relevance of an attribute by computing the Gini index of the class distribution, if the given example set would have been split according to the feature.
- *Weight by Uncertainty:* This operator calculated the relevance of an attribute by measuring the symmetrical uncertainty with respect to the class.
- *Weight by Relief:* This operator measured the relevance of features by sampling examples and comparing the value of the current feature for the nearest example of the same and of a different class. This version also worked for multiple classes and regression data sets. The resulting weights were normalized into the interval between 0 and 1.

- *Weight by SVM (Support Vector Machine):* This operator used the coefficients of the normal vector of a linear SVM as feature weights.

- *Weight by PCA (Principle Component Analysis):* This operator used the factors of the first of the principal components as feature weights.

3. ATTRIBUTE SELECTION

After attribute weighting models were ran on the dataset, each protein attribute or feature gained a value between 0 to 1; showing the importance of that attribute regarded to target attribute (optimum temperature of enzymes). All variables with weights higher than 0.50 were selected and 10 new datasets (with 2057 records in each dataset) created.

4. UNSUPERVISED CLUSTERING ALGORITHMS

The following clustering algorithms applied on 10 newly created datasets from attribute weighing:

- *K-Means:* This operator uses kernels to estimate distance between objects and clusters. Because of the nature of kernels, it is necessary to sum over all elements of a cluster to calculate one distance.
- *K-Medoids:* This operator represents an implementation of k-Medoids. This operator will create a cluster attribute if not present yet.
- *Support Vector Clustering (SVC):* An implementation of Support Vector Clustering based on. This operator will create a cluster attribute if not present yet.
- *Expectation Maximization (EM):* This operator represents an implementation of the EM-algorithm

RESULTS

The first dataset had 2090 records with 852 protein attributes; 75% of them (1573 records) classified as T class and the rest (517 or 25% of records) as F class. Data cleansing (removing duplicates, useless attributes and removing correlated features) reduced the numbers of records and features to 2057 and 794, respectively.

1. ATTRIBUTE WEIGHTING

Before running the models, data were normalized, so it would be reasonable to expect that all weights should be between 0 to 1.

- *Weighting by PCA*

Nine attributes weighed equal to or higher than 0.80; they were the counts of Asp, Glu, Phe, His, Met, Ser, Val and Tyr.

- *Weighting by SVM*

The frequencies of hydrophilic residues, Gly, Asn, Tyr, Asp – Pro, Glu – Ile, Glu – Gln, Met – Leu, Met –Thr, Asn – Asn, Pro – Tyr, Thr – Lys, Trp – Leu and Tyr – Val, the counts of Gln, Glu – Gln, Lys – Gln, Asn – Asn, Trp – Leu and frequency of other residues and the percentages of Thr and Val were 27 attributes selected by this model.

- *Weighting by Relief*

When this model applied on dataset, 7 attributes showed weights higher than 0.50. The counts of Met – Gln and the frequency of other residues, the percentage of Gln and the frequencies of hydrophilic residues, Asn, Asn – Ile and Gln – Leu were selected protein attributes.

- *Weighting by Uncertainty*

The frequency of hydrophilic residues, the count of other residues, the count of Gln, the frequency of Asn, the frequency of Arg, the percentage of Glu, the percentage of Gln, the count of Asn – Asn, the count of Gln – Asn, the frequency of Asn – Asn and the frequency of Gln – Asn were the protein attributes gained weighs higher than 0.50.

- *Weighting by Gini index*

Again the count of Met – Tyr was the only attributes weighed equal to 1.00. Three, one, four, six and seven protein attributes weighed equal to or higher than 0.90, 80, 0.70, 0.60 and 0.50, respectively (Table 1).

- *Weighting by Chi Squared*

Just five attributes weighed higher than 0.50. The frequencies of hydrophilic residues, and Asn, the count of other residues and the percentages of Glu and Gln were selected by this model.

- *Weighting by Deviation*

Ten counts of dipeptides (His – Cys, His –Trp, Met – Cys, Met – Trp, Gln – Cys, Gln –Trp, Trp – Cys, Trp – His, Trp – Met and Trp – Trp) and the frequencies of seventeen dipeptides (His – Cys, His – Trp, Lys – Cys, Met – Cys, Met – Trp, Gln – Cys, Gln – Trp, Arg – Cys, Trp – Cys, Trp – His, Trp – Met, Trp – Pro, Trp – Gln, Trp – Thr, Trp – Trp, Trp – Tyr and Tyr – Cys) were 27 attributes weighed higher than 0.50.

