Amino Acids (2013) 44:573–580
DOI 10.1007/s00726-012-1374-z

ORIGINAL ARTICLE

A similarity distance of diversity measure for discriminating mesophilic and thermophilic proteins

Yong-Chun Zuo · Wei Chen · Guo-Liang Fan · Qian-Zhong Li

Received: 28 November 2011 / Accepted: 17 July 2012 / Published online: 1 August 2012

Abstract The successful prediction of thermophilic proteins is useful for designing stable enzymes that are functional at high temperature. We have used the increment of diversity (ID), a novel amino acid composition-based similarity distance, in a 2-class *K*-nearest neighbor classifier to classify thermophilic and mesophilic proteins. And the KNN-ID classifier was successfully developed to predict the thermophilic proteins. Instead of extracting features from protein sequences as done previously, our approach was based on a diversity measure of symbol sequences. The similarity distance between each pair of protein sequences was first calculated to quantitatively measure the similarity level of one given sequence and the other. The query protein is then determined using the *K*-nearest neighbor algorithm. Comparisons with multiple recently published methods showed that the KNN-ID proposed in this study outperforms the other methods. The improved predictive performance indicated it is a simple and effective classifier for discriminating thermophilic and mesophilic proteins. At last, the influence of protein length and protein identity on prediction accuracy was discussed further. The prediction model and dataset used in this article can be freely downloaded from http://wlxy.imu.edu.cn/college/biostation/fuwu/KNN-ID/index.htm.

Keywords Thermophilic protein · Increment of diversity · *K*-nearest neighbor · Amino acids · Prediction performance

Electronic supplementary material The online version of this article (doi:10.1007/s00726-012-1374-z) contains supplementary material, which is available to authorized users.

Y.-C. Zuo (✉) · G.-L. Fan · Q.-Z. Li (✉)
School of Physical Science and Technology,
Inner Mongolia University, Hohhot 010021, China
e-mail: yczuo@imu.edu.cn

Q.-Z. Li
e-mail: qzli@imu.edu.cn

W. Chen
Center of Genomics and Computational Biology,
College of Sciences, Hebei United University,
Tangshan 063000, China

Introduction

The temperature of the environment plays a crucial role in the lives of the cell (Perutz and Raidt 1975; Thompson and Eisenberg 1999). It is an ongoing interest in understanding the stability mechanism of the proteins in the organisms which are living in the so-called harsh environments such as high pressure, high temperature, and non-physiological pH (Pokala and Handel 2001; Bommarius et al. 2006; Huber et al. 2000; Kawashima et al. 2000). The protein production of thermophilic organisms is extremely stable, tolerating up to the temperature of more than 80 °C. But the mesophilic proteins are unstable under high temperature (Sadeghi et al. 2006; Vieille and Zeikus 2001; Li et al. 2005; Haney et al. 1999; Szilagyi and Zavodszky 2000). The protein thermostability refers to the stability of the unique chemical and spatial structure of a polypeptide chain under the extreme temperature conditions (Zhou et al. 2008; Bartesaghi et al. 2007; Schmidinger et al. 2006; Kumar et al. 2000; Elcock 1998). Discrimination of thermophilic and mesophilic proteins by computational recognition algorithms provided a new thought for the theoretical description of protein folding and stability (Zhou et al. 2008).

Data mining technique can quickly provide insights when doing basic research. In theory, it is also helpful for understanding the mechanism of protein thermostability,

providing a quantitative model to explain the sequence—characteristic relationship (Fukuchi and Nishikawa 2001; Montanucci et al. 2008; Zeldovich et al. 2007). Many effective computation classifiers have been developed in the past decade. The correlation between the optimal growth temperatures of the genomes and the occurrences of the amino acid coupling patterns was found by Liang et al. (Liang et al. 2005). Zhang and Fang constructed the first benchmark dataset of thermophilic and mesophilic proteins based on Swiss-Prot database (Zhang and Fang 2006a). In the following, the four pattern recognition methods, namely, principal component analysis (PCA), stepwise regression (SR), partial least-square regression (PLSR), and backpropagation neural network, were developed to discriminate thermophilic and mesophilic proteins based on amino acid contents (Zhang and Fang 2006b). Subsequently, the LogitBoost classifier and support vector machine (SVM) program were proposed based on the same dataset. The best accuracy achieved to 87.39 % (Zhang and Fang 2007). Gromiha and Suresh constructed the first low-similarity dataset with 40 % sequence identity. And the overall accuracy increased to 89.40 % using several machine learning algorithms implemented in Waikato environment for knowledge analysis (WEKA) program based on amino acid composition(Gromiha and Suresh 2008). Recently, some other studies that obtained a slight improvement in prediction accuracy (Li and Fang 2010; Lin and Chen 2011; Wang et al. 2011; Nakariyakul et al. 2012) were reported. All of above computational methods demonstrated that the basic amino acid composition provides sufficient accuracy for thermostability prediction.

