Amino Acids (2012) 42:1947–1953
DOI 10.1007/s00726-011-0923-1

ORIGINAL ARTICLE

Detecting thermophilic proteins through selecting amino acid and dipeptide composition features

Songyot Nakariyakul · Zhi-Ping Liu · Luonan Chen

Received: 25 January 2011 / Accepted: 20 April 2011 / Published online: 6 May 2011

Abstract Detecting thermophilic proteins is an important task for designing stable protein engineering in interested temperatures. In this work, we develop a simple but efficient method to classify thermophilic proteins from mesophilic ones using the amino acid and dipeptide compositions. Since most of the amino acid and dipeptide compositions are redundant, we propose a new forward floating selection technique to select only a useful subset of these compositions as features for support vector machine-based classification. We test the proposed method on a benchmark data set of 915 thermophilic and 793 mesophilic proteins. The results show that our method using 28 amino acid and dipeptide compositions achieves an accuracy rate of 93.3% evaluated by the jackknife cross-validation test, which is higher not only than the existing methods but also than using all amino acid and dipeptide compositions.

Keywords Amino acid composition · Dipeptide composition · Feature selection · Floating search method · Protein thermostability

Introduction

Protein thermostability plays a crucial role in protein engineering and biotechnological research (Pokala and Handel 2001; Bommarius et al. 2006). Proteins produced by thermophilic organisms are extremely stable and can tolerate up to the temperature of more than 80°C, whereas mesophilic proteins are unstable under high temperature. Many experiments have been carried out to study the chemical properties that influence the stability of thermophilic proteins (Szilagyi and Zavodsky 2000; Kumar and Nussinov 2001; Yano and Poulos 2003; Razvi and Scholtz 2006). Gromiha et al. (1999) showed that the Gibbs free energy change of hydration and shape influenced the thermostability of proteins. The number of salt bridges (Kumar et al. 2000) and ion pairs (Kumar et al. 2001) in thermophilic proteins could also enhance the stability. Furthermore, protein stability depends linearly on the chain length (Ghosh and Dill 2009) and protein rigidity (Radestock and Gohlke 2008). However, experimental determination of the protein thermostability is time-consuming and labor-intensive. Thus, a computational method to determine thermophilic proteins is demanded.

Several methods have been proposed to determine the thermostability of a given protein from its primary sequence (Zhang and Fang 2006a, 2006b, 2007; Gromiha and Suresh 2008; Montanucci et al. 2008; Wu et al. 2009). Zhang and Fang (2006b) developed a statistical method for discriminating thermophilic proteins from mesophilic ones based on the dipeptide compositions of a given protein.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-011-0923-1) contains supplementary material, which is available to authorized users.

S. Nakariyakul · Z.-P. Liu · L. Chen (✉)
Key Laboratory of Systems Biology,
SIBS-Novo Nordisk Translational Research Centre
for PreDiabetes, Shanghai Institutes for Biological Sciences,
Chinese Academy of Sciences, Shanghai 200031, China
e-mail: lnchen@sibs.ac.cn

S. Nakariyakul (✉)
Department of Electrical and Computer Engineering,
Thammasat University, Khongluang,
Pathumthani 12120, Thailand
e-mail: nsongyot@engr.tu.ac.th

The method yielded an overall accuracy of 86% for classifying 3,521 thermophilic and 4,895 mesophilic protein sequences. Gromiha and Suresh (2008) analyzed the amino acid compositions of 1,609 thermophilic and 3,075 mesophilic proteins and found that many charged and hydrophobic residues have higher occurrence in thermophilic proteins than in mesophilic ones. The neural network-based method based on the amino acid compositions could successfully obtain a fivefold cross-validation accuracy of 89%. Montanucci et al. (2008) employed a support vector machine (SVM)-based method to predict whether a given protein mutant is thermostable. The method correctly classified 12 mutated proteins out of 14 (86% accuracy). Wu et al. (2009) considered both structure and sequence features of proteins to predict protein thermostability. Using a decision tree, they achieved an overall accuracy of more than 80%. Although the sequence and structural models obtained a slightly higher accuracy, Wu et al. suggested that sequence-only models can provide sufficient accuracy for thermostability prediction.

