ELSEVIER

Process Biochemistry 41 (2006) 552–556

PROCESS BIOCHEMISTRY

www.elsevier.com/locate/procbio

Short communication

Discrimination of thermophilic and mesophilic proteins via pattern recognition methods

Guangya Zhang, Baishan Fang *

Institute of Industrial Biotechnology, Huaqiao University, Quanzhou 362021, Fujiang, P.R. China

Received 5 May 2005; received in revised form 5 September 2005; accepted 6 September 2005

Abstract

Four pattern recognition methods, namely, principal component analysis (PCA), stepwise regression (SR), partial least-square regression (PLSR), and backpropagation neural network, were used to discriminate thermophilic and mesophilic proteins. And four models were made to classify between these two kinds of proteins. To some degree the prediction accuracy of the methods was encouraging except for principal component analysis. Results showed that the average fitting accuracy of the four methods was 92%, 96%, 95% and 98%, respectively. And the average prediction reliability was 60%, 67.5%, 72.5% and 72.5%, respectively, the best prediction reliability for thermophilic proteins was 75%, and for mesophilic proteins was 85%.

Keywords: Protein thermostability; Sequence–characteristic relationship; Principal component analysis; Stepwise regression; Partial least-square regression; Backpropagation neural network

1. Introduction

Protein thermostability has been vigorously studied in the biophysical and biotechnological research areas [1,2], because protein instability at high temperature is one of the main bottlenecks in extending the application of protein [3]. The properties of thermophilic proteins have also been examined extensively. In particular, it has been investigated whether thermophily can be detected at the amino acid level. Such studies have detected some preferences of thermophilic proteins for particular amino acids, but general rules have not yet emerged [4].

Pattern recognition may be summarized as the categorization of some input data into identifiable classes via the extraction of significant features or attributes of the data from a background of irrelevant details. Pattern recognition was very much an interdisciplinary subject, covering developments in the areas of statistics, artificial intelligence, computer science, psychology, and physiology, among others. It has a large numbers of applications, ranging from the classical ones such as automatic character recognition and medical diagnosis to the more recent ones in data mining [5]. But the application in the classification of thermophilic and mesophilic proteins has not been reported before.

In this work, based on principal component analysis (PCA), stepwise regression (SR), partial least-square regression (PLSR) and principal component artificial neural networks (PC-ANN), the amino acid contents were computed and used as factors to discriminate thermophilic proteins from mesophilic proteins, the results were encouraging; some discriminating models were established and the biological meaning of them were expatiated on and the real difference of amino acid components between these two kinds of proteins were found.

2. Materials and methods

2.1. Data set construction

Seventy-six pairs set of thermophilic and mesophilic proteins for training were downloaded from Swiss-Prot (http://au.expasy.org), for the reason of its non-redundancy. The training sample was constructed as following procedures. At first, all the thermophilic proteins were investigated in Swiss-Prot database using "thermo" and "pyro" as search keywords. Secondly, the sequences with the annotation of putative, probable or fragment were eliminated. Thirdly, only one protein sequence was chosen among the protein sequences with the same name. Fourthly, each name of the selected thermophilic proteins was used as the

Abbreviations: PCA, principal component analysis; SR, stepwise regression; PLSR, partial least-square regression; PC-ANN, principal component artificial neural networks; DPS, data processing system; LVs, latent variables; CPU, central processing unit; PCs, principal components

* Corresponding author. Fax: +86 595 269 1560.
E-mail address: zhgyghh@hqu.edu.cn (B. Fang).

