Identifying the mesophilic and thermophilic proteins from their amino acid composition with ν-Support Vector Machines

Ding Y. R.[1,2*], Cai Y. J.[2,3], Sun J.[1], Xu W. B.[1]

[1] School of Information Technology, Jiang Nan University, Wuxi, 214122, China
2 Key Laboratory of Industrial Biotechnology, Jiang Nan University, Wuxi, 214122, China
3 School of Biotechnology, Jiang Nan University, Wuxi, 214122, China

Abstract:

Many researchers had proved that both single amino acid composition and dipeptide composition can influence protein thermostability. We use ν-support vector machines approach to predict hyperthermophilic protein, thermophilic protein and mesophilic protein from single amino acid composition, dipeptide composition and the combination of the two factors. For the prediction accuracies, we conclude that, single amino acid composition is suitable for prediction of mesophilic protein; dipeptide composition is suitable for prediction of hyperthermophilic protein and thermophilic protein; when considering the combination of the two factors, the prediction accuracy of hyperthermophilic protein is 84.1%, thermophilic protein is 83.4%, mesophilic protein is 84.4%, average accuracy is 84.0%. It shows that the protein thermostability can be predicted properly based on the combination of single amino acid composition and dipeptide composition. Obviously, dipeptide composition is correlative significantly to protein thermostability based on the prediction accuracies.

1. Introduction

In 2006, Japanese researchers found a protein called "CutA1", which can act in 148.5℃. As we know, both mesophilic proteins and thermophilic proteins are composed of the same kinds of amino acids. Why thermophilic proteins can maintain their activities at high temperatures? There are many factors that influence the thermostability of proteins[1-13], Such as single amino acid composition[1], disulfide bond[2,3], hydrophobic interactions[4~6], aromatic interactions[7], hydrogen bond[4,5,8,9], ion pairs[4,5,8,10~12], prolines and decreasing the entropy of unfolding[14,15], intersubunit interactions and oligomerization[16], packing and reduction in solvent-accessible hydrophobic surface[5,17,18].

Among these factors, single amino acid composition has long been thought to be correlated significantly to its thermostability [19,20]. Several investigations [19~24] have been carried out to illustrate the influence of amino acid composition on protein thermostability. These studies showed that thermophilic protein prefers to contain charged, aromatic, and hydrophobic residues comparing to mesophilic protein.

From the facts that mutation of the residues in the mesophilic enzyme to those observed in the thermophilic enzyme (i.e. Ser->Ala and Thr->Ala) produces a mutant enzyme which is 20℃ more stable than the wild type [23], and the tertiary structures of pig heart (37℃) and Thermoplasma acidophilum (55℃) citrate synthases have a high degree of structural homology but only 20% sequence identity[18], we can also know that single amino acid composition play an dominant role on protein thermostability.

In our previous work [13], we studied the influence of dipeptide composition on protein thermostability. At the same time, the influence of single amino acid composition also was studied for comparison. We found the influence of single amino acid composition could be deduced from the influence of dipeptide composition. The characteristic dipeptides not only describe the dipeptide that influence protein thermostability significantly but also show the relationship among significant single amino acids that influence protein thermostability.

Support Vector Machines (SVMs) is a good supervised machine learning technology, which can get high prediction accuracy using fewer data than Neural Network or Genetic Algorithm. In this paper, the use of the SVMs approach to predict protein thermostability from single amino acid composition, dipeptide composition, and the combination of the two factors is described. From the prediction accuracy, we not only know if the SVMs can predict protein thermostability from these factors, but also can deduce which factor that examined is correlative significantly to protein thermostability.

2. Material and Method

2.1 Dataset

At present, there are 10 hyperthermophilic organisms, 3 thermophilic organisms and 52 mesophilic organisms in NCBI COG database[25~27]. We selected the prokaryotic organisms from them and retrieved protein sequences of each organism from NCBI database (http://www.ncbi.nlm.nih.gov/COG). To make the sequences' similarity less than 30%, we use ClustalW program (http://www.ebi.ac.uk/clustalw) to remove the redundant sequences. Then, the final dataset was composed of 15187 hyperthermophilic protein sequences, 3974 thermophilic protein sequences and 101868 mesophilic protein sequences.