To further improve the accuracy, we develop a new computational method to discriminate thermophilic proteins from mesophilic ones by considering to use both amino acid and dipeptide compositions, since these compositions give useful information for classification (Zhang and Fang 2006b, 2007; Gromiha and Suresh 2008). However, many of the 420 compositions (20 amino acids and 400 dipeptides) are redundant and cannot significantly contribute to the thermostability. We thus propose a simple but efficient method using feature selection to identify only a small number of amino acid and dipeptide compositions for classification. Our proposed method, i.e., improved forward floating selection (IFFS) algorithm (Nakariyakul and Casasent 2009), is employed for feature selection, and an SVM classifier is built as a predictor. We find that only 28 features of these compositions are needed and that the accuracy of jackknife cross-validation test achieves 93.3%, which is considerably higher not only than the existing methods but also than using all amino acid and dipeptide compositions. In addition, we discuss the significance of features selected by our method and their functional implications.

Materials and methods

Datasets

We used the benchmark thermophilic and mesophilic dataset provided by Lin and Chen (2011), which is available at http://cobi.uestc.edu.cn/people/hlin/tools/ThermoPred. The dataset was extracted from the Universal Protein Resource (UniProt) (http://www.uniprot.org) containing protein sequences of 136 prokaryotic organisms (17 archaea and 119 bacteria). 60°C was used as the lower limit of optimal growth temperature for thermophilic organisms, and 30°C was set as the upper limit of optimal growth temperature for non-thermophilic organisms. To produce a reliable dataset, protein sequences containing ambiguous residues (such as "X", "B" and "Z") were removed. Proteins that are fragment of other proteins or that infer from prediction or homology were also excluded. After the described procedure, 1,329 thermophilic and 1,250 mesophilic proteins were obtained. Furthermore, the CD-HIT program (Huang et al. 2010) was used to remove redundancy and homology bias, so that no two sequences are similar more than 40%. The final dataset contains 915 thermophilic and 793 mesophilic proteins for testing. The length of thermophilic protein sequences ranges from 27 to 1,853 with an average of 318.32 and a standard deviation of 221.10, while the length of mesophilic protein sequences ranges from 31 to 3,567 with an average of 334.85 and a standard deviation of 286.61. Prior work (Zhang and Fang 2006b; Gromiha and Suresh 2008) did not remove highly similar or homologous sequences from their data sets. Thus, the data set used in this study is deemed to be more reliable.

Amino acid and dipeptide compositions

The amino acid composition for each amino acid of a protein is computed as the number of that amino acid divided by the total number of all residues in the protein. In other words, it is defined as:

$$\mathrm{Comp}(i) = \frac{n_i}{\sum_{i=1}^{20} n_i},\ 1 \le i \le 20 \tag{1}$$

where i stands for the 20 amino acids and n_i is the number of residues of amino acid i in the protein. The composition of all the 400 dipeptides of each protein sequence is computed using the following expression (Shen and Chou 2008):

$$\mathrm{Comp_d}(i,j) = \frac{n_{ij}}{L-1},\ 1 \le i,j \le 20 \tag{2}$$

where i, j stand for the distribution of amino acid i followed by amino acid j, n_{ij} is the number of residues of amino acid i followed by amino acid j, and L is the total number of residues in the protein sequence. The amino acid and dipeptide compositions have been widely used in prior protein-related work (Zhang and Fang 2006b, 2007; Gromiha and Suresh 2008).