1359-5113/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.procbio.2005.09.003

Table 1
The accession number of the training samples

	Thermophilic protein	Mesophilic protein
Training sample	Q9V0L2, O93730, Q9UY47, Q58549, Q9V2I6, O58362, Q8ZVE4, Q9V1I5, Q9V1I6, Q8ZUA0, Q8ZV07, Q8ZU95, Q8ZW80, Q8ZW90, Q8ZW59, Q8U0A6, Q8U0A5, Q8ZU97, P81413, Q9X1B7, Q9HIY2, O58111, P95474, Q47950, Q9HHC4, P77916, Q8DL74, O59605, Q9V1P1, Q9UXW3, Q9HHB6, Q9V0T9, Q9WY82, Q8ZTZ0, O58097, Q8TZI8, Q8U039, Q9UYR1, Q8ZY36, Q8TH25, Q9V1R3, Q9YB30, Q9UZ09, Q8ZYU6, P19514, Q8ZZX3, Q8U0F3, Q8ZU24, Q8ZW35, Q8U0C0, O32450, Q8ZVB2, O59488, Q8U381, Q9WY74, Q8TZL3, O57765, Q8RBA4, Q8U4I9, Q8U3K8, O58429, Q8U263, Q51742, O58665, Q9V0N0, P58202, Q8U0G6, P61883, Q8U111, O58050, O57979, Q9UY56, Q8ZZK5, Q8U3Z2, Q8U491, Q8U4A0	P53001, P36333, Q9VF36, P54570, P71295, P53582, Q82XS4, P68729, Q88QQ6, P59308, P36839, Q8G5F3, P31102, P63609, P05194, P43904, P34003, O34347, O89033, O04928, Q8FC88, P36561, P37306, Q43314, P44121, P21189, P77488, P40370, O34425, P11537, P56091, Q97FQ7, Q9KRB5, P14742, Q9LVI8, Q8FBC3, Q9HTE9, Q91Z53, P60757, Q82WM3, P14891, Q38929, P05793, Q9I6E0, P00817, P29364, P46086, P32895, P09151, P30127, P15977, P22133, P17109, P59286, P52085, P17443, P10902, Q9I4W9, Q9HX21, Q9HUP3, Q99JR6, P39207, P68739, Q8FAE1, Q9I3C3, P38787, Q9NXJ5, Q8KEX0, P00558, P32662, P35558, P00496, Q05728, Q9UUB4, Q9Y0Y2, P35421

keyword to search in the same database to find its counter part mesophilic protein. Finally, if the counter part mesophilic protein cannot find in Swiss-Prot, then the thermophilic protein was eliminated. We each selected 76 thermophilic and mesophilic proteins, respectively, as base data set to make discriminating models based on pattern recognition algorithms (see Table 1).

In order to check the accuracy of each model, 20 pairs set of thermophilic and mesophilic proteins were selected as testing dataset. The dataset for testing came from reference [6]. The sequences of the testing sample were found in Protein Data Bank (PDB) via the PDB code provided by reference [6]. These 20 pairs set of thermophilic and mesophilic proteins were different from the proteins used in the training sample. All the proteins for training and testing were presented in Table 1 and Table 2.

Bioedit was used to calculate the contents of 20 amino acids in each sequence, and the contents of residues were used as variant to calculate principal components (PCs) by the principal component analysis of SPSS10.0. Stepwise regression, partial least-square regression and principal component artificial neural networks were performed by data processing system (DPS) [7]. The plots were constructed by Origin 7.0.

2.2. Principal component analysis

Principal component analysis is a linear analysis technique that finds the most efficient representation (in the least-square sense) of a data set in several dimensions. PCA is best employed as a tool to reduce the dimensionality of a set of data. This reduction of dimensionality can allow one to visualize a multivariable data set more easily, and to employ traditional statistical methods that might otherwise be impossible to use. For more detailed information please see reference [8].

2.3. Partial least-square regression

Partial least-square regression is conceptually similar to PCA, except that it reduces the dimensions of two sets of data (an $m \times n$ X input data set X and an $m \times n$ Y output data set Y) simultaneously, finding the directions (latent variables, LVs) in the input space that are most predictive for the output space. A detailed description of the PLSR algorithm and its mathematical formulation are provided by Geladi and Kowalski [9].