2.2ν-Support Vector Machines

2.2.1 Basic theory

SVMs are a discriminative supervised machine learning technology based on statistical learning theory[28]. SVMs have many attractive features, including effective avoidance of overfitting, the ability to handle large feature spaces, information condensing of the given data set, etc. Then this method has been shown to be an effective bioinformatics tool in multiple areas of biological analysis including identifying splicing sites of eukaryotic RNA[29], protein fold recognition[30], protein-protein interactions prediction[31] protein secondary structure prediction[32], evaluating microarray expression data[33, detecting remote protein homologies[34], protein subcellar localization prediction[35].

Here, we briefly describe the basic ideas of ν-SVMs for pattern recognition. Let us consider a binary classification task with data points $X_i \in R^d \, (i = 1, ..., m)$ with corresponding labels $y_i \in \{-1, +1\} \, (i = 1, ..., m)$ where +1 and -1 are used to stand respectively for the two classes.

The decision function implemented by SVMs can be written as:

$$f(X) = \text{sgn}(\sum_{i=1}^{m} y_i \alpha_i .(\Phi(X).\Phi(X_i)) + b)$$

$$= \text{sgn}(\sum_{i=1}^{m} y_i \alpha_i .K(X, X_i) + b) \qquad (1)$$

Where the coefficients α_i are obtained by solving the following convex Quadratic Programming (QP) problem:

Maximize: $W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(X_i, X_j)$ (2)

Subject to $0 \le \alpha_i \le \frac{1}{m}$, $\sum_{i=1}^{m} \alpha_i y_i = 0$, and $\sum_{i=1}^{m} \alpha_i \ge \nu$ (3)

In the equations (1) and (2), $K(X_i, X_j)$ is kernel function that determines a non-linear mapping of the input vectors from the Euclidean space R^d into a higher dimensional Hilbert space H. In this paper, we adopt the Radial Basic Function (RBF) kernel:

$$K(X_i, X_j) = \exp(-\gamma \left\| X_i - X_j \right\|^2) \qquad (4)$$

As we know, the ν-support vector machines use a new parameter υ in equation (3) which let one control the number of support vectors and errors. The parameter $\nu \in (0,1]$ is an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors. It is different from traditional C-support vector machines. The details can be found in reference [36].

It is obviously that it is easier and more intuitive to deal with $\nu \in [0,1]$ than $C \in [0, \infty)$. So we use ν-SVMs to prediction the protein thermostability.

2.2.2 ν-SVMs Multi-class Method

Support Vector Machines were originally designed for binary classification. How to effectively extend it for multi-class classification is still an on-going research issue. Currently there are two types of approaches for multi-class SVMs. One is by constructing and combining several binary classifiers while the other is by directly considering all data in one optimization formulation. The first method is including one-against-all, one-against-one and DAGSVM methods. Some researchers [37] show that one-against-one is more suitable for practical use than other methods. This method constructs *k*(*k*-1)/2 classifiers where each one is trained on data from two classes. Here *k* is the number of class.

In this paper, OSU_SVM classifier matlab toolbox (http://www.ece.osu.edu/~maj/osu_svm/) which supports one-against-one multi-class classification is used to predict the protein thermostability. In another word, for user, this toolbox can directly deal with multi-class problem.

2.2.3 ν-SVMs input vector and labels

In this paper, there are three kinds of vectors.

The first kind of vector is defined as: $X_1 = (x_1, x_2, x_3, \ldots, x_{20})^T$ where x_i (i=1,2,3,…,20) is the composition of each amino acid in the protein.

The second kind of vector is defined as $X_2 = (x_1, x_2, x_3, \ldots, x_{400})^T$ where x_i (i=1,2,3,…,400) is the composition of dipeptide (AA, AC, AD, …, AY, CA, CC, CD, …,CY, …, YA, YC, YD, ... , YY) in the protein.

The third kind of vector is defined as $X_3 = (x_1, x_2, x_3, \ldots, x_{400}, x_{401}, \ldots x_{420})^T$ where x_i (i=1,2,3,…,420) is the composition of dipeptide and single amino acid (AA, AC, AD, …, AY, CA, CC, CD,…, CY,…YA, YC, YD, …, YY, A, C, D, …, Y) in the protein.