Feature selection

Feature selection refers to search algorithms that select a subset of m features from an initial set of n features, where a criterion function J is used to assess the quality of each

candidate subset. Feature selection methods can be roughly categorized as filter (Yu and Liu 2003) or wrapper (Kohavi and John 1997). Filter methods select feature subset based mainly on the intrinsic properties of the data such as distance, dependency, and consistency and without any knowledge of the learning algorithm. On the other hand, wrapper methods find subsets that maximize the performance of a predetermined learning algorithm. The wrapper method generally achieves better performance than the filter method, but it is also more computationally expensive. For our work, since there are a total of 420 initial features (20 amino acids and 400 dipeptides), we propose a wrapper method to perform feature selection. The fivefold cross-validation accuracy rate is used as the criterion function because our dataset is balanced (915 thermophilic and 793 mesophilic proteins). In many applications with imbalanced datasets, different criterion functions such as the area under the receiver operating characteristic (AUC) curve and the area under the precision-recall curve (PRC) should be considered (Wasikowski and Chen 2010).

Several feature selection techniques have been proposed in the literature. Sequential forward selection (SFS) (Whitney 1971) and sequential backward selection (SBS) (Marill and Green 1963) algorithms are widely used for their simplicity and speed. The SFS method is a greedy search strategy that starts with an empty set and at each iteration, one feature is added to the subset, so that the resultant subset yields the best criterion function value. The SBS algorithm starts with all input features n and removes one feature at a time from the feature set until the desired number of features m is obtained. These methods are attractive but suffer from the "nesting effect", i.e., once the features are selected, they cannot be discarded from the current feature subset, and vice versa.

Sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS) methods (Pudil et al. 1994) efficiently overcome the nesting problem by dynamically backtracking after each sequential step. The SFFS algorithm starts the search with an empty feature set and uses the SFS algorithm to add one feature at a time to the selected feature subset. Every time a new feature is added to the current feature subset, the algorithm backtracks using the SBS algorithm to remove one feature at a time from the subset to check whether a better subset can be located. The search terminates when the size of the current feature set is larger than the desired number m of features to allow sufficient backtracking. The SBFS method starts with all input features n, removes one feature at a time, and conditionally adds a feature to the resultant subset as long as a better subset can be located.

Our proposed IFFS algorithm is actually an improvement of the SFFS algorithm, and its simplified flowchart is shown in Fig. 1. First, it starts with an empty feature subset

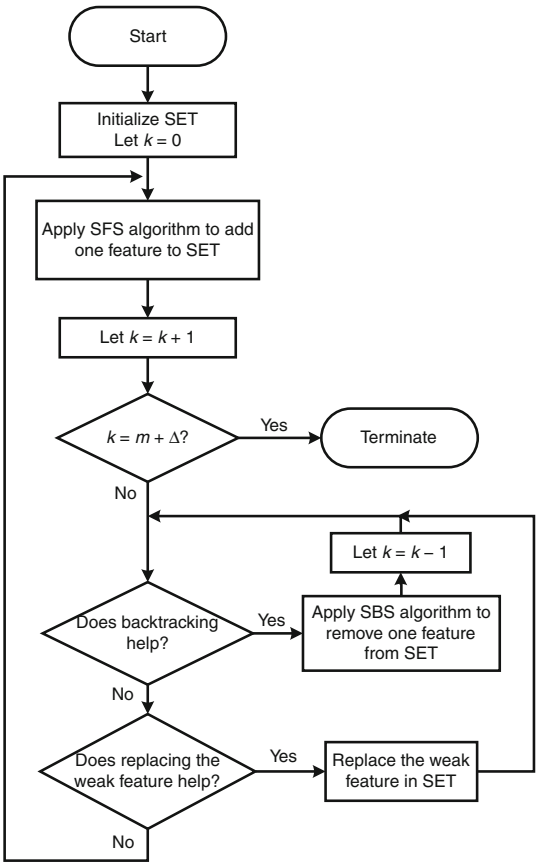

Fig. 1 The simplified flowchart of the IFFS algorithm

and applies the greedy SFS algorithm to add one feature at a time, so that the resultant subset yields the best cross-validation accuracy rate. Like the SFFS algorithm, every time a new feature is added to the current feature subset, the IFFS algorithm backtracks by applying the SBS algorithm to check whether removing a feature from the selected subset can improve the result at the prior iteration. As a new step in the IFFS algorithm, after the search stops backtracking, IFFS attempts to exchange a feature in the currently selected subset with any discarded feature to improve the accuracy rate. The IFFS algorithm was shown to perform better than prior search algorithms on many datasets (Nakariyakul and Casasent 2008, 2009).

