Journal of Microbiological Methods 84 (2011) 67–70

Contents lists available at ScienceDirect

Journal of Microbiological Methods

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

Prediction of thermophilic proteins using feature selection technique

Hao Lin [a,*], Wei Chen [b]

[a] Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China
[b] Department of Physics, School of Basic Medical Sciences, Hebei United University, Tangshan, China

ARTICLE INFO

Article history:
Received 17 August 2010
Received in revised form 15 October 2010
Accepted 19 October 2010
Available online 31 October 2010

Keywords:
Protein thermostability
Support vector machine
Amino acid
Feature selection

ABSTRACT

The thermostability of proteins is particularly relevant for enzyme engineering. Developing a computational method to identify mesophilic proteins would be helpful for protein engineering and design. In this work, we developed support vector machine based method to predict thermophilic proteins using the information of amino acid distribution and selected amino acid pairs. A reliable benchmark dataset including 915 thermophilic proteins and 793 non-thermophilic proteins was constructed for training and testing the proposed models. Results showed that 93.8% thermophilic proteins and 92.7% non-thermophilic proteins could be correctly predicted by using jackknife cross-validation. High predictive successful rate exhibits that this model can be applied for designing stable proteins.

1. Introduction

Protein thermostability is an important aspect of protein biochemical and biotechnological research (Bommarius et al., 2006). Commonly, for chemical reactions, high temperature could increase reaction activity and decrease reaction time. However, many proteins are unstable under the condition of high temperature. In this sense, it is urgent and important to develop a validated method that can predict the stability of a given protein from its primary sequence. Currently, several researches have focused on the study of thermophilic properties of protein (Ding et al., 2004; Liang et al., 2005; Zhang and Fang, 2006a,b,c; Chen et al., 2007; Zhou et al., 2008; Gromiha and Suresh, 2008; Wu et al., 2009; Taylor and Vaisman, 2010). Results have shown that amino acid composition, dipeptide composition, number of ion pairs and salt bridges of proteins are correlated to their thermostability (Fukuchi and Nishikawa, 2001; Dominy et al., 2004; Ibrahim and Pattabhi, 2004; Sadeghi et al., 2006). It has also been confirmed that some single point mutations could influence the thermostability of proteins (Capriotti et al., 2005; Montanucci et al., 2008; Huang and Gromiha, 2009). For example, Ile, Arg, Glu, Lys and Pro residue content were found to be higher, while Ser, Asn, Gln, Thr and Met were lower in thermophilic proteins when compared with mesophilic ones (Zhang and Fang, 2006a; Gromiha and Suresh, 2008). In addition, Gromiha et al. (1999) showed that the Gibbs free energy of hydration and shape were also coupled with the stability of thermophilic proteins. Zhang and Fang (2006a,b) found that the occurrences of some dipeptides were significantly different between thermophilic proteins and mesophilic proteins.

Based on properties of protein sequences, thermophilic proteins can be predicted. Liang et al. (2005) used amino acid coupling patterns to distinguish between thermophilic proteins and their mesophilic orthologs. Zhang and Fang (2006a,b, 2007) employed dipeptide composition and amino acid composition for discriminating between mesophilic proteins and thermophilic proteins. The five-fold cross-validated accuracy achieved 86.6%. Subsequently, Gromiha and Suresh (2008) removed the redundancy of Zhang and Fang's dataset. The overall accuracy of five-fold cross-validation increased to 89% using amino acid composition based on neural network. Montanucci et al. (2008) used support vector machine (SVM) to predict protein thermostability. The accuracy is 88% by using jackknife cross-validation. Recently, Wu et al. (2009) proposed a decision tree to predict protein thermostability. The accuracies of >80% were achieved. Although these works obtain good results, the accuracy is also required to improve.

In this work, we constructed a reliable benchmark dataset containing 915 thermophilic proteins and 793 non-thermophilic proteins. The SVM combined with amino acid composition and dipeptide composition was used to discriminate between thermophilic and non-thermophilic proteins. The feature selection technique was used to improve predictive accuracy. As a result, the jackknife cross-validated accuracy of 93.3% is achieved by use of 30 optimal parameters. Furthermore, the influence of parameters on predictive performance was discussed.