In ν-SVMs, three labels (1, 2, 3) were used to represent hyperthermophilic proteins, thermophilic proteins and mesophilic proteins separately.

2.3 The Training and Predicting Process

The regularization parameter controls the complexity of the learning machine to a certain extent and influences the training speed. To solve the classification problem properly, it's important to select optimal regularization parameters. In addition, if the data came from different class for training is unbalanced, the prediction system would not good. The 10 fold cross-validation procedure is employed to estimate the classification accuracy for selecting suitable regularization parameter and examining the influence of unbalance data on prediction accuracy. A grid search on regularization parameter using 10 fold cross-validation was carried out on training data. The training data were selected from dataset randomly. The proportion of hyperthermophilic proteins, thermophilic proteins, and mesophilic proteins of training data was examined here. The training data was divided into 10 subsets of (approximately) equal size. Sequentially one subset is tested using the classifier trained on the remaining 10-1 subset. Thus, each instance of the whole training set is predicted once so the cross validation accuracy is the percentage of data which are correctly classified. Basically pairs of (υ, γ) are tried and the one with the best accuracy of 10 fold cross-validation is picked. Figure 1 describes the process of ν-SVMs training and predicting protein thermostability.

Here, the average accuracy (AA) and the prediction accuracy of each class are used to assess the prediction system.

$$accuracy_i = \frac{p_i}{n_i} \tag{5}$$

$$AA = \frac{\sum_{i=1}^{k} \frac{p_i}{n_i}}{k} \tag{6}$$

Where, p_i is the number of correctly predicted proteins in class i, n_i is the number of proteins in class i, k is the class number.

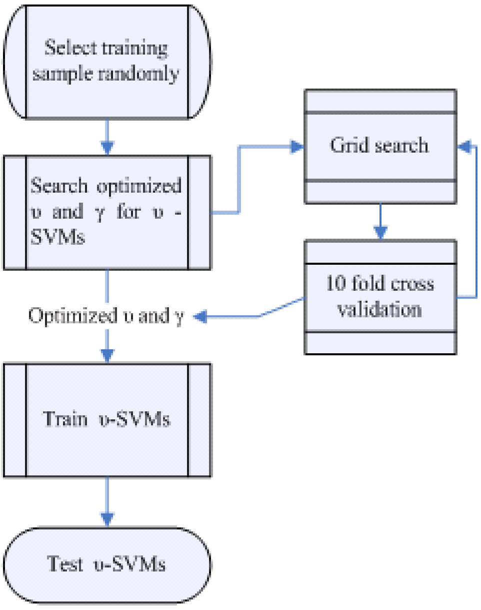

Figure1. Training and Predicting Process of ν-SVMs

3. RESULT AND DISCUSSION

3.1 Regularization parameter selection

As we mentioned, 10 fold cross-validation is used to select the optimal parameters, for a certain training sample size, the optimal parameters is selected through "grid search" method, the highest prediction accuracy of 10 fold cross-validation, the most optimal parameters. The result is listed in table 1.

Table 1. The prediction accuracies of 10 fold cross-validation and optimal parameters of different training samples

No	Training size	Sample	Amino acid composition AA (%)	Dipeptide composition AA (%)	The combination of single amino acid composition and dipeptide composition AA (%)
1	1000:1000:1000		79.8	79.9	80.7
2	2000:2000:2000		80.6	82.4	82.8
3	3000:3000:3000		81.2	83.1	83.6
4	4000:3000:4000		81.9	82.9	83.4
5	5000:3000:5000		81.7	82.1	82.8
6	6000:3000:6000		80.7	81.3	82.0
7	7000:3000:7000		80.1	80.5	81.3
8	8000:3000:8000		79.7	80.1	80.8
9	9000:3000:9000		79.7	79.9	80.1

From table 1, we know all the prediction accuracies of 10 fold cross-validation are larger than 79%. This indicates that when the regularization parameter were selected properly, the hyperthermophilic proteins, thermophilic proteins, and mesophilic proteins can be well separated based on single amino acid composition, dipeptide composition, or the combination of the dipeptide composition and single amino acid composition. All prediction accuracies based on single amino acid composition are smaller than the others, this shows that the influence of single amino acid composition on protein thermostability is smaller than dipeptide composition. From No. 1 to No. 3 the prediction accuracies of 10 fold cross-validation ascend, while from No. 4 to No. 9, the prediction accuracies of 10 fold cross-validation decrease, when the training sample size is 3000:3000:3000, most of the prediction accuracies are highest. It is a good proportion for training sample to get a good prediction system.