2.4. Principal component artificial neural networks

Since the high number of nodes in the input layer of the network (i.e. number of amino acid for each sequence) increases the central processing unit (CPU) time for ANN modeling, and to decrease the redundancy existed in the descriptors data, the data matrix was firstly analyzed by principal component analysis. The principal components, which can explain more than 85% of variances in the original descriptors data matrix, were selected and used as the input variables of the ANN models. A feed-forward neural network with backpropagation of an error algorithm was used for modeling. Our network has one input layer, one hidden layer, and one output layer. The input layer was the scores of the PCs, and the output layer was the type of the protein and had

Table 2
The predicted results of the four methods

PDB ID	Actual type	Predicted type			
		PCA	SR	PLS	PC-ANN
1zin	T	T	M	M	M
1tmy	T	T	T	T	T
1aj8	T	T	T	T	T
1tfe	T	T	T	T	T
1yna	T	M	M	M	M
1gtm	T	T	T	T	T
1hdg	T	T	T	T	M
2prd	T	T	T	T	T
1ldn	T	T	T	T	T
1bdm	T	T	T	T	T
3mds	T	T	T	T	T
1xgs	T	T	T	T	T
3pfk	T	M	T	T	M
1php	T	T	T	T	T
1ebd	T	T	T	T	T
1ril	T	T	M	M	M
1caa	T	M	M	M	M
1thm	T	M	M	M	M
1lnf	T	M	M	M	T
1btm	T	T	M	M	M
Accuracy		75%	65%	65%	60%
1aky	M	T	M	M	M
3chy	M	T	M	M	T
1csh	M	M	M	M	M
1efu	M	T	T	T	M
1xnb	M	M	M	M	M
1hrd	M	M	T	T	T
1gad	M	M	M	M	M
1ino	M	T	T	T	T
1ldg	M	T	T	M	M
4mdh	M	T	M	M	M
1qmn	M	T	M	M	M
1mat	M	T	T	M	M
2pfk	M	T	T	T	M
1qpg	M	T	M	M	M
1lpf	M	T	M	M	M
2rn2	M	M	M	M	M
8rxn	M	M	M	M	M
1st3	M	M	M	M	M
1npc	M	M	M	M	M
1ypi	M	M	M	M	M
Accuracy		45%	70%	80%	85%

T: thermophilic protein; M: mesophilic protein.

only one node. The ANN models were confined to a single hidden layer because the network with more than one hidden layer is harder to train [10].

3. Results and discussion

3.1. Pattern recognition of thermophilic and mesophilic proteins via PCA

After performing a PCA on the dataset it was found that PC_1 described 27.7% and PC_2 17.7% of the total variance present in the data set. As can be seen in the score plot (Fig. 1A), most of the samples were clustered in two groups in training sample. The average accuracy was 92%. The relationship between the first two PCs and the amino acid can be described as following:

$$\begin{aligned}PC_1 = {} & 0.273A + 0.253C + 0.47D - 0.295E - 0.112F \\ & + 0.038G + 0.257H - 0.342I - 0.425K + 0.23L \\ & + 0.055M - 0.111N + 0.194P + 0.364Q + 0.203R \\ & + 0.047S - 0.008T - 0.152V - 0.082W - 0.288Y\end{aligned}$$

$$\begin{aligned}PC_2 = {} & -0.267A + 0.282C + 0.179D - 0.235E + 0.109F \\ & - 0.209G + 0.158H + 0.172I + 0.056K - 0.047L \\ & + 0.041M + 0.399N - 0.066P + 0.156Q - 0.326R \\ & + 0.338S + 0.349T - 0.338V + 0.001W + 0.011Y\end{aligned}$$

A discriminating inequality was got as the following: when $PC_2 > -1.5PC_1 + 6$, the protein was mesophilic, and when $PC_2 < -1.5PC_1 + 6$, it was thermophilic. It meant that when the amino acid composition of a protein satisfied the following inequality, it would be a thermophilic protein.