2. Materials and methods

2.1. Datasets

The foundation in developing an accurate model is to construct a reliable benchmark dataset. In present research, thermophilic

[*] Corresponding author. Tel.: +86 28 83208232; fax: +86 28 83208238.
E-mail address: hlin@uestc.edu.cn (H. Lin).

0167-7012/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2010.10.013

proteins and non-thermophilic proteins are extracted from thermophilic organisms and non-thermophilic organisms, respectively. In order to guarantee the non-thermophilic proteins to be denaturation when temperature rises to the temperature of thermophilic organisms, we used 60 °C and 30 °C as the lower limit of optimal growth temperature for thermophilic organisms and the upper limit of optimal growth temperature for non-thermophilic organisms, respectively. By examining the optimal growth temperatures of 1126 complete microbial genomes in National Center for Biotechnology Information (NCBI), 136 prokaryotic genomes (17 archaea and 119 bacteria) meet the requirement (shown in Table S1).

The protein sequences of 136 prokaryotic organisms were extracted from the Universal Protein Resource (UniProt). The following steps were used to guarantee the reliable of protein data.

(1) The proteins must be manually annotated and reviewed.
(2) The protein sequences containing ambiguous residues (such as “X”, “B” and “Z”) were excluded.
(3) The sequences which are fragment of other proteins were excluded.
(4) The proteins which infer from prediction or homology were excluded because of lacking confidence.

After strictly following the above procedures, 1329 thermophilic proteins and 1250 non-thermophilic proteins were obtained. A list of organisms, optimal growth temperatures, number of proteins and domain of life were recorded in Table S1. To get rid of redundancy and homology bias, the CD-HIT program (Huang et al., 2010) was utilized to remove the highly similar sequences using 40% sequence identity as the cutoff. The final data set comprises 915 thermophilic proteins and 793 non-thermophilic proteins which can be downloaded from our web site.

2.2. Parameter selection

One of the most important parts in prediction is to generate a set of informative parameters. The amino acid composition (AAC) and dipeptide composition are accepted parameters which have been widely applied in the area of protein prediction (Zhang and Fang, 2006a,b; Montanucci et al., 2008; Gromiha and Suresh, 2008; Wu et al., 2009). Here, we extended dipeptide composition to *g*-gap dipeptide composition (Lin, 2008). Therefore, our parameters reflect not only the difference on composition and sequence order between two types of proteins, but also on residue correlation. The AAC and *g*-gap dipeptide composition for each sequence can be defined as:

$$f_{20}(i) = \frac{x_{20}(i)}{\sum_i x_{20}(i)} \tag{1}$$

$$f_{400}^{g}(j) = \frac{y_{400}^{g}(j)}{\sum_j y_{400}^{g}(j)} \tag{2}$$

here $x_{20}(i)$ and $y_{400}^{g}(j)$ denote the number of residues of types i and the number of *g*-gap dipeptide of types j in a protein sequence, respectively.

Many works have found that some of the parameters may be redundant to each other which can reduce the predictive accuracies. Thus it is necessary to use feature selection techniques to remove redundant parameters. At present, principal component analysis, genetic algorithm and minimal-redundancy-maximal-relevance were proposed for feature selection (Yuan et al., 2009; Wang and Yang, 2009). Other selection procedures were done according to forward or backward selection (Park et al., 2005; Lin et al., 2009a; Yuan et al., 2009). However, forward or backward selection is time-consuming. Hence we proposed analysis of variance (ANOVA) technique to perform feature selection.

2.3. Support vector machine

The freely available package LibSVM (Chang and Lin, 2001) was used to implement SVM. Four kinds of kernel functions (linear, polynomial, sigmoid and radial basis function) can be chosen to obtain the best classification hyperplane. Here, we used radial basis function to perform the classification. The regularization parameter C and kernel parameter γ must be determined in advance. The Libsvm package offers grid search program to optimize parameters C and γ. For economizing time of calculation, we performed grid search on parameters C and γ using 5-fold cross-validation.