- *Weighting by Rule*

Aliphatic index, non-reduced absorption at 280 nm, the percentage of Thr, the frequency of Asn – Asn and the counts of Asn – Asn, Pro – Gln and Gln – Asn weighed equal to or higher than 0.50 when rule algorithm run on dataset.

- *Weighting by Correlation*

From 11 protein attributes weighed equal to or higher than 0.50, the frequency of hydrophilic, the count of other residues, the count of Gln, the frequency of Asn, the frequency of Arg, the percentage of Glu and Gln, the count of Asn – Asn, the count of Gln – Asn, the frequency of Asn – Asn and the frequency of Gln – Asn were the selected protein attributes.

- *Weighting by Information Gain*

Just three features, the frequency of Glu, the percentage of Asn and the percentage of Val gained weights equal to or higher than 0.50, when this algorithm applied on dataset.

2. UNSUPERVISED CLUSTERING ALGORITHMS

As mentioned before, four different algorithms (K-Means, K-Medoids, SVC and EMC) applied on ten datasets created by attribute selection algorithms. As seen in Table 1. Some models such as EMC algorithm on Chi squared, Gini Index, Information Gain, Relief, Rule and PCA were unable to identify T enzyme from F enzyme (all proteins were selected as F class). EMC algorithm on Deviation dataset was unable to cluster any protein into the right classes. Some other algorithms, such as K-Medoids on Chi Squared dataset, identified most of proteins from T class as f class. Just two clustering methods, uncertainly and correlation clustered were able to categorize the most attributes into the right clusters; selecting the right classes of T and F (1544 and 513, respectively).

DISCUSSION

Engineering thermostable enzymes are of wide industrial and biotechnical interest, because they are more suited to harsh conditions and therefore they are appointed as one of the best candidates in enzyme engineering [15]. In thermostable enzymes, the activities are known to increase with increasing temperature up to the temperature at which inactivation starts to occur [16]. Finding a successful method to discriminate between thermophilic and mesophilic proteins is an important problem and it would help greatly in designing stable proteins [17]. Different models have been proposed to determine the most important attributes that contribute to stability of proteins at higher temperatures such as crystal structure of thermostable proteins [18], logistic model tree extraction [19], mutant position [1], machine learning algorithms [20], characteristic patterns of codon usage, amino acid composition and nucleotide content [21], disulphide bridge [15], analyses of three-dimensional structures [22], salt bridges [23], aromatic interactions [24], content of Arg, Pro, His, Try [25], the isoelectric points [26], hydrophobicity [27] and the content of electric charges [2]. To determine the most important feature contributing to stability of proteins in harsh thermal conditions, various modelling techniques applied to study more than eight hundreds attributes of mesophilic and thermophilic proteins.

When the numbers of variables or attributes are large enough, the abilities of processing units reduce significantly. Data cleansing algorithms were used to remove correlated, useless or duplicated attributes which results in shrinked database [28]. About 10% of unimportant attributes discarded when these algorithms applied on the original dataset.

Each attribute weightings uses specific pattern to define the most important features; so the results could be different [29] and the same has been highlighted in previous studies [30, 31]. The frequency of hydrophilic residues, the percentage of Gln, and the count of other residues were the most important feature defined by 70% of attribute weighting algorithms to distinguish between meso- and thermostable enzymes. This finding is in-line with previous reports and confirms the importance of hydrophilicity in compacting the enzymes and increasing their capacities to resist higher temperatures [32]. It has also been shown the thermostable dehydrins in plant mitochondria are highly hydrophilic and their accumulation in some species such as wheat and rye induce more heat-stable capacity. It has been supposed these proteins' hydrophilicity character stabilize proteins in the membrane or in the matrix when heat increase [33]. Site-directed mutagenesis has been used to understand thermostable capacity of xylanase enzymes; confirming the importance of hydrophilicity and salt- bridge [23]. Although the importance of hydrophobicity (not hydrophilicity) have been highlighted in some studies [20], in a prominent study in this field, the hydration entropy introduced as the major contributor to the stability of surface mutations in

helical segments; and confirmed the inverse hydrophobic effect was generally applicable only to coil mutations [1]. An unusually large proportion of surface ion-pairs involved in networks that cross-link sequentially separate structures on the protein surface, and an unusually large number of solvent molecules buried in hydrophilic cavities between sequentially separate structures in the protein core of thermostable proteins have been reported [18]. Increase in the number of polar amino acids such as the count of Gln (as confirmed in this study) may contribute to increase in hydrophilic properties of thermostable proteins.

