Computational Biology and Chemistry 46 (2013) 16–22

Computational Biology and Chemistry

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

Research article

Support vector machine with a Pearson VII function kernel for discriminating halophilic and non-halophilic proteins

Guangya Zhang*, Huihua Ge

Department of Biotechnology and Bioengineering, Huaqiao University, Xiamen 361021, Fujian, PR China

ARTICLE INFO

Article history:
Received 12 April 2013
Received in revised form 24 April 2013
Accepted 3 May 2013

Keywords:
Halophile
Pearson VII function kernel
Support vector machine
Amino acid composition
Hypersaline adaptation

ABSTRACT

Understanding of proteins adaptive to hypersaline environment and identifying them is a challenging task and would help to design stable proteins. Here, we have systematically analyzed the normalized amino acid compositions of 2121 halophilic and 2400 non-halophilic proteins. The results showed that halophilic protein contained more Asp at the expense of Lys, Ile, Cys and Met, fewer small and hydrophobic residues, and showed a large excess of acidic over basic amino acids. Then, we introduce a support vector machine method to discriminate the halophilic and non-halophilic proteins, by using a novel Pearson VII universal function based kernel. In the three validation check methods, it achieved an overall accuracy of 97.7%, 91.7% and 86.9% and outperformed other machine learning algorithms. We also address the influence of protein size on prediction accuracy and found the worse performance for small size proteins might be some significant residues (Cys and Lys) were missing in the proteins.

1. Introduction

Halophilic microorganisms, which can survive in media with high salt concentrations, have generated many scientific interests. As proteins (enzymes) from the halophilic microorganisms have the ability to adapt to the extreme conditions of some industrial processes, such as high salt concentrations, and wide range of pH, thus offering important biotechnological potentials (Delgado-García et al., 2012). Designing proteins with improved halo-stability has been a main focus of protein engineering because of its theoretical and practical significance. So identifying the principles that rule protein halo-stability is of great interest both in basic research and industrial applications. To identify the features in proteins of halophilic organisms, many previous studies have been performed and revealed many important factors such as amino acid composition (Satoshi et al., 2003), dipeptide composition (Ebrahimie et al., 2011), lower propensities for helix formation (Sandip et al., 2008), highly negatively charged surfaces and weak hydrophobic cores (Kastritis et al., 2007) that contributed to the halo-stability of proteins.

However, there have been few parallel progresses about theoretical predictions for halo-stabilization (Ebrahimie et al., 2011), while many previous methods used sequence or structure-dependent information to predict protein thermostability. For example, Mozo-Villarías used a simple electrostatic criterion named the quasi-electric dipole profile to predict the thermal stability o proteins (Mozo-Villarías et al., 2003). Huang and Gromiha presented a weighted decision table method to predict the stability changes of 180 mutants obtained from thermal denaturation, the prediction accuracy was 82.2% for the 10-fold cross-validation (Huang and Gromiha, 2009). And recently, researchers develop PROTS, a sequential and spatial fragment based potential, for classifying thermophilic proteins/mesophilic proteins and stability changes upon mutations. The approach exhibits good performances in both classification and regression (Li et al., 2012). Some of these methods were successfully applied to design of thermo-stable mutants of several proteins (Dalluge et al., 2007; Bae et al., 2008). Thus, to design resistant proteins under high-salt concentrations, effective and robust computational algorithms for designing halo-stable proteins are in critical demand.

In the last few years, applying support vector machines (SVMs) for solving biological classification and regression problems has grown substantially duo to its attractive modeling features, promising generalization performances and robustness. SVMs are becoming established as a standard tool in bioinformatics (Ward et al., 2003; Kandaswamy et al., 2011). One of the main reasons for

Abbreviations: Weka, Waikato environment for knowledge analysis; RBF neural network, radial basis function neural network; SVMs, support vector machine; PUFK, Pearson VII universal function kernel; SE, sensitivity; SP, specificity; ACC, accuracy; MCC, Matthew's correlation coefficient; ROC, receiver operating characteristic; TP, true positives; FN, false negatives; TN, true negatives; FP, false positives.

* Corresponding author. Tel.: +86 592 616 2300.
E-mail address: zhgyghh@hqu.edu.cn (G. Zhang).