The steps in the IFFS algorithm to select the best subset of m features from the set Y of n features can be summarized as follows.

- SET: the current feature subset being evaluated.
- k: the number of features in SET.
- J_k: the best fivefold cross-validation accuracy rate found so far for a subset of k features.

Step 0. (Initialization): Set $k = 0$ and SET $= \emptyset$ (empty set).

Step 1. (Adding a feature to the set): Use the SFS method to add a feature to SET, increase k by 1, and update J_k if necessary. If $k = m + \Delta$, terminate the algorithm.

Step 2. (Backtracking): Conditionally remove the least significant feature from SET by applying the SBS method to SET. If the resultant subset has the best J_{k-1} found so far, update J_{k-1} and SET, decrease k by 1, and repeat Step 2. Else, return the conditionally removed feature to SET and go to Step 3.

Step 3. (Replacing the weak feature): Conditionally remove feature x_i from SET, apply the SFS method to add a new feature to the resultant subset to obtain a new SET_i, and compute J of SET_i, for $1 \leq i \leq k$. Let SET_s be the subset that yields the largest J value among the new k SET_i subsets. If SET_s has the best J_k found so far, update J_k and SET, and go to Step 2. Otherwise, go to Step 1.

Here, Δ is the user-specified parameter that allows the search to adequately backtrack. We set Δ to five for the present work. A more detailed description of IFFS can be found in (Nakariyakul and Casasent 2009).

Machine learning technique and performance evaluation

We choose the SVM classifier with a radial basis function to perform the classification (Chen et al. 2009, 2010). The regularization parameter C and kernel parameter γ in the SVM are optimized using a grid search approach. To minimize the overfitting of the prediction model, a fivefold cross-validation process is implemented. The software LIBSVM version 3.0 is employed in this work and is freely available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. We use sensitivity (Sn), specificity (Sp), and accuracy (Acc) to measure the performance of our method, which are defined as follows.

$$\text{Sn} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

where true positive (TP) is the number of correctly classified thermophilic proteins. True negative (TN) is the number of correctly classified mesophilic proteins. False positive (FP) is the number of mesophilic proteins misclassified as thermophilic proteins, and False negative (FN) is the number of thermophilic proteins misclassified as mesophilic proteins.

Results and discussion

Feature selection results

In this subsection, we discuss the experimental results using the proposed IFFS algorithm. We compare our results with those for the SFS algorithm, the SFFS algorithm, and the minimum redundancy maximal relevance (mRMR) method. The mRMR algorithm is a well-known filter method that selects feature subsets based on mutual information (Peng et al. 2005). After the mRMR method selected a feature subset, samples with the selected features were applied into the SVM classifier for classification. We ran all of our experiments using MATLAB 7.10 on an Intel Core i7-860 computer with 4 GB of RAM. Figure 2 shows the results of the four algorithms evaluated by fivefold cross-validation. As expected, the mRMR method gives the lowest Acc rates among the four algorithms in almost all m cases, since it is a filter method. The SFFS algorithm achieves the Acc rates that are higher than or equal to those of the SFS algorithm when $1 \leq m \leq 11$. When $m \geq 12$, the SFS algorithm performs slightly better than the SFFS algorithm. As shown in Fig. 2, the proposed IFFS algorithm is consistently superior to other search algorithms. The Acc rates of IFFS increase as m increases from 1 to 28 and saturate when m is larger than 28. Thus, we chose to keep 28 features for testing and obtained an Acc rate of 93.9%.