$$\begin{aligned}0.678E + 0.059F + 0.152G + 0.341I + 0.582K + 0.021R \\ + 0.566V + 0.421Y + 6 > 0.142A + 0.661C + 0.249D \\ + 0.544H + 0.298L + 0.123M + 0.232N + 0.225P \\ + 0.701Q + 0.408S + 0.337T + 0.124W\end{aligned}$$

From the inequality above, when proteins had higher frequency of Glu, Lys, Val and Tyr, while lower frequency of Gln, Cys, His and Ser, the inequality was tenable, meaning that the proteins were thermophilic. This was accord with the conclusions of some early studies. Chakravarty and Vardarajan [11] have reported that the increase in proportion of charged residues (Arg, Lys, His, Asp, Glu) and decrease in proportion of uncharged polar residues (Ser, Thr, Gln, Asn, Cys) in thermophilic proteins were statistically significant. Thompson and Eisenberg [12] have also observed that thermophilic proteins contained more Glu, Val, Arg, and Gly residues and less Gln, Ser, and Asp residues. This showed that the inequality we established here could explain the mechanism of protein thermostability to some extent. Using above function, we checked its accuracy by calculating proteins in the testing data set (Fig. 1B), the accuracy for thermophilic protein was 75%, and mesophilic protein only 45%. The average accuracy was only 60%.

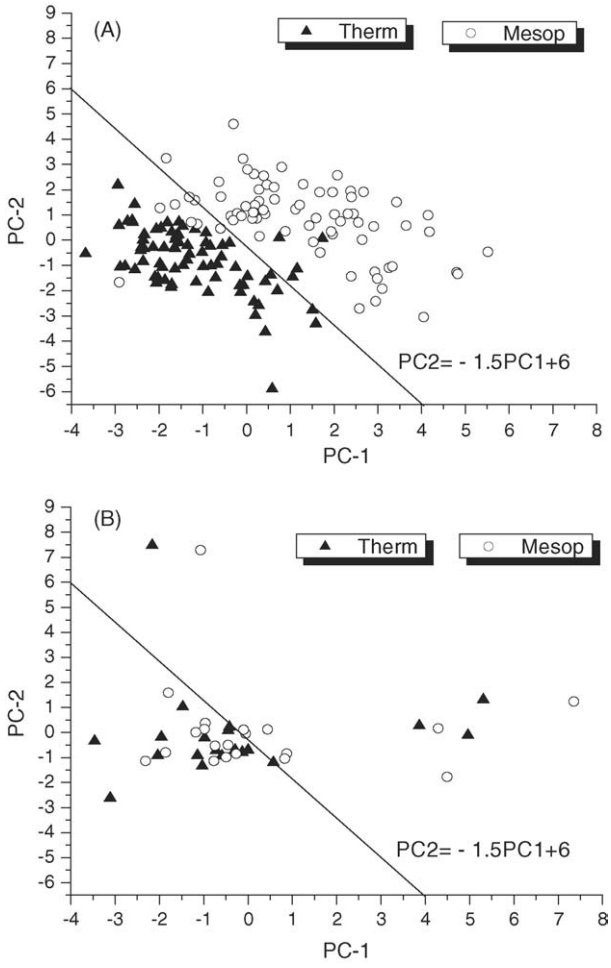

Fig. 1. The map of classification between principal component 1 and principal component 2. (A) The training sample; (B) the testing sample.

3.2. Pattern recognition of thermophilic and mesophilic proteins via SR

Thermophilic proteins were assigned a value of 1, and mesophilic proteins were assigned a value of 0, using the stepwise regression procedure of DPS, a discriminating function was got as following (coefficient of each variant was retained with three decimal place accuracy, and the model reached significant $p < 0.0001$, $R = 0.87$):

$$\begin{aligned}Z = {} & 0.363 - 0.068C - 0.059D + 0.067I - 0.089Q + 0.026R \\ & - 0.076S - 0.066T + 0.064V + 0.099Y\end{aligned}$$

To discriminate thermophilic protein from mesophilic protein, we set the value as 0.48 according to the calculated value in the training data set. If the Z value was above 0.48, it was considered as thermophilic protein, and if the value below 0.48, it was considered as mesophilic protein.