3.2 Prediction result based on single amino acid composition

From table 2, we can easily find the prediction accuracies for mesophilic protein are higher than the other proteins, this shows the single amino acid composition of mesophilic protein is very different from hyperthermophilic proteins and thermophilic proteins.

Table 2. Prediction result based on single amino acid composition

No.	Training Sample Size	Hyperthermophilic protein Accuracy (%)	Thermophilic protein Accuracy (%)	Mesophilic protein Accuracy (%)	AA(%)	Optimized (ν，γ) pair
1	1000:1000:1000	77.1	72.3	82.6	77.3	(0.5,185)
2	2000:2000:2000	78.2	73.8	85.7	79.2	(0.5,155)
3	**3000:3000:3000**	**80.0**	**75.7**	**85.8**	**80.5**	(0.5,100)
4	4000:3000:4000	80.4	63.3	88.6	77.4	(0.5,135)
5	5000:3000:5000	81.7	57.4	89.4	76.2	(0.5,155)
6	6000:3000:6000	85.4	59.2	89.1	77.9	(0.5,130)
7	7000:3000:7000	86.7	54.2	89.4	76.8	(0.5,130)
8	8000:3000:8000	85.7	49.9	90.1	75.2	(0.5,140)
9	9000:3000:9000	87.5	46.9	89.9	74.8	(0.5,140)

Train sample size=hyperthermophilic:thermophilic:mesophilic

Because the number of thermophilic protein sequences is very small comparing with hyperthermphilic protein sequences and thermophilic protein sequences, we have to increase the amounts of hyperthermophilic and mesophilic protein sequences to consider the influence of training sample size on prediction accuracy. Although, average accuracy has only a little change, the prediction accuracies for thermophilic protein decreased dramatically. The sequence amount is more unbalanced, the accuracies for the thermophilic protein are lower. From No. 3 to No. 9, the average accuracy decreased only 5.7%, but the accuracies for thermophilic protein decreased 28.8%. As we know, microorganisms can be classified according to their optimal growth temperature [38], T_{opt}, roughly into four groups: psychrophilic (0< T_{opt} <20℃), mesophilic (20< T_{opt} <50℃), thermophilic (50< T_{opt} <80℃) and hyperthermophilic (80< T_{opt} <120℃). Obviously, thermophilic protein is a transition protein between mesophilic protein and hyperthermophilic protein, and if the training data is unbalanced, ν-SVMs will receive more information from hyperthermophilic and mesophilic protein and less 'noise' from thermophilic, then ν-SVMs can predict hyperthermophilic and thermophilic protein with the accuracy around 90%, but the average accuracy is relative lower. Also, we had checked the selection process of ν-SVMs parameters carefully. The prediction results under unbalanced dataset were not suffered from overtraining. Considering better prediction accuracy of each class and average accuracy, when the training size is 3000:3000:3000, the prediction result is best.

3.3 Prediction result based on dipeptide composition

In our previous work [13], we had found that the dipeptide composition could provide more information of protein thermostability than single amino acid composition. In order to compare with results in table 2, we use the same protein sequences as in table 2 to train the ν-SVMs.