In terms of computational complexity, the IFFS algorithm is noticeably more time-consuming than other search algorithms. To select 40 out of 420 features, IFFS demanded more than 2 days, SFFS needed about 4 h, SFS required approximately 90 min, and mRMR took only

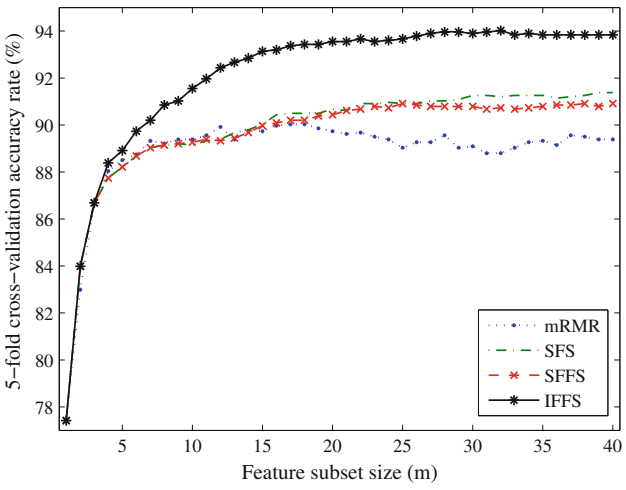

Fig. 2 The fivefold cross-validation accuracy rates obtained by mRMR, SFS, SFFS, and IFFS for different numbers of selected features

1 min (since it is a filter method). We note that feature selection is usually performed off-line and that time required by IFFS is not a major issue for this dataset. However, the computational complexity of the IFFS algorithm increases exponentially with the total number of features. Thus, use of IFFS becomes impractical for very high-dimensional datasets such as microarray data. In such cases, filter methods such as the mRMR algorithm should be considered.

The 28 features selected by our IFFS algorithm are now described. These features include 9 amino acids and 19 dipeptides as shown in Table 1. We found that amino acids Lys (K) and Glu (E) were chosen. They are known to participate in salt bridges (charge–charge interactions between oppositely charged residues), which contribute significantly to protein thermostability (Querol et al. 1996). The residues Ala (A) and Gln (Q) were found to be statistically different between mesophilic and thermophilic proteins (Gromiha and Suresh 2008). The box plots of these compositions in thermophilic and mesophilic proteins for our dataset are shown in Fig. 3. Cys (C) was also reported to occur less frequently in thermophiles than mesophiles (Saraboji et al. 2005). Furthermore, many selected dipeptide compositions in Table 1 contain these particular amino acids; in these 19 dipeptides, Cys (C) and Glu (E) appear five and three times, respectively.

Comparison with other attributes

To assess the performance of our selected features, we compared the results with other attributes, i.e., 20 amino acids, 400 dipeptides, and 420 amino acids and dipeptides. The SVM classifier with a radial basis function was used to perform the classification for all models. We analyzed the performance of each attribute using fivefold, tenfold, and jackknife cross-validation tests. Among these tests, the jackknife cross-validation is deemed the most objective and rigorous one. The results are presented in Table 2. From Table 2, the 20 amino acid compositions give higher Acc rates than the dipeptide compositions and the combination of amino acid and dipeptide compositions for all three tests. These similar findings were also reported in (Gromiha and Suresh 2008). Our method outperformed other models in terms of Sn, Sp, and Acc by approximately 1–4% for all tests. Generally, even a slight increase in Acc rate is always desirable and crucial in many applications. When the jackknife cross-validation test was performed, our method gave a high Acc rate of 93.3%. The SVM parameters used for each attribute are provided in Supplementary Material.

In addition, we considered applying feature selection to select features only from amino acid compositions. Our IFFS algorithm selected 15 out of 20 amino acid compositions and obtained a fivefold cross-validation Acc rate of 92.6%, which is 0.6% higher than using all 20 amino acid compositions for classification. However, this result is lower than 93.9% obtained by using our 28 selected amino acid and dipeptide compositions. This clearly indicates that dipeptide features contribute significantly to discriminate thermophilic proteins from mesophilic ones.