Recent studies have shown that amino acid composition, especially some polar amino acids such as Gln, exert distinguishable effect on thermostability [21]. The most dramatic effect was a two-fold decrease in the frequency of Gln residues among thermophiles [34]. In another study [35], it has been shown that the frequency of Gln is playing an important role in enzyme thermostability and our results confirmed it again [36]. Large number of polar molecules which may increase molecular hydrophilicity has been found in crystal structure of thermostable beta-glycosidase [18]. It has been proposed that amino acid substitution (toward polar amino acids such as Gln) changes protein stability in harsh thermal conditions can be useful for protein engineering in designing novel proteins with increased stability and altered function [1]. The same has been noted that in thermophiles charged residues such as Glu as well as the hydrophobic residues have higher occurrence than mesophiles [20]. A direct correlation of higher optimum pH with an increase in the number of electrically charged residues such as Arg has been observed [25]. This may justify why the count of Gln as well as other residues have gained higher weights in attribute weighting algorithms applied in this study. The roles of other protein residues in enzymes thermostability have been reported before [37, 38]. Therefore, thermostable proteins and enzymes also can be distinguished from meso-stable proteins based on their proteomes amino acid composition, as several authors related these differences to functional adaptation [21]. It has been shown that significant changes in the frequencies of some amino acids and increases in the their proportions in thermostable enzymes up to two-fold change in the frequency of Gln can be traced [21]. Also the residues of some amino acids (as well as Gln) showed significant differences ($p < 0.01$) between meso-stable and thermos-stable enzymes [20].

Clustering algorithms have been widely used in various areas of biological sciences such as image processing and diagnostics [39], EST [10], cancer detection [40], promoter analysis [10], gene and protein bioinformatics [23], and so on [8]. Herein we applied four different unsupervised clustering (K-Means, K-Medoids, SVC and MEMC) on 10 datasets created from protein attributes gained higher weights. As seen in Table 1, the performances of these algorithms were significantly different. Some were unable to select even one T protein into the right class (as EMC algorithm on most datasets, except Correlation and Uncertainty). Some others could not put F proteins into their classes (such as SVC

algorithm on Deviation dataset). The results showed EMC algorithm was able to classify T and F proteins into the right classes when it runs on Correlation and Uncertainty datasets. The numbers of proteins in each class were exactly the same as original dataset, showing 100% performances and accuracies of this algorithm on the mentioned datasets. To our knowledge, this is the first report on applying these algorithms to classify thermo- and meso-stable proteins. The current findings add to a growing body of literature on engineering new thermostable enzymes which in urgent needs by industries. The methods used in this study may be applied to other proteins to explore more attributes contribute to enzymes' thermostability.

Here the various attribute weightings and clustering employed to find a suitable model to clearly discriminate between mesophilic and thermophilic proteins. The results showed some amino acid compositions (such as the count of Gln and the frequency of hydrophilic and other residues) can be used to differentiate between enzymes from meso- and thermo-stable groups. The combination of attribute weighting and clustering algorithms can effectively be used to classify thermo- and meso-stable enzymes. The findings may pave road to engineer new thermostable proteins.

REFERENCES

[1] M. M. Gromiha, M. Oobatake, H. Kono, H. Uedaira and A. Sarai, Importance of mutant position in Ramachandran plot for predicting protein stability of surface mutations, Biopolymers, 64(2002), 210-20.

[2] H. M. Yang, B. Yao and Y. L. Fan, Recent advances in structures and relative enzyme properties of xylanase, Sheng Wu Gong Cheng Xue Bao, 21(2005), 6-11.

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

[4] M. M. Gromiha and Y. Yabuki, Functional discrimination of membrane proteins using machine learning techniques, BMC Bioinformatics, 9(2008), 135.

[5] E. Ebrahimie, M. Ebrahimi and M. Rahpayma, Investigating protein features contribute to salt stability of halolysin proteins, Journal of Cell and Molecular Research, 2(2010), 15-28.

[6] E. Ashrafi, A. Alemzadeh, M. Ebrahimi, E. Ebrahimie, N. Dadkhodaei and M. Ebrahimi, Determining specific amino acid features in P1B-ATPase heavy metals transporters which provides a unique ability in small number of organisms to cope with heavy metal pollution Bioinformatics and Biology Insights, Accepted(2011),

[7] K. M. Thai and G. F. Ecker, Similarity-based SIBAR descriptors for classification of chemically diverse hERG blockers, Mol Divers, 2009),

[8] G. J. McLachlan, R. W. Bean and S. K. Ng, Clustering, Methods Mol Biol, 453(2008), 423-39.