From the equation above, we can see that amino acids such as Ile, Val, Arg and Tyr were positive correlated with Z, Gln,

Asp, Thr, Ser and Cys were negative correlated with *Z*, while the left 11 amino acids were excluded in the equation, suggesting that these amino acids may have no difference between these two kinds of proteins. When proteins have higher frequency of Ile, Val, Arg and Tyr, while lower frequency of Gln, Asp, Thr, Ser and Cys, the *Z* value will be larger, meaning that the proteins are thermophilic. This result agreed with the conclusions of some systematical analyses. In the analyses, any difference has not been reported about the aliphatic amino acids in thermophilic proteins except that thermophilic proteins have higher frequency of Ile compared with mesophilic ones [6]. Several properties of Arg residues suggest that they would be well adapted to high temperatures: the Arg δ-guanido moiety has a reduced chemical reactivity due to its high pKa and its resonance stabilization. The δ-guanido moiety can provide more surface area for charged interactions. The Arg side chain contains one fewer methylene group than Lys, it has the potential to develop less unfavorable contacts with the solvent. Last, because its pKa (approximately 12) is 1 unit above that of Lys (11.1), Arg more easily maintains ion pairs and a net positive charge at elevated temperatures (pKa values drop as the temperature increases) [13]. Thermophilic proteins have higher frequency of Tyr compared with mesophilic ones because Tyr was known as the participant for cation–pi interaction, which is an interaction for maintaining conformational stability in protein structure [14]. Gln was known as thermolabile amino acid due to their tendency to undergo deamination at high temperature, but thermophilic proteins were characteristically reduced in Gln.

Using the equation above, we checked its accuracy by calculating proteins in the testing data set (Table 2), the accuracy for thermophilic protein was 65% and mesophilic protein 70%. The average reliability was 67.5%.

3.3. Pattern recognition of thermophilic and mesophilic proteins via PLSR

Thermophilic proteins and mesophilic proteins were assigned a value of 1 and 0, respectively. PLSR was performed to identify underlying factors in the training data set. Plotting all proteins in reduced dimensions (Fig. 2) produces a clear separation of thermophilic and mesophilic proteins. Then, using the PLSR procedure of DPS, a discriminating function was got as following (coefficient of each variant was retained with three decimal place accuracy):

$$\begin{aligned} Z =& 0.462 - 0.001A - 0.116C - 0.041D + 0.018E - 0.001F \\ &+ 0.011G - 0.033H + 0.035I + 0.013K - 0.002L \\ &- 0.025M - 0.025N + 0.012P - 0.069Q + 0.028R \\ &- 0.055S - 0.07T + 0.053V - 0.012W \\ &+ 0.069Y \ (R = 0.86) \end{aligned}$$

To discriminate thermophilic protein from mesophilic protein, we set the value as 0.48 according to the calculated value in the training data set. If the *Z* value was above 0.48, it was considered as thermophilic protein, and if the value was below 0.48, it was considered as mesophilic protein.

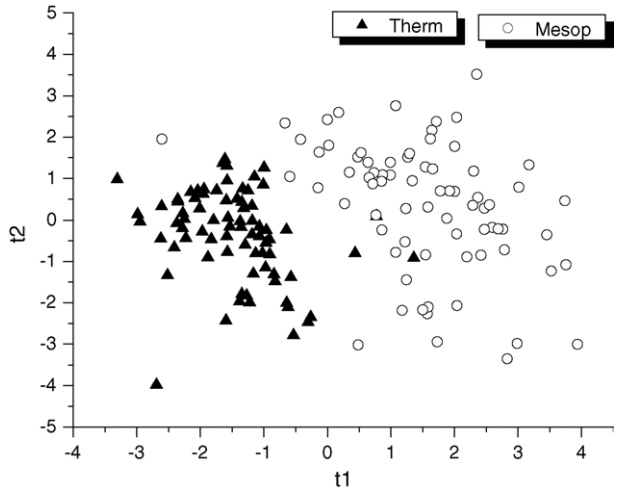

Fig. 2. The two dimensional map of *t*1 and *t*2 of partial least-square regression.