To verify the reliability of the proposed method, we tested it with an independent dataset that has never been trained before. This data set was used in prior work (Zhang and Fang, 2006a) and contains 76 thermophilic and 81 mesophilic proteins. Our model correctly identified 71 out of 76 thermophilic proteins with the Sn of 93.4%. 65 of 81

Table 1 Twenty-eight amino acid and dipeptide compositions selected by the IFFS algorithm

Types	Features
Amino acids	A, C, D, E, G, K, Q, S, T
Dipeptides	CP, CW, DM, EC, EI, GN, HH, HW, IP, MA, QW, RI, RK, SN, VC, VI, WE, WV, YC

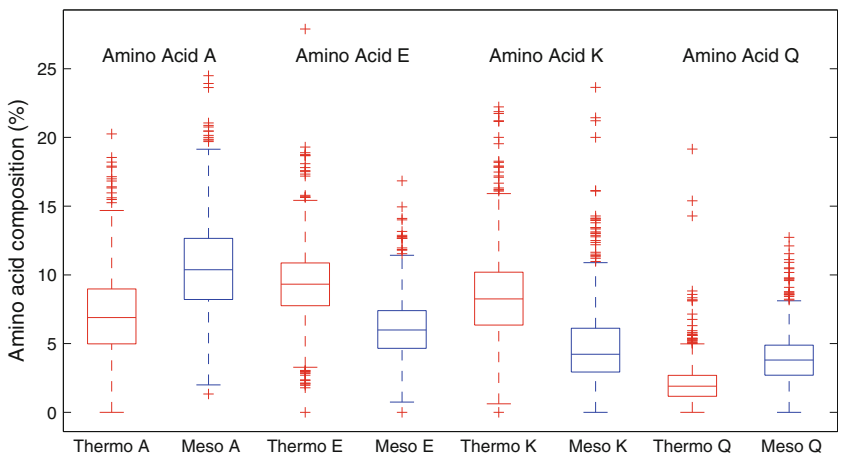

Fig. 3 The box plots of the amino acid compositions A, E, K, and Q in thermophilic and mesophilic proteins

Table 2 Test results for the prediction of thermophilic and mesophilic proteins using different models

Models	Number of features	Cross-validation test (%)								
		Fivefold			Tenfold			Jackknife		
		Sn	Sp	Acc	Sn	Sp	Acc	Sn	Sp	Acc
Amino acids	20	92.1	91.9	92.0	92.1	91.8	91.9	92.4	92.3	92.3
Dipeptides	400	90.0	91.9	90.9	90.2	92.4	91.2	90.8	91.7	91.2
Amino acids and dipeptides	420	90.0	93.0	91.4	90.2	92.3	91.2	91.5	91.8	91.6
IFFS-selected amino acids and dipeptides	28	93.8	94.1	93.9	93.1	93.2	93.1	93.0	93.7	93.3

mesophilic proteins were correctly classified with the Sp of 80.3%. The overall Acc of 86.6% was obtained. These results demonstrate that our proposed method generalizes well with independent data. Note that more tests should be carried out with larger datasets to further validate the effectiveness of our method and that considering heterogeneous information (Chen et al. 2009, 2010) may also improve the accuracy.

Conclusions

We proposed a novel method by applying feature selection to choose only a useful subset of the amino acid and dipeptide compositions for classifying thermophilic and mesophilic proteins. Our IFFS feature selection algorithm is an improvement on the SFFS algorithm. This simple but efficient algorithm was shown to outperform other wrapper methods and the mRMR algorithm. The SVM-based predictor using our selected 28 amino acids and dipeptides distinguished the thermophilic and mesophilic proteins at the jackknife cross-validation accuracy of 93.3%, which is superior to other existing methods. The software used in this work is available upon request.