On the basis of the Shannon entropy definition, the diversity measure was introduced to describe the information on discrete state space and whole uncertainly of system (Laxton 1978). And the increment of diversity (ID) was defined to compare the difference between the total diversity measure of two systems and the diversity measure of the mixed system (Li and Lu 2001). And the increment of diversity (ID) has been successfully applied in biological data classification extensively, e.g., the promoter prediction (Zuo and Li 2011), the recognition of protein structural class (Lin and Li 2007), the protein superfamily classification (Zuo and Li 2009), the protein subnuclear location (Li and Li 2008), and the prediction of secretory proteins (Zuo and Li 2010).

In this study, based on the *K*-nearest neighbor (KNN) method and the Shannon entropy definition, we first developed the *K*-nearest neighbor increment of diversity (KNN-ID) classifier to discriminate thermophilic and mesophilic proteins based on amino acid contents. The prediction performance outperformed the other current methods. The best overall accuracy increased to 91.02 % when using the 20 amino acid composition. The influence of sequence length and sequence similarity on predictive performance was also discussed. The good results indicated that the KNN-ID method is an efficient program for discriminating thermophilic and mesophilic proteins. We believed that it is also useful for other protein function classification.

Materials and methods

Datasets

To have a consistent comparison between different approaches, two public benchmark datasets were selected for evaluating the performance of the proposed method. The first dataset contains 4,895 mesophilic proteins and 3,522 thermophilic proteins, which was constructed by Zhang and Feng (2006a). Protein sequences in this dataset were retrieved from 15 thermophilic (hyperthermophilic) organisms and nine mesophilic organisms based on Swiss-Prot database. The second dataset was constructed by Gromiha and Suresh contains 3,075 mesophilic proteins and 1,609 thermophilic proteins (Gromiha and Suresh 2008). These datasets have the proteins with less than 40 % sequence identity by using the CD-HIT program (Li and Godzik 2006).

The introduction of diversity measure

For a discrete state space X with d dimensions X:{n_1, n_2, …, n_i, …, n_d}, n_i denotes the times of ith state, the Shannon information entropy (Shannon 1948), a measure of uncertain and denoted by $H(X)$, is defined as:

$$H(X) = -\sum_{i=1}^{d} P_i \log_b P_i \tag{1}$$

where $P_i = n_i/N$, P_i indicates probability of ith state.

According to the definition of information, the quantity of the measured diversity named diversity measure, denoted by $D(X)$, can be defined as

$$\begin{aligned} D(X) &= -\sum_{i=1}^{d} n_i \log_b P_i = -\sum_{i=1}^{d} n_i \log_b \frac{n_i}{N} \\ &= N \log N - \sum_{i=1}^{d} n_i \log_b n_i \end{aligned} \tag{2}$$

According to the definition of information entropy, combining the formula (1), we get

$$H(X) = -\sum_{i=1}^{d} P_i \log_b P_i = -\sum_{i=1}^{d} \frac{n_i}{N} \log_b \frac{n_i}{N} = \frac{1}{N} D(X) \tag{3}$$

So we have

$$D(X) = N \cdot H(X) \qquad N = \sum_{i=1}^{d} n_i. \tag{4}$$

$H(X)$ is the information entropy, which indicates a measure of the uncertainty associated with a random variable. The measure of diversity $D(X)$ in formula (4) means a kind of information description on state space and a measure of whole uncertainly and total information of a system (Laxton 1978).

The similarity distance derived from increment of diversity

For comparing our method of information distance with other method, we also select the amino acid compositions (AAC) as the inputting vector of a diversity source, which is defined in discrete state space with 20 dimensions, formulated as X: $\{n_1, n_2, \ldots, n_i, \ldots, n_{20}\}$. The n_i is the absolute occurrence frequency of each type of 20 amino acid composition for each protein.