Table 3. Prediction result based on dipeptide composition

No.	Training Sample Size	Hyperthermophilic protein Accuracy (%)	Thermophilic protein Accuracy (%)	Mesophilic protein Accuracy (%)	AA(%)	Optimized(ν，γ）pair
1	1000:1000:1000	78.8	83.9	77.5	80.1	(0.5,300)
2	2000:2000:2000	81.7	84.6	79.1	81.8	(0.5,340)
3	**3000:3000:3000**	**83.0**	**85.1**	**83.0**	**83.7**	(0.5,280)
4	4000:3000:4000	86.9	77.1	86.4	83.5	(0.5,260)
5	5000:3000:5000	89.3	71.6	87.9	82.9	(0.5,280)
6	6000:3000:6000	90.7	68.9	88.1	82.6	(0.5,280)
7	7000:3000:7000	91.7	64.4	88.7	81.6	(0.5,280)
8	8000:3000:8000	91.9	58.6	89.0	79.8	(0.5,320)
9	9000:3000:9000	92.3	57.7	89.1	79.7	(0.5,320)

Train sample size=hyperthermophilic:thermophilic:mesophilic

Table 3 shows there are higher prediction accuracies for hyperthermophilic protein and thermophilic protein than mesophilic protein. Comparing with table 2, the unbalanced data have the same influence on prediction accuracies as that in table 2. When the sample size is balance, the predict accuracies for mesophilic protein in table 3 is lower than that in table 2, but when the sample size is unbalance, the predict accuracies for mesophilic protein in table 3 is similar as that in table 2. Because the dipeptide composition is 400 dimensions, it includes more information than single amino acid composition. Then, average accuracy based on dipeptide composition have an improvement than those based on single amino acid composition. Obviously, the training sample size, 3000:3000:3000 is the best.

3.4 Prediction result based on the combination of dipeptide composition and single amino acid composition

From the above results, we can find single amino acid composition is better to predict mesophilic proteins and dipeptide composition is better to predict hyperthermophilic and thermophilic proteins. Here, we combined these two factors to predict protein thermostability. The protein sequences for training and predicting in table 4 are same as that in table 2 and in table 3. The results were list in table 4.

Table 4: Prediction result based on the combination of dipeptide composition and single amino acid composition

No.	Training Sample Size	Hyperthermophilic protein Accuracy (%)	Thermophilic protein Accuracy (%)	Mesophilic protein Accuracy (%)	AA(%)	Optimized(ν，γ）pair
1	1000:1000:1000	80.4	81.2	79.8	80.5	(0.5,500)
2	2000:2000:2000	82.9	81.7	83.3	82.6	(0.5,560)

3	**3000:3000:3000**	**84.1**	**83.4**	**84.4**	**84.0**	(0.5,560)
4	4000:3000:4000	87.2	79.1	87.2	**84.5**	(0.5,740)
5	5000:3000:5000	88.9	71.0	88.8	82.9	(0.5,660)
6	6000:3000:6000	90.1	70.2	89.0	83.1	(0.5,580)
7	7000:3000:7000	91.3	66.6	89.4	82.4	(0.5,620)
8	8000:3000:8000	90.9	62.6	90.3	81.3	(0.5,680)
9	9000:3000:9000	92.1	62.6	90.3	81.7	(0.5,680)

Train sample size=hyperthermophilic:thermophilic:mesophilic

We find the predict accuracies for three kinds of protein is balance and the average accuracy is higher than that in table 2 and table 3 with balanceable data. For the unbalanced data, although the prediction for hyperthermophilic and mesophilic proteins in table 4 is similar to table 3, the prediction for thermophilic proteins in table 4 is higher than those in table 2 and table 3. Obviously, the overall prediction result based on the combination of dipeptide composition and single amino acid composition is highest. For the training sample size, 3000:3000:3000, the prediction accuracy of hyperthermophilic protein is 84.1%, thermophilic protein is 83.4%, mesophilic protein is 84.4%, and average accuracy is 84.0%. It's a better result for predicting protein thermostability using ν-SVMs. After all, there are many factors influence protein thermostability.

We consider the training sample size is 3000 is enough for predicting each kind of proteins, larger sample size will improve the prediction CPU time significantly.

4. CONCLUSION

Many researchers had analysis the mesophilic and thermophilic proteins based on the same families [39]. They found different thermophilic proteins family carried different information about the protein stability. In this article, we predicted the thermostability of protein by collected together all the sequence. The high prediction accuracy proved that there was the overall trend in mesophilic and (hyper)thermophilic proteins which is implicated in the protein primary structure.