[9] Y. Lin, G. C. Tseng, S. Y. Cheong, L. J. Bean, S. L. Sherman and E. Feingold, Smarter clustering methods for SNP genotype calling, Bioinformatics, 24(2008), 2665-71.

[10] T. Abeel, Y. Saeys, P. Rouze and Y. Van de Peer, ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles, Bioinformatics, 24(2008), i24-31.

[11] S. A. Billings, H. L. Wei and M. A. Balikhin, Generalized multiscale radial basis function networks, Neural Netw, 20(2007), 1081-94.

[12] M. E. Sparks and V. Brendel, MetWAMer: eukaryotic translation initiation site prediction, BMC Bioinformatics, 9(2008), 381.

[13] V. De Bruyne, F. Al-Mulla and B. Pot, Methods for microarray data analysis, Methods Mol Biol, 382(2007), 373-91.

[14] S. Waydo and C. Koch, Unsupervised learning of individuals and categories from images, Neural Comput, 20(2008), 1165-78.

[15] W. W. Wakarchuk, W. L. Sung, R. L. Campbell, A. Cunningham, D. C. Watson and M. Yaguchi, Thermostabilization of the Bacillus circulans xylanase by the introduction of disulfide bonds, Protein engineering, 7(1994), 1379-86.

[16] M. Paloheimo, A. Mantyla, J. Kallio, T. Puranen and P. Suominen, Increased production of xylanase by expression of a truncated version of the xyn11A gene from Nonomuraea flexuosa in Trichoderma reesei, Applied Environmental Microbiology, 73(2007), 3215-24.

[17] M. W. Adams and R. M. Kelly, Finding and using hyperthermophilic enzymes, Trends Biotechnol, 16(1998), 329-32.

[18] C. F. Aguilar, I. Sanderson, M. Moracci, M. Ciaramella, R. Nucci, M. Rossi and L. H. Pearl, Crystal structure of the beta-glycosidase from the hyperthermophilic archeon Sulfolobus solfataricus: resilience as a key factor in thermostability, J Mol Biol, 271(1997), 789-802.

[19] D. Dancey, Z. A. Bandar and D. McLean, Logistic model tree extraction from artificial neural networks, IEEE Trans Syst Man Cybern B Cybern, 37(2007), 794-802.

[20] M. M. Gromiha and M. X. Suresh, Discrimination of mesophilic and thermophilic proteins using machine learning algorithms, Proteins, 70(2008), 1274-9.

[21] G. A. Singer and D. A. Hickey, Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content, Gene, 317(2003), 39-47.

[22] O. Bogin, M. Peretz, Y. Hacham, Y. Korkhin, F. Frolow, A. J. Kalb and Y. Burstein, Enhanced thermal stability of Clostridium beijerinckii alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic Thermoanaerobacter brockii alcohol dehydrogenase, Protein Sci, 7(1998), 1156-63.

[23] J. Georis, F. de Lemos Esteves, J. Lamotte-Brasseur, V. Bougnet, B. Devreese, F. Giannotta, B. Granier and J. M. Frere, An additional aromatic interaction improves the thermostability and thermophilicity of a mesophilic family 11 xylanase: structural basis and molecular study, Protein Sci, 9(2000), 466-75.

[24] Y. Asada, S. Endo, Y. Inoue, H. Mamiya, A. Hara, N. Kunishima and T. Matsunaga, Biochemical and structural characterization of a short-chain dehydrogenase/reductase of Thermus thermophilus HB8: a hyperthermostable aldose-1-dehydrogenase with broad substrate specificity, Chem Biol Interact, 178(2009), 117-26.

[25] A. Masui, N. Fujiwara and T. Imanaka, Stabilization and rational design of serine protease AprM under highly alkaline and high-temperature conditions, Appl Environ Microbiol, 60(1994), 3579-84.

[26] S. Berens, H. Kaspari and J. H. Klemme, Purification and characterization of two different xylanases from the thermophilic actinomycete Microtetraspora flexuosa SIIX, Antonie Van Leeuwenhoek, 69(1996), 235-41.

[27] K. Miyazaki, M. Takenouchi, H. Kondo, N. Noro, M. Suzuki and S. Tsuda, Thermal stabilization of Bacillus subtilis family-11 xylanase by directed evolution, Journal of Biological Chemistry, 281(2006), 10236-42.