In general, for the diversity sources of two different protein chains with 20 dimensions X:$\{n_1, n_2, \ldots, n_i, \ldots, n_{20}\}$ and Y:$\{m_1, m_2, \ldots, m_i, \ldots, m_{20}\}$, the combination diversity source can be described as $X + Y$:$\{n_1 + m_1, n_2 + m_2, \ldots, n_i + m_i, \ldots, n_{20} + m_{20}\}$, and the diversity measure of mix source can be calculated as

$$D(X+Y) = (N+M)\log(N+M) - \sum_{i=1}^{d}(n_i+m_i)\log_b(n_i+m_i) \times \left(N = \sum_{i=1}^{d} n_i, M = \sum_{i=1}^{d} m_i\right). \tag{5}$$

And it can be proved the combination measure of diversity is no smaller than the sum of every diversity measure:

$$D(X+Y) \geq D(X) + D(Y). \tag{6}$$

Thus we define the increment of diversity (ID) to quantitatively measure the similarity level of two different sources X and Y, denoted by ID(X, Y) as follows:

$$ID(X,Y) = D(X+Y) - D(X) - D(Y). \tag{7}$$

It can be proved that the *ID(X, Y)* satisfies:

$$0 \leq ID(X,Y) \leq D(N,M) \quad (D(N,M) = (N+M)\log(N+M) - N\log N - M\log M). \tag{8}$$

We can see that the increment of diversity (ID(X, Y)) satisfies nonnegative and symmetry; therefore, the increment of diversity is a quantitative measure of the similarity level for two diversity sources. The higher the similarity of two sources, the smaller the ID value (Zuo and Li 2010).

The K-nearest neighbor increment of diversity (KNN-ID) classifier

The K-nearest neighbor (K-NN) technique has become extremely popular for a variety of forest inventory mapping and estimation applications, such as protein subcellular localization (Chou and Shen 2006; Chou and Shen 2007a, b; Shen and Chou 2007a), protein structural classification (Shen et al. 2005; Zhang et al. 2008), protein fold pattern (Shen and Chou 2006), membrane protein type (Shen and Chou 2005; Shen et al. 2006), and enzyme classification (Shen and Chou 2007b). Much of this popularity may be attributed to the non-parametric, multivariate features of the technique, its intuitiveness, and its ease of use. The query protein should be classified by a majority vote of its neighbors, with the protein being assigned to the class most common amongst its K-nearest neighbors. K is a positive integer, typically small. If $K = 1$, then the protein is simply assigned to the class of its nearest neighbor. Although different distance measures can be used for this, such as Euclidean distance, Hamming distance, and Mahalanobis distance, in this paper, the similarity distance measure of increment of diversity is first introduced for predicting query protein.

Suppose there are N proteins (x_1, x_2 …, x_n), which have been classified into M categories (c_1, c_2,…, c_m). According to the KNN rule, the query protein X should be assigned to the category represented by a majority of its K-nearest neighbors. The similarity distance between X and x_i ($i = 1, 2, \ldots, n$) is defined as

$$ID(X, x_i) = D(X + x_i) - D(X) - D(x_i) \quad (i = 1, 2, \ldots, n). \tag{9}$$

The candidate protein will be predicted to belong to the μth category if

$$\mu = \arg\max_{c}\left\{\sum_{i=1}^{K} \Delta(X_i^*, x_m)\right\} \tag{10}$$

where μ is the argument of m that maximizes $\{\sum_{i=1}^{K} \Delta(X_i^*, x_m)\}$, and

$$\Delta(X_i^*, x_m)\begin{cases} 1, \ If\ \mathrm{ID}(X_i^*, x_c) = \mathrm{Min}(\mathrm{ID}(X_i^*, x_1), \mathrm{ID}(X_i^*, x_2), ..\mathrm{ID}(X_i^*, x_c,)) \\ 0, \ \text{Otherwise} \end{cases}. \tag{11}$$

Performance measures and assessments

According to the Chou's review, the jackknife test is deemed the most objective and being able to yield a unique result in the statistical prediction (Chou and Shen 2007c; Chou 1995). The jackknife test was used to examine the power of the proposed method. The prediction performance was evaluated by the sensitivity (Sn), specificity (Sp), positive prediction value (PPV), accuracy (Ac), and Mathew's correlation coefficient (MCC), which are defined as follows:

$$\mathrm{Sn} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{12}$$

$$\mathrm{Sp} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \tag{13}$$

$$\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{14}$$

$$\mathrm{Ac} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}) \tag{15}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FP})}} \tag{16}$$

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

Evaluating the proposed method by comparing with other different machine learning approaches

Protein amino acid composition has long been thought to be correlated with its function biology (Gromiha et al. 1999; Das and Gerstein 2000; Cambillau and Claverie 2000; Karshikoff and Ladenstein 1998). Recently, more and more experimental, accumulation, and statistical analyses have also demonstrated that the properties of side chain of amino acids were determinants for the protein thermostability, though thermophilic and mesophilic proteins have both similar polar and nonpolar contribution to the surface area and compactness (Zhou et al. 2008; Farias and Bonato 2003; Suhre and Claverie 2003). For example, Ile, Arg, Glu, Lys, and Pro residue contents were found to be higher, while Ser, Asn, Gln, Thr, and Met were lower in thermophilic proteins (Vieille et al. 2001; Zhang and Fang 2006a; Gromiha and Suresh 2008). Thus, the amino acid composition from primary sequence is the most important feature parameter for the existing theoretical predictor.