Thermophilic microorganisms are the source of novel thermostability enzymes. Some thermophilic enzymes such as DNA polymerases, amylases from thermophilic microorganisms had been used successfully. For enzymes which can't be found in thermophilic microorganisms, modern techniques like mutation genesis and gene shuffling will lead to convert mesophilic enzyme to thermophilic enzyme. Here, we provide a powerful method (ν-SVMs) which is easily to predict thermostability of protein from primary structure.

5. ACKNOWLEGEMENTS

This work is supported by the innovation team-building project of JiangNan University, JNIRT0702.

References:

[1] Suhre, K. and Claverie, J. M., Genomic correlates of hyperthermostability, an update, Journal of Biological Chemistry, 2003, 278(19), 17198-17202.

[2] Cacciapuoti, G., Porcelli, M., Bertoldo, C., Rosa, M.D. and Zappia, V., Purification and characterization of extremely thermophilic and thermostable 5'-methylthioadenosine phosphorylase from the archaeon Sulfolobus solfataricus. Purine nucleoside phosphorylase activity and evidence for intersubunit disulfide bonds, Journal of Biological Chemistry, 1994, 269(40), 24762-24769.

[3] Matsumura, M., Signor, G. and Matthews, B.W., Substantial increase of protein stability by multiple

disulphide bonds, Nature, 1989, 342, 291-293.

[4] Robinson-Rechavi, M., Alibes, A. and Godzik, A., Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of Thermotoga maritime, Journal of Molecular Biology, 2006, 356(2), 547-557.

[5] Sadeghi, M., Naderi-Manesh, H., Zarrabi, M. and Ranjbar, B., Effective factors in thermostability of thermophilic proteins, Biophysical Chemistry, 2006, 119(3), 256-270.

[6] Saraboji, K., Gromiha, M.M. and Ponnuswamy, M.N., Importance of main-chain hydrophobic free energy to the stability of thermophilic proteins, International Journal of Biological Macromolecules, 2005, 35(3-4), 211-220.

[7] Serrano, L. and Fersht, A.R., Aromatic-aromatic interactions and protein stability. Investigation by double-mutant cycles, Journal of Molecular Biology, 1991, 218, 465-475.

[8] Vogt, G., Woell, S. and Argos, P., Protein thermal stability, hydrogen bonds, and ion pairs, Journal of Molecular Biology, 1997, 269(4), 631-643.

[9] Querol, E., Perez-Pons, J.A. and Mozo-Villarias, A., Analysis of protein conformational characteristics related to thermostability, Protein Engineering, 1996, 9(3), 265-271.

[10] Bae, E. and Phillips, GN Jr., Structures and analysis of highly homologous psychrophilic, mesophilic, and thermophilic adenylate kinases, Journal of Biological Chemistry, 2004, 279(27), 28202-28208.

[11] Fish, A., Danieli, T., Ohad, I., Nechushtai, R. and Livnah, O., Structural basis for the thermostability of ferredoxin from the cyanobacterium Mastigocladus laminosus, Journal of Molecular Biology, 2005, 350(3), 599-608.

[12] Makhatadze, G.I., Loladze, V.V., Ermolenko, D.N., Chen, X. and Thomas. S, T., Contribution of surface salt bridges to protein stability: guidelines for protein engineering, Journal of Molecular Biology, 2003, 327(5), 1135-1148.

[13] Ding, Y.R., Cai, Y.J., Zhang, G.X. and Xu, W.B., The influence of dipeptide composition on protein thermostability, FEBS letters, 2004, 569, 284-288.

[14] Li, C., Heatwole. J., Soelaiman, S. and Shoham, M., Crystal structure of a thermophilic alcohol dehydrogenase substrate complex suggests determinants of substrate specificity and thermostability, Proteins: Structure, Function, and Genetics, 1999, 37, 619-627.

[15] Matthews, B.W., Nicholson, H. and Becktel, W.J., Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding, Proceedings of the National Academy of Sciences, 1987, 84, 6663-6667.

[16] Moriyama, H., Onodera, Ko., Sakurai, M., Tanaka, N., Kirino, H., Oshima, T. and Katsubet, Y., The crystal structures of mutated 3-isopropylmalate dehydrogenase from Thermus thermophilus HB8 and their relationship to the thermostability of the enzyme, The Journal of Biochemistry, 1995, 117, 408-413.