1476-9271
http://dx.doi.org/10.1016/j.compbiolchem.2013.05.001

the popularity of SVMs is its ability to model complex non-linear relationships by selecting a suitable kernel function. Some popular kernels are the linear kernel, polynomial kernel and radial basis function (RBF) kernel. The present study was initiated in an attempt to introduce a new kernel, the so-called "Pearson VII universal function kernel (PUFK)" (Uestuen et al., 2006), to discriminate halophilic and non-halophilic proteins based only on protein primary structure information.

2. Materials and methods

2.1. Datasets construction

To get high-quality and unbiased dataset, the data were strictly screened according to the following procedures. (1). The extremely halophilic archaeon *Halorhabdus tiamatea* (Antunes et al., 2011) and two non-halophilic archaeon *Methanococcus maripaludis* (Hendrickson et al., 2004) and *Cenarchaeum symbiosum* (Hallam et al., 2006) were chosen, the proteomic sequences were from UniProt. (2). Sequences which have fewer than 100 residues were removed because they might be partial or just be fragments. (3). Sequences which contain three or more consecutive uncertain amino acids (i.e. "XXX," "XXXX," and so on) were also removed. (4). To avoid any homologous bias, a redundancy cutoff was imposed by Blastclust to exclude those sequence that have ≥25% sequence identity to any other in the same subset according to Chou's work (Chou and Shen, 2008). Finally, we got 2121 halophilic proteins and 2400 non-halophilic proteins.

2.2. Normalized amino acid composition

In previous studies (Satoshi et al., 2003; Sandip et al., 2008), we found the average amino acid composition of all sequences was not considered in comparing the difference of amino acid composition between halophilic and non-halophilic proteins. In Uniprot database, average amino acid composition in percent for the complete database is listed. Some amino acids such as Cys and Trp have a small composition in protein sequences, while some amino acids such as Leu and Ala have a high composition. So, when analyzing the influence of amino acid composition, the result would be better if considering the average amino acid composition of all related proteins (Ding et al., 2004).

To achieve the goal, we calculated the normalized amino acid composition (*Nacc*) of each halophilic and non-halophilic protein with Eqs. (1) and (2):

$$Nacc_H^i = \frac{Comp_H^i - \overline{Comp^i}}{\overline{Comp^i}} \tag{1}$$

$$Nacc_N^i = \frac{Comp_N^i - \overline{Comp^i}}{\overline{Comp^i}} \tag{2}$$

where $Nacc_H^i$ and $Nacc_N^i$ are normalized composition of amino acid i for halophilic and non-halophilic proteins, $Comp_H^i$ and $Comp_N^i$ are the composition of amino acid i for halophilic and non-halophilic proteins, $\overline{Comp^i}$ is the average composition of amino acid i for all proteins in Uniprot. The differences of the amino acid composition between halophilic and non-halophilic proteins were calculated according to Eq. (3).

$$D_{H-N}^i = Nacc_H^i - Nacc_N^i = \frac{Comp_H^i - Comp_N^i}{\overline{Comp^i}} \tag{3}$$

The overall amino acids for halophilic and non-halophilic proteins are 671 230 and 717 749, respectively.

2.3. Pearson VII universal function kernel (PUFK)

The general form of the Pearson VII function for curve fitting purpose is give by

$$f(x) = \frac{H}{[1 + ((2(x - x_0)\sqrt{2^{(1/\omega)} - 1})/\delta)^2]^{\omega}} \tag{4}$$

where H is the peak height at the center x_0 of the peak, and x represents the independent variable. The parameters σ and ω control the half-width and the tailing factor of the peak. However, a function belongs to the class of valid kernel functions if and only if its corresponding kernel matrix is symmetric and positive semi-definite. To show that the PUFK indeed satisfies these conditions, Uestuen rewritten Eq. (1) into a function of two vectors (Uestuen et al., 2006):

$$K(x_i, x_j) = \frac{1}{[1 + ((2\sqrt{|x_i - x_j|^2\sqrt{2^{(1/\omega)} - 1}})/\sigma)]^{\omega}} \tag{5}$$

where x_i and x_j are two vector arguments. The peak off-set term x_0 in Eq. (1) is removed and the peak height H is simply replaced by 1, this without loss of generality. In this way, the Pearson VII function kernel will lead to a symmetric matrix with ones on the diagonal and all other entries ranging between the values 0 and 1 for any arbitrary pair (x_i and x_j). The PUFK is robust and has an equal or even stronger mapping power as compared to the standard kernel functions, which lead to an equal or better generalization performance of SVMs.