During the process of jackknife testing, each protein in the benchmark dataset is singled out in turn as a test sample; the remaining proteins are used as training set to calculate test sample's membership and predict the category. After identifying the location of each protein using KNN algorithm, the prediction accuracy for the category will be calculated. The jackknife test is used to examine our method in the following study.

To investigate the best K value for predicting the thermophilic proteins, tests have been done with various values of nearest neighbors K (from 1 to 40), and the prediction accuracies obtained with jackknife test for the first data set are depicted in Fig. 1. The prediction results of jackknife test compared with SVM models based on the 20 amino acid compositions (AAC) are shown in the Table 1. For different values of K, it is shown that the prediction accuracy is improved along with the K increase, up to the best when K equals to nine, and the prediction accuracy has little reduction when the K continued to increase. The performance of prediction achieved 90.66 % accuracy (Ac) with 0.81 when $K = 9$, better than the best results achieved by the LogitBoost, AdaBoost, neural network, radial basis function (RBF), and SVM program classifiers. Therefore, in the following calculations, the $K = 9$ is used as the operation parameter.

It is interesting to know how the proposed method evaluated subsets of residue types. Further, the simplified amino acid alphabets were further applied to discriminate the mesophilic and thermophilic proteins (Murphy et al. 2000). The prediction results showed that the prediction accuracy of 4-letter alphabet performed better than the 5-letter alphabet and 6-letter alphabet (Supplementary Table). The 4-letter alphabet contains four amino acids groups, including large hydrophobic residues (LVIM), large aromatic residues (FYW), hydrophilic residues (EDNQKRH), and small residues (AGSTP). It demonstrated that the reduced alphabets have the ability in finding structurally conserved regions of protein with different optimum temperature.

The good performances of the datasets with lower sequence similarity

Since the sequence similarity may affect prediction accuracy, the highly similar data prefer leading to performance overestimation of the proposed methods. We further test the performance of our method on the second datasets with 40 % sequence similarity. Figure 2 shows the performance of various values of nearest neighbor K (from 1 to 40) based on jackknife test. Better results were obtained than the original dataset. The best results were also obtained when $K = 9$. The proposed KNN-ID method achieved

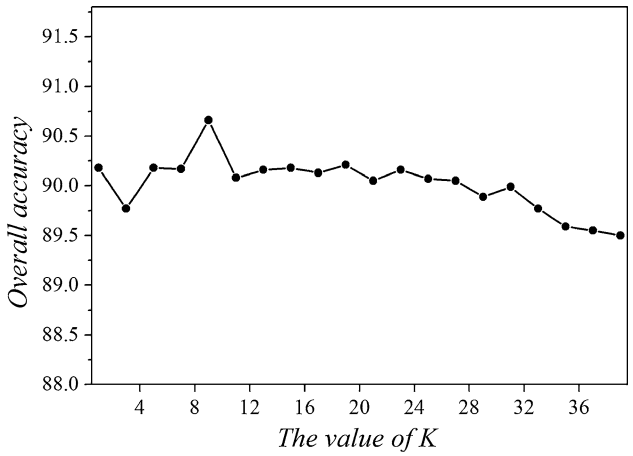

Fig. 1 Prediction accuracy of KNN-ID method by using different values of nearest neighbors K

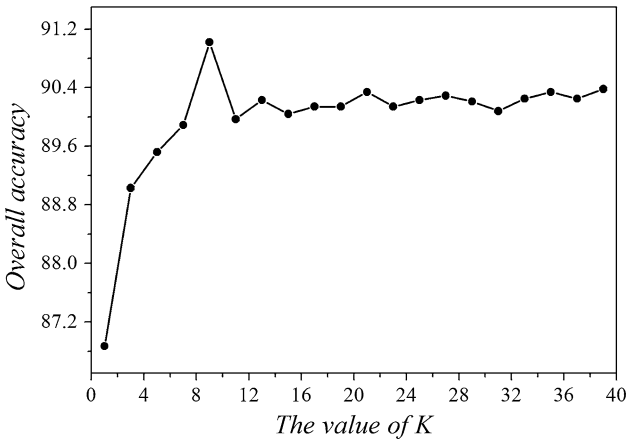

Fig. 2 Predictive trend of over accuracy at different values of K using KNN-ID method for low identity dataset