The strongest contributions to *Z* come from Cys, Thr, Gln, Tyr, Ser, and Val, whereas Cys and Thr exhibit high negative correlations to *Z*. It means that there should be an apparent tendency for less Cys and Thr in thermophilic proteins. Cys is known as thermolabile amino acids duo to its tendency to undergo oxidation at high temperature, but systematical analyses reported that thermophilic proteins have lower frequency of Cys compared with mesophilic ones, Thr and Ser are known as the best residue for interacting with the water surrounding protein structure, but the water would be released at higher temperature, the local protein structure around water-binding site such like Thr and Ser could be changed to be unstable enough to evoke protein instability, so the thermophilic proteins have very low frequency of Thr and Ser compared with mesophilic proteins [15]. It has been widely accepted that the aliphatic amino acids (such as Val) would contribute to the hydrophobic interaction, which is the main force for maintaining conformational stability in inner part of protein. This was approved by the fact that Val exhibited high positive correlations to *Z*. Using the equation above, we checked its accuracy by calculating proteins in the testing data set (Table 2), the accuracy for thermophilic protein was 65% and mesophilic protein 80%. The average accuracy was 72.5%.

3.4. Pattern recognition of thermophilic and mesophilic proteins via PC-ANN

To reduce the dimension of the node of the input layer and consequently to increase the speed of calculation and overall performances of the ANN models, the data set were subjected to PCA. As observed, the first 13 PCs explained 88% of variances in the descriptors data matrix and were selected. So the node of the input layer was 13, the node of the output layer was 1. The maximum iterations were 1000, and the learning rate (η), momentum parameter and Sigmoid parameter was 0.1, 0.6 and 0.9, respectively. In general, the optimal number of hidden nodes and the training error is difficult to determine. Empirically, the number of the hidden nodes is about 75% of the number of the input nodes. We assigned the number of the

hidden layer nodes as 8, 9 and 10, the training error was appointed as 0.001 and 0.005, and found that when the number of the hidden layer was 9, the training error was 0.005; it gave out the best result among the six combinations. So the architecture of the ANN was chosen as '13-9-1'.

Thermophilic proteins and mesophilic proteins were assigned a value of 1 and 0, respectively. The training data set was used to train the network, after training, the fitting result was very encouraging; the accuracy was 98.0%. The testing data set was used to examine the validity of the model in predicting the outputs, the predicting results were shown in Table 1, and the accuracy for the thermophilic protein was 60% and mesophilic protein 85%. The average accuracy was 72.5%. Comparison of the results of the ANN method with those of the PLSR model revealed the lower prediction errors in mesophilic proteins but higher in thermophilic protein.

The advantage of PCA was mentioned above. But the percentages of variance explained by PC_1 and PC_2 of our model were low. Maybe these low percentages have resulted in the rather low accuracy of prediction as shown in Fig. 1B. Whereas, in another study [16], the first three components together only encoded 48.9% of the variance (22.9%, 14.9% and 11.1%) of the original 12-variable data set, however, the results showed a clear distinction between the biotic and abiotic categories. The stepwise regression method could find out the responsible amino acid for protein thermostablility, but the prediction accuracy was not satisfying. The most important advantage of PLSR reported to the non-problematic handling of multi-collinearities relying on an iterative algorithm, which made possible the treatment of data with more features than objects. While, in general, ANNs gain an advantage over PCA, SR and PLSR techniques because the transfer function may be nonlinear. A neural network can discover nonlinear interactions between variables in the data set that would be missed by a linear technique. The prediction accuracy of PLSR and PC-ANN was ideal. But PC-ANN functioned largely as a black box and understanding of the acquired knowledge was not always possible. It implied that the meaning of PC-ANN result was difficult to understand. While, the inherent advantage of the PLSR model over the ANN model was that the PLSR computations were simpler and require shorter computation time, and the biological meaning of PLSR was easy to understand and seemed to explain the thermostability of proteins.