The algorithms implementations were achieved using the Weka package (Inamdar et al., 2004); all the running parameters of the classifiers were set as the defaults.

2.4. Validation check methods

The performance and robustness of the model was evaluated by three different validation check approaches, as shown below.

Firstly, we have used the datasets of 2121 halophilic proteins (HPs) and 2400 non-halophilic proteins (NPs) to train the model, these same proteins have been used to predict whether each protein is halophilic or non-halophilic. This method is called back-check prediction (or self-consistency test).

Secondly, we adopted the jackknife test (leave-one-out), which is deemed the most rigorous and objective with the least arbitrariness that can always yield a unique result for a given benchmark dataset as discussed by many investigators (Chou, 2011; Chou and Shen, 2008; Mohabatkar, 2010; Hayat and Khan, 2011; Jahandideh et al., 2012; Chou, 2001; Kandaswamy et al., 2010; Chen and Li, 2013; Sahu and Panda, 2010) and a review (Chou and Zhang, 1995). However, to reduce the computational time, we choose the 10-fold cross-validation to test the accuracy of our method. It was carried out by taking the total available set of the training datasets and partitioning it into 10 approximately equal-sized sets (212 halophilic proteins and 240 no-halophilic proteins). The protein sequences in each partition were randomly selected. Then the jackknife test was used. Nine partitions were used to train the model and then tested with the remaining partition. This was repeated 10 times, leaving in turn a different partition of the data out of the training set and using it to validate the resulting models.

Finally, to provide a more precise assessment of the reliability and the generalization capacity of the method, we carried out an independent test. The testing datasets contained completely new halophilic and non-halophilic protein. There were 2350 HPs and 1565 NPs, which came from an extreme halophile *Salinibacter ruber* DSM 13855 and a non-halophile *Pelodictyo luteolum* DSM 2379

(Sandip et al., 2008). These proteins have less than 25% identity with the training dataset sequences.

2.5. Evaluation of the performance

The final performance of the method was determined by measuring the sensitivity (SE), specificity (SP), accuracy (ACC), Matthew's correlation coefficient (MCC) and the receiver operating characteristic (ROC) score. The ROC score is the area under the ROC curve (AUC) and were calculated automatically by the *Weka* software. The SE, SP, ACC and MCC parameters were calculated using Eqs. (6)–(9), respectively.

$$SE = \frac{TP}{TP + FN} \tag{6}$$

$$SP = \frac{TN}{TN + FP} \tag{7}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}} \tag{9}$$

where *TP* are true positives (HPs predicted as halophilic); *FN* are false negatives (HPs predicted as non-halophilic); *TN* are true negatives (NPs predicted as non-halophilic) and *FP* are false positives (NPs predicted as halophilic). To make the equations more intuitive, please see Eqs. (9)–(13) in (Xu et al., 2013) and Eqs. (10)–(14) in (Chen et al., 2013), which should be better and clearer although the final results would be the same.

3. Results and discussion

3.1. Differences of amino acid composition in HPs and NPs

Fig. 1A shows the difference of individual amino acid between HPs and NPs. From it, we can see there are marked, significant amino acid composition differences in the HPs and NPs. Here, if $|D^i_{H-N}| > 0.5$, we regard *i* as the significant amino acids. Based on it, the significant amino acids were Asp (D), Lys (K), Ile (I), Cys (C) and Met (M), while HPs contained more Asp at the expense of Lys, Ile, Cys and Met. In general, halophilic proteins contained an excess of negatively charged amino acids (Asp) over positively charged residues (Lys) (Tokunaga et al., 2008), our result confirms the previous reports. Cys residues are usually overrepresented in non-flexible regions due to the formation of rigid disulfide-bridges. Avoidance of Cys in halophilic proteins might give them more flexibility in high salt environment (Sandip et al., 2008). Ile and Met are significantly underrepresented in halophilic proteins; this was mainly because they are large hydrophobic amino acids and halophilic proteins have weak hydrophobic cores (Kastritis et al., 2007). Further, Ile and Met are strong β-pleated sheet formers (Chou and Fasman, 1974); however, halophilic proteins have lower propensities for the formation of sheet in their secondary structure and adopted a narrower β-pleated sheet (Kastritis et al., 2007).