2.4. Performance evaluate

The jackknife cross-validation (Chou and Zhang, 1995; Chou and Shen, 2007; Lin et al., 2009a,b) was used to examine the power of the proposed method. The performance can be measured in term of sensitivity (S_n), specificity (S_p) and accuracy. These parameters can be defined by following equations:

$$Sn = TP\,/\,(TP + FN) \tag{3}$$

$$Sp = TN\,/\,(TN + FP) \tag{4}$$

$$Accuracy = (TP + TN)\,/\,(TP + TN + FP + FN) \tag{5}$$

here *TP*, *TN*, *FP* and *FN* represent the number of the correctly recognized thermophilic proteins, the number of the correctly recognized non-thermophilic proteins, the number of non-thermophilic proteins recognized as thermophilic proteins and the number of thermophilic proteins recognized as non-thermophilic proteins, respectively.

3. Results and discussion

3.1. Discrimination of thermophilic and non-thermophilic proteins

The jackknife cross-validated accuracies of residue composition and *g*-gap dipeptide composition were recorded in Table S2. We found that both residue and *g*-gap dipeptide compositions discriminated thermophilic from non-thermophilic proteins with the accuracy in the range of 89–93%. The accuracy by using residue composition is higher than that of *g*-gap dipeptide composition in spite of the fact that dipeptides contain more parameters. This indicates that some redundant information exists in dipeptide composition. For the analysis of contribution of 20 amino acids, accuracy of each amino acid was calculated and shown in Fig. 1. From this figure, we could deduce that the residue compositions of Glu, Lys, Gln, Ala, ILe are dramatically different between thermophilic and non-thermophilic proteins. In fact, statistics shows that thermophiles have high content of Glu, Lys. These charged residues are easily to form hydrogen bonds or salt bridges which contribute to protein thermostability. It also implies that the thermal denaturation occurs easily for Ala, Gln rich proteins with less charged residues. According to the above analysis, we used these peculiar resides Glu, Lys, Gln, Ala, Ile as parameters to train and test SVM. Jackknife cross-validated results show that the accuracy of peculiar resides is 88.65% which is lower than 92.56% of full amino acid composition. Thus, we used 20 amino acid compositions as initial parameter subset in the following prediction.

Former researches (Zhang and Fang, 2006a,b; Montanucci et al., 2008) have found that residue pairs contain important information for the prediction of thermophilic proteins. However, information

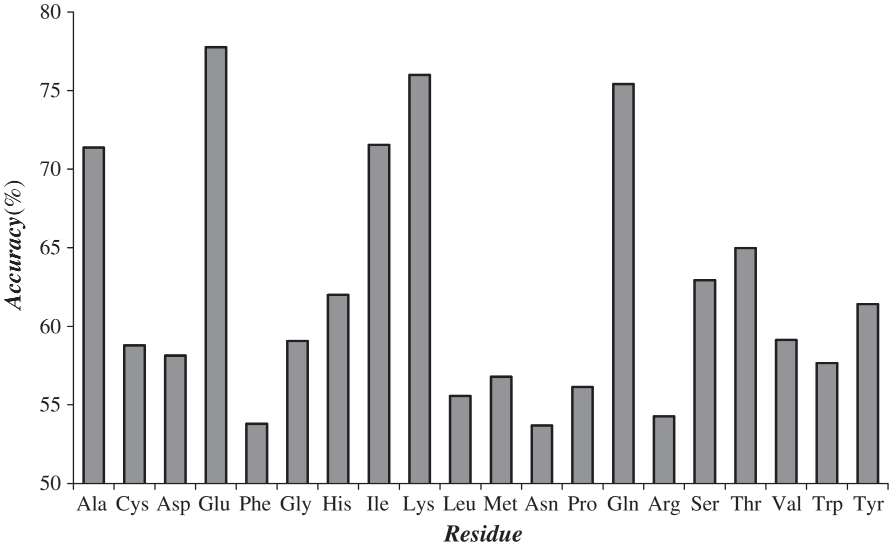

Fig. 1. The accuracies of 20 amino acids.

redundancy in dipeptide set may reduce the cluster-tolerant capacity so as to lower down the predictive accuracy.