Discrimination of thermophilic and mesophilic proteins via pattern recognition algorithms provided a new thought for analyzing the difference between thermophilic and mesophilic proteins. In theory, it can help to understand the mechanism of protein thermostability and give a quantitative model to explain the sequence–characteristic relationship. In practical application, it may be used to aid in protein redesign and reduce the screening burden for rapid optimization of protein properties at high temperature. This may provide an alternative approach to expensive assays or unreliable high-throughput surrogate screens.

4. Conclusion

Using four pattern recognition algorithms, proteins were classified into thermophilic or mesophilic ones, the accuracy of classification was rather moderate. The highest accuracy for the thermophilic proteins was 75%, which was predicted via PCA, while the highest accuracy for the mesophilic proteins was 85%, which was predicted via PC-ANN. However, the highest average accuracy was 72.5%, which was predicted via PLSR and PC-ANN. These models seemed to explain some of the biological meanings may have practical application.

Acknowledgement

The authors gratefully thank Dr. Yizhou Zheng for helpful comments on the manuscript. We acknowledge support from the Science Foundation of Overseas Chinese Affairs Office of the State Council of China (Grant No. 05QZR06).

References

[1] McCarthy T, Hanniffy O, Lalor E, Savage AV, Tuohy MG. Evaluation of three thermostable fungal endo-β-glucanases from *Talaromyces emersonii* for brewing and food applications. Process Biochem 2005;40(5):1741–8.
[2] Sharma R, Soni SK, Vohra RM, Gupta LK, Gupta JK. Purification and characterisation of a thermostable alkaline lipase from a new thermophilic *Bacillus* sp. RSJ-1. Process Biochem 2002;37(10):1075–84.
[3] Fagain CO. Understanding and increasing protein stability. Biochem Biophys Acta 1995;1252:1–14.
[4] Kreil DP, Ouzounis CA. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res 2001;29:1608–15.
[5] Pal NR, Pal S. Computational intelligence for pattern recognition. Int J Pattern Recognit Artif Intell 2002;16:773–9.
[6] Seung PP, Yoo YJ. Protein thermostability: structure-based difference of amino acid between thermophilic and mesophilic proteins. J Biotechnol 2004;1111:269–77.
[7] Tang QY, Feng MG. Practical statistics and DPS data processing system. Peking, China: Science Press, 2002. p. 241–90.
[8] Duttaa D, Mohanty AK, Choudhury RK, Chand P. Pattern recognition of particle tracks using principal component analysis and artificial neural network. Nucl Instrum Methods Phys Res A 1998;404:445–54.
[9] Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Anal Chim Acta 1986;185:1–17.
[10] Bahram H, Mohammad AS, Ramin M, Nasim N. Toward an optimal procedure for PC-ANN model building: prediction of the carcinogenic activity of a large set of drugs. J Chem Inf Model 2005;45:190–9.
[11] Chakravarty S, Varadarajan R. Elucidation of determinants of protein stability through genome sequence analysis. FEBS Lett 2000;470:65–9.
[12] Thompson MJ, Eisenberg D. Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. J Mol Biol 1999; 290:595–604.
[13] Vieille C, Zeikus GJ. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev 2001; 65(1):1–43.
[14] Ma JC, Dougherty DA. The cation–pi interaction. Chem Rev 1997; 97:1303–24.
[15] Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. Protein Eng 2000;13:179–91.
[16] Dorn ED, McDonald GD, Storrie-Lombardi MC, Nealson KH. Principal component analysis and neural networks for detection of amino acid biosignatures. Icarus 2003;166:403–9.