To get more information about the composition differences of HPs and NPs, we classified the amino acids into 13 groups according to the website http://www.russelllab.org/aas/ (Betts and Russell, 2003) (for simplicity, we used the one letter code of amino acid.). The 13 groups include the charged (DEKHR), aliphatic (ILV), aromatic (FWY), polar (DERKQN), neutral (AGHPSTY), hydrophobic (CVLIMFW), positively charged (HKR), negatively charged (DE), tiny (ACDGST), small (EHILKMNPQV); large (FRWY), sulfur (CM) and amide (NQ) residue (Betts and Russell, 2003) and the acidic amino acids minus basic (D + N + E + Q − K − R). We also calculated

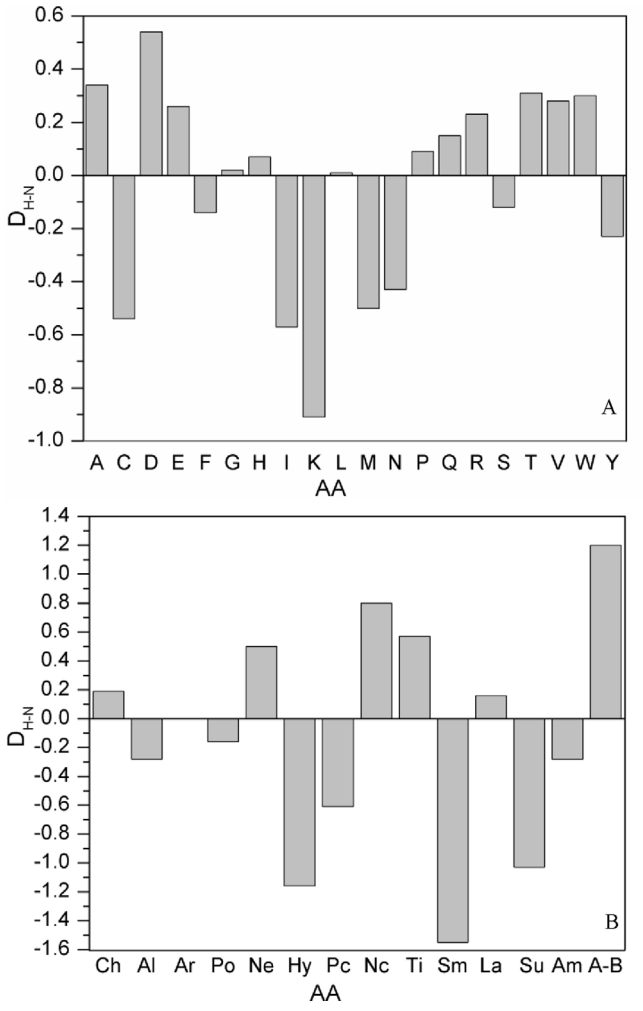

Fig. 1. Differences of amino acid composition between HPs and NPs. (A) Amino acids and (B) 14 kinds of amino acids. D_{H-N}: the differences of the normalized amino acid composition between halophilic and non-halophilic proteins; Ch: charged (DEKHR); Al: aliphatic (ILV); Ar: aromatic (FHWY); Po: polar (DERKQN); Ne: neutral (AGHPSTY); Hy: hydrophobic (CVLIMFW); Pc: positively charged (HKR); Nc: negatively charged (DE); Ti: tiny (ACDGST); Sm: small (EHILKMNPQV); La: large (FRWY), Su: sulfur (CM) and Am: amide (NQ) residue; A-B: acidic minus basic (D + N + E + Q − K − R). In (A) and (B), the upper half shows the dominance of amino acids in HPs and the negative values indicate higher occurrence of amino acids in NPs than in HPs.

the differences of the 13 groups' residues and showed the results in Fig. 1B. As shown in Fig. 1B, halophilic proteins contained more negative charge residues, while the apparent excess of small, hydrophobic and sulfur residues in the non-halophilic proteins. Some earlier works have revealed the higher content of negative charged residues (often on protein surface) as one of the most prominent features of halophilic protein (Mevarech et al., 2000; Kennedy et al., 2001; Madern et al., 2000). Such halophilic proteins have a strong negative charge at the physiological pH, which the proteins may be adapted to function in a high-salt environment (Mongodin et al., 2005). Earlier works also showed that relatively low hydrophobicity was another adaptation to hypersaline condition (Kastritis et al., 2007), which is consisted with our result. Although our results shared a common tendency to some amino acids in halophilic proteins, some differences exist among the cases. For example, small residues are found significantly higher in non-halophilic proteins, however, a statistical analysis of 26 soluble halophilic proteins showed an increase in small residues (Madern et al., 1995). This might be non-halophilic protein contain more K, I, M and N (Asn), while the first three are the significant residues

Table 1
Performances of different algorithms.