ANOVA technique can be used to select informative *g*-gap dipeptides. It initially evaluates each *g*-gap dipeptide and selects the one with the maximum difference between two types of proteins. This selected *g*-gap dipeptide then combines with 20 amino acids as new parameter subset to predict thermophilic proteins. Subsequently, the *g*-gap dipeptide with the second maximum difference is selected out and merged into the parameter subset. We repeated this process until increasing the size of the current parameter subset leads to a lower prediction rate. By examining a great number of parameters, we found that an optimized feature set including 30 parameters achieves the highest predictive accuracy. Table 1 shows that the overall accuracy improves from 92.56% to 93.27%. The informative *g*-gap dipeptides are EE, KE, EI, I*K, I*E, E**K, E**E, K**E, Q**A, E***K (here * denotes gap of residues). Residues E, K, I play important roles in enhancing thermostability of proteins. These results are consistent with former results (Ding et al., 2004; Dominy et al., 2004; Zhang and Fang, 2006a,b; Gromiha and Suresh, 2008).

3.2. Comparison with other methods

It is important to compare our method with other machine learning method using the same benchmark dataset. WEKA (Waikato environment for knowledge analysis) (Witten and Frank, 2005) program including several machine learning techniques such as Bayes Net, Naïve Bayes, Random Forest was used to execute comparisons. Results are recorded in Table 1. Obviously, SVM can achieve the highest predictive successful rate.

Table 1
Results for the prediction of thermophilic proteins and non-thermophilic proteins.

Methods	*Sn* (%)	*Sp* (%)	Accuracy (%)
SVM	93.77	92.69	93.27
Bayes Net	84.92	88.78	86.71
Naïve Bayes	81.75	90.04	85.60
Random Forest	91.37	88.65	90.11
Decision tree J4.8	83.65	81.83	82.33
Bagging meta learning	88.85	87.77	88.35
Logistic function	90.93	90.92	90.93
RBF network	87.98	89.41	88.64
Classification via Regression	88.31	85.88	87.18
NBTree	86.01	82.35	84.31

Recently, Gromiha and Suresh (2008) have predicted a non-redundant dataset including 1609 thermophilic proteins and 3075 mesophilic proteins. The sensitivity, specificity and accuracy of 5-fold cross-validation are 82.4%, 93.0% and 89.4%, respectively. We used this dataset to train and test our model. The sensitivity, specificity and accuracy of 5-fold cross-validation achieved 85.4%, 93.6% and 90.8%, respectively. The accuracy is comparable or superior to that reported by Gromiha and his colleague (Gromiha and Suresh, 2008).

4. Conclusion

We have systematically analyzed the predictive performance of each amino acid for thermophilic and non-thermophilic proteins. By use of ANOVA, ten specific dipeptides (EE, KE, EI, I*K, I*E, E**K, E**E, K**E, Q**A, E***K) were mined to improve performance of SVM. The jackknife cross-validated accuracy is 93.27% which demonstrates that this method can be used to discriminate between thermophilic and non-thermophilic proteins. The predictor based on the proposed model can be freely downloaded from http://cobi.uestc.edu.cn/people/hlin/tools/ThermoPred/.

Acknowledgements

The authors are grateful to the Editors and anonymous reviewers for their valuable suggestions and comments, which have led to the improvement of this paper. This work was supported by the Fundamental Research Funds for the Central Universities (ZYGX2009J081).

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.mimet.2010.10.013.

References

Bommarius, A.S., Broering, J.M., Chaparro-Riggers, J.F., Polizzi, K.M., 2006. High-throughput screening for enhanced protein stability. Curr. Opin. Biotechnol. 17, 606–610.

Capriotti, E., Fariselli, P., Casadio, R., 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 33, W306–W310.

Chang, C.C., Lin, C.J., 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/_cjlin/libsvm.

Chen, C., Li, L., Xiao, Y., 2007. All-atom contact potential approach to protein thermostability analysis. Biopolymers 85, 28–37.

Chou, K.C., Shen, H.B., 2007. Review: recent progresses in protein subcellular location prediction. Anal. Biochem. 370, 1–16.