Table 1 The prediction results between the proposed method with the other methods on the first benchmark dataset

Method	Sn (%)	Sp (%)	PPV (%)	Acc (%)	MCC
KID ($K = 1$)	87.99	91.70	88.02	90.18	0.80
KID ($K = 9$)	**88.37**	**92.24**	**88.76**	**90.66**	**0.81**
LogitBoost	84.21	88.31	–	86.60	0.76
AdaBoost	77.19	86.58	–	82.65	0.64
RBF	84.01	86.39	–	85.40	0.70
SVM	83.87	89.93	–	87.39	0.74

The bold values show the best results

91.02 % accuracy (Ac). And their prediction accuracy does not vary significantly when $K > 10$.

This low sequence similarity dataset was first constructed by Gromiha and Suresh. Several machine learning programs had been evaluated on this dataset with 40 % sequence identity, such as Bayes rules, logistic functions, neural networks, support vector machines, decision trees and so forth (Gromiha and Suresh 2008). Most of the prediction accuracies range from 84.00 to 89.40 %. The best overall accuracy increased to 89.40 % when using neural network and logistic function method based on the amino acid compositions. The comparison results of our method with the other methods are shown in Table 2. Based on the same amino acid contents as the only input vectors, the traditional K-nearest neighbor (KNN) method obtained 84.80 % accuracy (Ac). The accuracy of the KNN-ID method proposed in this report was 91.02 %; more than 6 % improvement was obtained. Recently, Lin and Chen proposed the ANOVA feature selection technique to increase the accuracy of thermophilic proteins; the accuracy achieved was 90.8 %, based on 30 selected parameters (Lin and Chen 2011). Table 2 shows that the proposed method ($K = 9$) achieved 91.02 % accuracy, based on only 20 amino acids' composition, better than the results of Lin and Chen. This surprising predictive performance indicated that the KNN-ID method is indeed a good predictor for thermophilic proteins annotation.

The performance of the KNN-ID method for the protein with different similarities and lengths of sequence

How the similarity and the length of sequence affect the prediction performance is worth discussing. The distribution of different sequence identities analyzed using the CD-HIT program and the length distribution is shown in Fig. 3. The prediction accuracy of KNN-ID method, evaluating on the datasets with different sequence similarities and sequence lengths, are shown in Table 3.

From the Table 3, we could find the good results were obtained on the subsets of training data with different sequence similarities. With the sequence identity increasing, there is no significant change for the overall accuracies (Ac). Highly similar data will surely lead to overestimation of the performance of the proposed methods. The results will be more objective and reliable if the cutoff of sequence identity set to a lower percentage (such as 40 %). From the prediction results of our testing, we are pleased to see that the better prediction performance of our method was obtained based on the low sequence similarity. The lower similarity of primary sequence, the larger divergence of the 20 discrete amino acids. And the total measure of whole uncertainly for two different systems can easily be calculated by the ID algorithm. The comparison demonstrated that the increment of diversity (ID) is superior to integrate useful information on discrete state space. All of the overall accuracies were achieved at >90 % successful rate, and the best prediction accuracy (Ac) achieved 91.02 %. It indicated the KNN-ID method have the ability for predicting the low similarity thermophilic proteins.

Table 2 Comparisons between the proposed method with the other methods on the second benchmark dataset with low sequence identity

Method	Sn (%)	Sp (%)	PPV (%)	Acc (%)
Bayesnet	81.40	90.60	–	87.40
Naive Bayes	83.50	88.80	–	87.00
Logistic function	82.80	92.80	–	89.40
Neural network	82.40	93.00	–	89.40
RBF network	80.70	89.60	–	86.50
Support vector machines (SVM)	82.20	92.90	–	89.20
K-nearest neighbor	77.30	88.70	–	84.80
Bagging meta learning	80.00	92.00	–	87.90
Classification via regression	79.30	91.00	–	87.00
Decision tree J4.8	75.80	88.40	–	84.00
NBTree	79.20	89.50	–	86.00
Partial decision tree	81.50	85.20	–	83.90
SVM (Lin and Chen 2011)	85.40	93.60		90.80
KID ($K = 9$)	**84.27**	**94.53**	**88.90**	**91.02**