[17] Chen, J. and Stites, W.E., Replacement of staphylococcal nuclease hydrophobic core residues with those from thermophilic homologues indicates packing is improved in some thermostable proteins, Journal of Molecular Biology, 2004, 344(1), 271-280.

[18] Berezovsky, I.N. and Shakhnovich, E.I., Physics and evolution of thermophilic adaptation, Proceedings of the National Academy of Sciences, 2005, 102(36), 12742-12747.

[19] Kumar, S., Tsai, C.J. and Nussinov, R., Factors enhancing protein thermostability, Protein Engineering, 2000, 13, 179-191.

[20] Vieilie, C. and Zeikus, G.J., Hyperthermophilic enzymes: Sources, uses, and molecular mechanisms for thermostability, Microbiology and Molecular Biology Reviews, 2001, 65, 1-43.

[21] Haney, P.J., Badger, J.H., Buldak, G.L., Reich, C.I. and Woese, C.R., Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species, Proceedings of the National Academy of Sciences, 1996, 96, 3578-3583.

[22] Szilagyi, A. and Zabodszky, P., Structural differences between mesophilic, moderately thermophilic and extremely thremophilic protein subunits: results of a comprehensive survey, Structure, 2000, 8, 493-504.

[23] Gromiha, M.M., Oobatake, M. and Sarai, A., Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins, Biophysical Chemistry, 1999, 82, 51-67.

[24] Russell, R.J. and Taylor, G.L., Engineering thermostability: lessons from thremophilic proteins, Current Opinion in Biotechnology, 1995, 6, 370-374.

[25] Tatusov, R.L., Koonin, E.V. and Lipman, D.J., A genomic perspective on protein families, Science, 1997, 278, 631-637.

[26] Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V., The COG database: a tool for

genome-scale analysis of protein functions and evolution, Nucleic. Acids Research, 2000, 28, 33-36.

[27] Tatusov, R.L., The COG database: new developments in phylogenetic classification of proteins from complete genomes, Nucleic Acids Research, 2001, 29, 22-28.

[28] Vapnik, V., Statistical Learning Theory, Wiley-interscience, New York, 1998.

[29] Sun, Y.F., Fan X.D. and Li, Y.D., Identifying splicing sites in eukaryotic RNA: support vector machine approach, Computers in biology and medicine, 2003, 33, 17-29.

[30] Cai, Y.D., Liu, X.J., Xu, X.B. and Chou, K.C., Support vector machines for the classification and prediction of β-turn types, Journal of Peptide Science, 2002, 8, 297-301.

[31] Bock, J.R. and Gough, D.A., Predicting protein-protein interactions from primary structure, Bioinformatics, 2001, 17, 455-460.

[32] Hua, S.J. and Sun Z.R., A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach, Journal of Molecular Biology, 2001, 308, 397-407.

[33] Brown, M. P. S., Grundy, W.N., Lin, D., Cristianini N., Sugnet, C. W., Furey, T.S., Jr, M.A. and Haussler, D., Knowledge-based analysis of microarray gene expression data by using support vector machines, Proceedings of the National Academy of Sciences, 2000, 97, 262-267.

[34] Jaakkola, T., Diekhans, M. and Haussler, D., Using the Fisher kernel method to detect remote protein homologies. In proceedings of the 7th international Conference on intelligent systems for molecular biology AAAi Press, Menlo Park, CA 1999.

[35] Hua, S.J. and Sun, Z.R., Support vector machine approach for protein subcellular localization prediction, Bioinformatics, 2001, 17, 721-728.

[36] Scholkopf, B., Smola, A.J., Williamson, R.C. and Bartlett, P.L. New support vector algorithms, Neural Computation, 2000, 12, 1207-1245.

[37] Hsu, C.W. and Lin, C.J., A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural networks, 2002, 13, 415-425.

[38] Andrey, K. and Rudolf, L., Ion pairs and the thermotolerance of proteins from hyperthermophiles: a 'traffic rule' for hot roads, Trends in biochemical sciences, 2001, 26, 550-556.

[39] Yano, J.K. and Poulos, T.L., New understandings of thermostable and peizostable enzymes, Current Opinion in Biotechnology, 2003, 14, 360-365.