Chou, K.C., Zhang, C.T., 1995. Review: prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

Ding, Y., Cai, Y., Zhang, G., Xu, W., 2004. The influence of dipeptide composition on protein thermostability. FEBS Lett. 569, 284–288.

Dominy, B.N., Minoux, H., Brooks III, C.L., 2004. An electrostatic basis for the stability of thermophilic proteins. Proteins 57, 128–141.

Fukuchi, S., Nishikawa, K., 2001. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J. Mol. Biol. 309, 835–843.

Gromiha, M.M., Suresh, M.X., 2008. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins 70, 1274–1279.

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

Huang, L.T., Gromiha, M.M., 2009. Reliable prediction of protein thermostability change upon double mutation from amino acid sequence. Bioinformatics 25, 2181–2187.

Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26, 680–682.

Ibrahim, B.S., Pattabhi, V., 2004. Role of weak interactions in thermal stability of proteins. Biochem. Biophys. Res. Commun. 325, 1082–1089.

Liang, H.K., Huang, C.M., Ko, M.T., Hwang, J.K., 2005. Amino acid coupling patterns in thermophilic proteins. Proteins 59, 58–63.

Lin, H., 2008. The modified Mahalanobis Discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J. Theor. Biol. 252, 350–356.

Lin, H., Ding, H., Guo, F.B., Huang, J., 2009a. Prediction of subcellular location of mycobacterial protein using feature selection techniques. Mol. Divers. doi:10.1007/s11030-009-9205-1.

Lin, H., Wang, H., Ding, H., Chen, Y.L., Li, Q.Z., 2009b. Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. Acta Biotheor. 57, 321–330.

Montanucci, L., Fariselli, P., Martelli, P.L., Casadio, R., 2008. Predicting protein thermostability changes from sequence upon multiple mutations. Bioinformatics 24, i190–i195.

Park, K.J., Gromiha, M.M., Horton, P., Suwa, M., 2005. Discrimination of outer membrane proteins using support vector machines. Bioinformatics 21, 4223–4229.

Sadeghi, M., Naderi-Manesh, H., Zarrabi, M., Ranjbar, B., 2006. Effective factors in thermostability of thermophilic proteins. Biophys. Chem. 119, 256–270.

Taylor, T.J., Vaisman, I.I., 2010. Discrimination of thermophilic and mesophilic proteins. BMC Struct. Biol. 10 (Suppl 1), S5.

Wang, T., Yang, J., 2009. Using the nonlinear dimensionality reduction method for the prediction of subcellular localization of Gram-negative bacterial proteins. Mol. Divers. 13, 475–481.

Witten, I.H., Frank, E., 2005. Data mining: practical machine learning tools and techniques, 2nd ed. Morgan Kaufmann, San Francisco.

Wu, L.C., Lee, J.X., Huang, H.D., Liu, B.J., Horng, J.T., 2009. An expert system to predict protein thermostability using decision tree. Expert Syst. Appl. 36, 9007–9014.

Yuan, Y., Shi, X., Li, X., Lu, W., Cai, Y., Gu, L., Liu, L., Li, M., Kong, X., Xing, M., 2009. Prediction of interactiveness of proteins and nucleic acids based on feature selections. Mol. Divers. doi:10.1007/s11030-009-9198-9.

Zhang, G., Fang, B., 2006a. Application of Amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins. Process Biochem. 41, 1792–1798.

Zhang, G., Fang, B., 2006b. Discrimination of thermophilic and mesophilic proteins via pattern recognition methods. Process Biochem. 41, 552–556.

Zhang, G., Fang, B., 2006c. Support vector machine for discrimination of thermophilic and mesophilic proteins based on amino acid composition. Protein Pept. Lett. 13, 965–970.

Zhang, G., Fang, B., 2007. LogitBoost classifier for discriminating thermophilic and mesophilic proteins. J. Biotechnol. 127, 417–424.

Zhou, X.X., Wang, Y.B., Pan, Y.J., Li, W.F., 2008. Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. Amino Acids 34, 25–33.