The bold values show the best results

Table 3 The prediction performance of the KNN-ID method for different sequence identities with $K = 9$

Identity (%)	Sn (%)	Sp (%)	PPV (%)	Acc (%)	MCC
40	84.27	94.53	88.90	91.02	0.80
50	85.07	93.04	87.64	90.12	0.79
60	85.68	92.76	88.29	90.00	0.79
70	87.01	92.12	88.18	90.06	0.79
80	87.77	92.00	88.37	90.27	0.80
90	88.15	91.92	88.32	90.38	0.80

The influence of protein length on discrimination accuracy was also discussed. We divided the proteins of first database into six groups and the prediction accuracy of jackknife test for each group is listed in Table 4. For the large-length proteins (>1,000 residues), the method achieved an overall accuracy of 92.91 %. For the proteins with residues between 500 and 800 amino acids, the accuracy of proposed method improved to 95.03 %. But the accuracy (Ac) for the small-size proteins (<100 residues) was rather moderate (81.04 %). The same results were obtained by the LogitBoost method in the study of Zhang and Fang (2007). Similar trends also existed in the process of discriminating globular and outer membrane (Gromiha 2005).

After analyzing the hydrophilicity and hydrophobicity of amino acid for the small protein (<100 AA) and large proteins (<500 AA), the results showed that the non-polar residues L, A, aromatic residues F, and polar charged (hydrophobicity) residues E, D are preferred to the large protein. And polar, uncharged residues C and polar charged residues *K*, *R* are preferred to the small protein (Supplementary Fig. 1). Compared with the large proteins (>500 AA), the statistical results of amino acids showed that the small proteins indeed have a lower percentage of hydrophobics residues (L, A, F etc.).

The bad moderate performance of KNN-ID for the small-size proteins might explain from the point of information science. The smaller size of a protein, the less information content it contains; using only the 20 feature vectors (20 AA compositions) might not represent its inherent characteristics. Based on this point, we consider it is reasonable to believe that the algorithm based on KNN-ID still has potential to improve as the other methods, especially for predicting the small-size proteins.

Further, we tested our KNN-ID classifier on 76 independent mesophilic-thermophilic protein pairs constructed by Zhang and Fang (2006b). In the previous study, the 92.11 % accuracies were obtained using LogitBoot and SVM programs. For our evaluation, each pair was decided

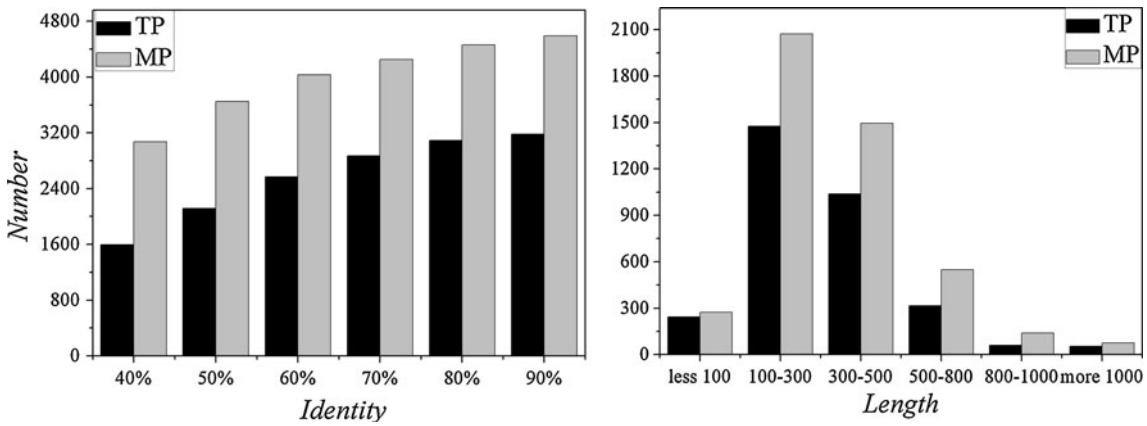

Fig. 3 The distributions of the identity and length of protein sequence

Table 4 The prediction performance of the KNN-ID method for different sequence lengths with $K = 9$

Length	Sn (%)	Sp (%)	PPV (%)	Ac (%)	MCC
Less 100	80.33	81.68	79.67	81.04	0.62
100–300	83.80	88.37	83.68	86.47	0.72
300–500	89.70	93.65	90.75	92.03	0.84
500–800	**90.54**	**97.63**	**95.67**	**95.03**	**0.89**
800–1,000	85.25	99.28	98.11	94.97	0.88
More 1,000	83.02	100.00	100.00	92.91	0.86

The bold values show the best results

by the minimum value calculated by $K - \mathrm{ID} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{ID}(X, Y)$, and our method obtained 97.37 % accuracy for the thermophilic protein and 92.11 % accuracy for the mesophilic protein. The accuracy achieved was 94.24 %, about 2 % improvement over the results of LogitBoot and SVM programs.