[28] G. Rustici, M. Kapushesky, N. Kolesnikov, H. Parkinson, U. Sarkans and A. Brazma, Data storage and analysis in ArrayExpress and Expression Profiler, Curr Protoc Bioinformatics, Chapter 7(2008), Unit 7 13.

[29] C. Baumgartner, G. D. Lewis, M. Netzer, B. Pfeifer and R. E. Gerszten, A new data mining approach for profiling and categorizing kinetic patterns of metabolic biomarkers after myocardial injury, Bioinformatics, 26(2010), 1745-51.

[30] M. Ebrahimi and E. Ebrahimie, Sequence-based prediction of enzyme thermostability through bioinformatics algorithms, Current Bioinformatics, 5(2010), 195-203.

[31] E. Bijanzadeh, Y. Emam and E. Ebrahimie, Determining the most important features contributing to wheat grain yield using supervised feature selection model, Australian Journal of crop science, 4(2010), 402-407.

[32] P. A. Barthelemy, H. Raab, B. A. Appleton, C. J. Bond, P. Wu, C. Wiesmann and S. S. Sidhu, Comprehensive analysis of the factors contributing to the stability and solubility of autonomous human VH domains, J Biol Chem, 283(2008), 3639-54.

[33] G. B. Borovskii, I. V. Stupnikova, A. I. Antipina, S. V. Vladimirova and V. K. Voinikov, Accumulation of dehydrin-like proteins in the mitochondria of cereals in response to cold, freezing, drought and ABA treatment, BMC Plant Biol, 2(2002), 5.

[34] S. Akanuma, C. Qu, A. Yamagishi, N. Tanaka and T. Oshima, Effect of polar side chains at position 172 on thermal stability of 3-isopropylmalate dehydrogenase from Thermus thermophilus, FEBS Lett, 410(1997), 141-4.

[35] M. Ebrahimi, E. Ebrahimi and M. Ebrahimi, Searching for patterns of thermostability in proteins and defining the main features contributing to enzyme thermostability through screening, clustering, and decision tree algorithms, EXCLI Journal, 8(2009), 218-233.

[36] A. Vitalis, N. Lyle and R. V. Pappu, Thermodynamics of beta-sheet formation in polyglutamine, Biophys J, 97(2009), 303-11.

[37] L. Bendova-Biedermannova, P. Hobza and J. Vondrasek, Identifying stabilizing key residues in proteins using interresidue interaction energy matrix, Proteins, 72(2008), 402-13.

[38] M. M. Gromiha, Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins, Biophys Chem, 91(2001), 71-7.

[39] D. Balasubramanian, P. Srinivasan and R. Gurupatham, Automatic classification of focal lesions in ultrasound liver images using principal component analysis and neural networks, Conf Proc IEEE Eng Med Biol Soc, 2007(2007), 2134-7.

[40] M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir and A. Schliep, Clustering cancer gene expression data: a comparative study, BMC Bioinformatics, 9(2008), 497.

Table 1. Clustering of 10 datasets into T and F classes by four different clustering algorithm (K-Means, K-Medoids, SVC and EMC); the actual numbers of T and F classes in original datasets were 1544 and 512, respectively.

	Chi Squared		Correlation		Deviation		Gini Index		Information Gain	
	T	F	T	F	T	F	T	F	T	F
K-Means	1461	596	1810	247	1222	835	1452	605	1333	724
K-Medoids	487	1570	1521	536	104	1953	1570	487	1152	905
SVC	363	1688	1701	328	1705	6	363	1688	570	1487
EMC	0	2057	1544	513	0	0	0	2057	0	2057

(comtinued from above)

Relief		Rule		PCA		SVM		Uncertainty	
T	F	T	F	T	F	T	F	T	F
1603	454	1601	456	434	1623	1076	981	372	1685
583	1474	1652	405	1768	289	892	1165	939	1118
529	1324	1561	4	631	1426	0	2057	1089	947
0	2057	4	2053	0	2057	0	2057	1544	513

Table 2. The most important protein attributes (features) selected by attribute weighting algorithms, # represents the number of algorithm selected this feature.

Attribute	# Repeat
The percentage of Gln	7
The frequency of hydrophilic residues	7
The count of other residues	7
The percentage of Glu	6
The frequency of Asn	6
The frequency of Asn-Asn	4
The count of Asn-Asn	4
The count of Gln	3
The Percentage of Thr	3
The count of Gln-Asn	3
The percentage of Val	2
The frequency of Arg	2
The count of Pro-Gln	2
The count of Lys-Gln	2

Received: March, 2011