Acknowledgments This work was supported by the Chinese Academy of Sciences Fellowship for Young International Scientist with Grant No. 2010Y1Sb10 and NSFC with Grant No. 31050110435 (S. Nakariyakul). This work was also supported by the Chief Scientist Program of Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences with Grant No. 2009CSP002 (L. Chen), and supported by NSFC under Grants No. 61072149 and No. 91029301 (L. Chen and Z.P. Liu), the Knowledge Innovation Program of CAS with Grant No. KSCX2-EW-R-01 (L. Chen and Z.P. Liu), and supported by the Key Project of Shanghai Education Committee (B.10-0412-08-001), Japan (JSPS) FIRST Program (L. Chen) and Shanghai Natural Science Foundation under Grant No. 11ZR1443100 (Z.P. Liu).

References

Bommarius AS, Broering JM, Chapparro-Riggers JF, Polizzi KM (2006) High-throughput screening for enhanced protein stability. Curr Opin Biotechnol 17:606–610

Chen L, Wang RS, Zhang X (2009) Biomolecular network: methods and applications in systems biology. Wiley, London

Chen L, Wang RQ, Li C, Aihara K (2010) Modelling biomolecular networks in cells: structures and dynamics. Springer, Berlin

Ghosh K, Dill KA (2009) Computing protein stabilities from their chain lengths. Proc Natl Acad Sci USA 106:10649–10654

Gromiha MM, Suresh MX (2008) Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins 70:1274–1279

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

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

Kohavi R, John G (1997) Wrappers for feature subset selection. Artif Intell 97:273–324

Kumar S, Nussinov R (2001) How do thermophilic proteins deal with heat? Cell Mol Life Sci 58:1216–1233

Kumar S, Tsai CJ, Nussinov R (2000) Factors enhancing protein thermostability. Protein Eng 13:179–191

Kumar S, Tsai CJ, Nussinov R (2001) Thermodynamic differences among homologous thermophilic and mesophilic proteins. Biochemistry 40:14152–14165

Lin H, Chen W (2011) Prediction of the thermophilic proteins using feature selection technique. J Microbiol Methods 84:67–70

Marill T, Green DM (1963) On the effectiveness of receptors in cognition system. IEEE Trans Inform Theory 9:11–17

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

Nakariyakul S, Casasent D (2008) Hyperspectral waveband selection for contaminant detection on poultry carcasses. Opt Eng 47:087202

Nakariyakul S, Casasent D (2009) An improvement on floating search algorithms for feature subset selection. Pattern Recog 42:1932–1940

Peng HC, Long FH, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intel 27:1226–1238

Pokala N, Handel TM (2001) Protein design-where we were, where we are, where we're going. J Struct Biolo 134:269–281

Pudil P, Novovicova J, Kittler J (1994) Floating search methods in feature selection. Pattern Recogn Lett 15:1119–1125

Querol E, Perez-Pons JA, Mozo-Villarias A (1996) Analysis of protein conformational characteristics related to thermostability. Protein Eng 9:265–271

Radestock S, Gohlke H (2008) Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. Eng Life Sci 8:507–522

Razvi A, Scholtz JM (2006) Lessons in stability from thermophilic proteins. Protein Sci 15:1569–1578

Saraboji K, Gromiha MM, Ponnuswamy MN (2005) Importance of main-chain hydrophobic free energy to the stability of thermophilic proteins. Int J Biol Macromol 35:211–220

Shen HB, Chou KC (2008) PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 373:386–388

Szilagyi A, Zavodsky P (2000) Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Struct Fold Des 8:493–504

Wasikowski M, Chen X-W (2010) Combating the small sample class imbalance problem using feature selection. IEEE Trans Knowl Data Eng 22:1388–1400

Whitney AW (1971) A direct method of nonparametric measurement selection. IEEE Trans Comput 20:1100–1103

Wu LC, Lee JX, Huang HD, Liu BJ, Horng JT (2009) An expert system to predict protein thermostability using decision tree. Expert Syst Appl 36:9007–9014

Yano JK, Poulos TL (2003) New understandings of the thermostable and peizostable enzymes. Curr Opin Biotechnol 14:360–365

Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of the 20th International Conference on Machine Learning, AAAI Press, Menlo Park, pp 56–63

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

Zhang G, Fang B (2006b) Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins. Process Biochem 41:1729–1798

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