At last, the 20 independent proteins randomly selected from 19 different mesophilic archeal organisms, including *Methanococcus maripaludis*, *Methanosphaerula palustris E1-9c*, *Methanobrevibacter smithii*, etc., were selected to evaluate the proposed method. The results showed that all of the mesophilic archeal proteins were correctly predicted. Meanwhile, 20 independent proteins randomly selected from 11 different eubacterial extreme thermophiles, such as *Thermotoga maritima MSB8*, *Thermoanaerobacter italicus Ab9*, *Thermus thermophilus HB8*, etc., were also selected to evaluate the proposed method. Out of 20 proteins 18 were correctly predicted. It is shown that the KNN-ID method has the ability for solving some extreme eubacterial thermophiles and mesophilic archaea. The organism list and all the protein sequences have been uploaded on our website.

Conclusion

For protein prediction and classification, traditional developments of the methods for predicting protein functions generally focus on investigating new and effective mathematical descriptors of protein sequences. In this study, a new classifier using only the similarity distance of diversity measure was introduced to predict thermophilic proteins. Based on comparisons with several current methods for the same datasets with different sequence length and identity, the successful prediction performance indicates that the KNN-ID is a promising classifier. The accuracy of the proposed method outperformed the Naive Bayes, Logistic function, Neural network, RBF network, Support vector machines, Decision tree J4.8, and the traditional k-nearest neighbor. We believe our method can play a complementary role to existing experimental and computational methods for understanding the sequence—characteristic relationship of protein thermostability. We have constructed the internet server to facilitate other researchers. All the training datasets and the predictor based on the KNN-ID method can also be freely downloaded from http://wlxy.imu.edu.cn/college/biostation/fuwu/KNN-ID/index.htm.

Acknowledgments The authors would like to thank the reviewers for their helpful comments on their manuscript and the R. Verma for sharing the datasets. This work was supported by the High-level Scientific Research Foundation for the introduction of talent, Inner Mongolia University (No. 115115), the Research Fund for the Doctoral Program of Higher Education of China (No. 20101501110004) and the National Natural Science Foundation of China (61063016, 31160188).

References

Bartesaghi S, Ferrer-Sueta G, Peluffo G, Valez V, Zhang H, Kalyanaraman B, Radi R (2007) Protein tyrosine nitration in hydrophilic and hydrophobic environments. Amino Acids 32:501–515

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

Cambillau C, Claverie JM (2000) Structural and genomic correlates of hyperthermostability. J Biol Chem 275:32383–32386

Chou KC (1995) A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. Proteins Struct Funct Genet 21:319–344

Chou KC, Shen HB (2006) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res 5:1888–1897

Chou KC, Shen HB (2007a) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6:1728–1734

Chou KC, Shen HB (2007b) Large-scale plant protein subcellular location prediction. J Cell Biochem 100:665–678

Chou KC, Shen HB (2007c) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16

Das R, Gerstein M (2000) The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct Integr Genomics 1:76–88

Elcock AH (1998) The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. J Mol Biol 284:489–502

Farias ST, Bonato MC (2003) Preferred amino acids and thermostability. Genet Mol Res 2:383–393

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

Gromiha MM (2005) Motifs in outer membrane protein sequences: applications for discrimination. Biophy Chem 117:65–71

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

Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ (1999) Thermal adaptation analyzed by comparison of protein

sequences from mesophilic and extremely thermophilic Methanococcus species. Proc Natl Acad Sci USA 96:3578–3583
Huber R, Huber H, Stetter KO (2000) Towards the ecology of hyperthermophiles: biotopes, new isolation strategies and novel metabolic properties. FEMS Microbiol Rev 24:615–623
Karshikoff A, Ladenstein R (1998) Proteins from thermophilic and mesophilic organisms essentially do not differ in packing. Protein Eng 11:867–872
Kawashima T, Amano N, Koike H, Makino S, Higuchi S, Kawashima-Ohya Y, Watanabe K, Yamazaki M, Kanehori K, Kawamoto T, Nunoshiba T, Yamamoto Y, Aramaki H, Makino K, Suzuki M (2000) Archaeal adaptation to higher temperatures revealed by genomic sequence of Thermoplasma volcanium. Proc Natl Acad Sci USA 97:14257–14262
Kumar S, Tsai CJ, Nussinov R (2000) Factors enhancing protein thermostability. Protein Eng 13:179–191
Laxton RR (1978) The measure of diversity. J Theor Biol 71:51–67
Li Y, Fang J (2010) Distance-dependent statistical potentials for discriminating thermophilic and mesophilic proteins. Biochem Biophys Res Commun 396:736–741
Li WZ, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659
Li FM, Li QZ (2008) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids 34:119–125
Li QZ, Lu ZQ (2001) The prediction of the structural class of protein: application of the measure of diversity. J Theor Biol 213: 493–502
Li WF, Zhou XX, Lu P (2005) Structural features of thermozymes. Biotechnol Adv 23:271–281
Liang HK, Huang CM, Ko MT, Hwang JK (2005) The amino acid coupling patterns in thermophilic proteins. Proteins 59:58–63
Lin H, Chen W (2011) Prediction of thermophilic proteins using feature selection technique. J Microbiol Meth 84:67–70
Lin H, Li QZ (2007) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 28:1463–1466
Montanucci L, Fariselli P, Martelli PL, Casadio R (2008) Predicting protein thermostability changes from sequence upon multiple mutations. Bioinformatics 24:i190–i195
Murphy LR, Wallqvist A, Levy RM (2000) Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 13:149–152
Nakariyakul S, Liu ZP, Chen L (2012) Detecting thermophilic proteins through selecting amino acid and dipeptide composition features. Amino Acids 42:1947–1953
Perutz MF, Raidt H (1975) Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. Nature 255:256–259
Pokala N, Handel TM (2001) Protein design-where we were, where we are, where we're going. J Struct Biolo 134:269–281
Sadeghi M, Naderi-Manesh H, Zarrabi M, Ranjbar B (2006) Effective factors in thermostability of thermophilic proteins. Biophys Chem 119:256–270
Schmidinger H, Hermetter A, Birner-Gruenberger R (2006) Activity based proteomics: enzymatic activity profiling in complex proteomes. Amino Acids 30:333–350
Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423
Shen HB, Chou KC (2005) Using optimized evidence-theoretic *K*-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 334:288–292
Shen HB, Chou KC (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22:1717–1722
Shen HB, Chou KC (2007a) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Des Sel 20:39–46
Shen HB, Chou KC (2007b) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun 364:53–59
Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 334:577–581
Shen HB, Yang J, Chou KC (2006) Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. J Theor Biol 240:9–13
Suhre K, Claverie JM (2003) Genomic correlates of hyperthermostability, an update. J Biol Chem 278:17198–17202
Szilagyi A, Zavodszky P (2000) Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Struct Fold Des 8:493–504
Thompson MJ, Eisenberg D (1999) *Trans*proteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. J Mol Biol 290:595–604
Vieille C, Zeikus GJ (2001) Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev 65:1–43
Vieille C, Epting KL, Kelly RM, Zeikus JG (2001) Bivalent cations and amino-acid composition contribute to the thermostability of *Bacillus licheniformis* xylose isomerase. Eur J Biochem 268:6291–6301
Wang D, Yang L, Fu Z, Xia J (2011) Prediction of thermophilic protein with pseudo amino acid composition: an approach from combined feature selection and reduction. Protein Pept Lett 18:684–689
Zeldovich KB, Berezovsky IN, Shakhnovich EI (2007) Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol 3:e5
Zhang GY, Fang BS (2006a) Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins. Proc Biochem 41:1792–1798
Zhang GY, Fang BS (2006b) Discrimination of thermophilic and mesophilic proteins via pattern recognition methods. Proc Biochem 41:552–556
Zhang GY, Fang BS (2007) LogitBoost classifier for discriminating thermophilic and mesophilic proteins. J Biotechnol 127:417–424
Zhang TL, Ding YS, Chou KC (2008) Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. J Theor Biol 250:186–193
Zhou XX, Wang YB, Pan YJ, Li WF (2008) Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. Amino Acids 34:25–33
Zuo YC, Li QZ (2009) Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet. Peptides 30:1788–1793
Zuo YC, Li QZ (2010) Using *K*-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids. Amino Acids 38:859–867
Zuo YC, Li QZ (2011) Identification of TATA and TATA-less promoters in plant genomes by integrating diversity measure, GC-Skew and DNA geometric flexibility. Genomics 97:112–120