Methods	Algorithms	SE	SP	ACC	MCC	ROC
Self-consistency check	SVM (PUFK)	97.5	97.9	97.7	0.950	0.977
	SVM (poly kernel, $E=2$)	95.3	96.0	95.7	0.913	0.957
	SVM (linear kernel)	93.8	94.9	94.4	0.887	0.943
	SVM (RBF kernel)	93.1	92.0	92.5	0.850	0.925
	RBF NN	93.7	95.2	94.5	0.889	0.986
	Logitboost	92.8	92.3	92.5	0.850	0.975
	Adboost	92.6	92.0	92.3	0.845	0.974
	Bayesnet	93.0	90.2	91.5	0.830	0.977
	Random forest	100.0	99.8	99.9	0.998	1.000
	J4.8	98.8	98.8	98.8	0.976	0.996
	Naïve Bayes	91.8	87.1	89.3	0.790	0.966
	Decision stump	90.1	76.5	82.9	0.667	0.822
10-Fold cross validation	SVM (PUFK)	95.3	96.3	95.8	0.917	0.958
	SVM (poly kernel, $E=2$)	94.9	95.7	95.3	0.906	0.953
	SVM (linear kernel)	93.5	94.6	94.1	0.880	0.941
	SVM (RBF kernel)	93.3	91.6	92.4	0.850	0.924
	RBF NN	93.1	94.5	93.8	0.880	0.983
	Logitboost	91.4	91.2	91.3	0.825	0.971
	Adboost	90.2	91.7	91.0	0.819	0.966
	Bayesnet	92.6	90.1	91.3	0.825	0.973
	Random forest	94.9	93.4	94.1	0.882	0.924
	J4.8	90.6	91.0	90.8	0.820	0.901
	Naïve Bayes	92.0	87.1	89.4	0.789	0.966
	Decision stump	90.1	76.5	82.9	0.667	0.822
Independent test	SVM (PUFK)	86.5	87.5	86.9	0.731	0.870
	SVM (poly kernel, $E=2$)	86.0	84.9	85.6	0.703	0.855
	SVM (linear kernel)	86.7	81.7	84.7	0.681	0.842
	SVM (RBF kernel)	90.5	78.5	85.7	0.699	0.845
	RBF NN	86.1	82.7	84.8	0.685	0.921
	Logitboost	82.4	83.1	82.7	0.647	0.908
	Adboost	77.5	87.2	81.4	0.634	0.894
	Bayesnet	90.6	77.2	85.2	0.689	0.931
	Random forest	83.7	85.6	84.5	0.684	0.920
	J4.8	70.8	81.2	74.9	0.509	0.737
	Naïve Bayes	91.2	65.8	81.1	0.601	0.908
	Decision stump	89.0	53.0	74.6	0.459	0.710

SE: sensitivity; SP: specificity; ACC: accuracy; MCC: Matthew's correlation coefficient; ROC: area under the receiver operating characteristics curve.

in non-halophilic proteins as mentioned above. We also calculated the acidic residues minus the basic residues, and the results was in accord with most earlier works that halophilic protein have a large excess of acidic residues over basic residues. However, Bardavid and Oren (2012) recently found that proteins from members of the Halanaerobiales, which are active in the presence of high intracellular KCl concentrations, did not have the typical acidic signature of the halophilic proteins.

In general, it should be noted that most of the trends (except more small residues in non-halophilic proteins) in our work have already been noticed and discussed in an excellent way by previous researchers (Kastritis et al., 2007; Sandip et al., 2008). However, they are much more pronounced in this study, which has the advantage to include many more sequences and refined calculate of normalized amino acid composition.

3.2. Performances of the SVMs with PUK

When using the amino acid composition as the attribute, we investigated the performance of SVMs with PUFK and showed the results in Table 1. During the self-consistency check, the so-called training dataset was also used as the testing one. It was observed that by a few iterations SVMs with PUFK had already achieved the 97.7% overall accuracy. The high success rate also suggests that SVMs with PUFK, after undergoing an efficient training, has grasped the complicated relationship between amino acid composition and the halo-stability. It did not achieve the 100% overall accuracy also indicates there may still exist some noisy sequences in the training dataset even if we have adopted some method to avoid them. The self-consistency check is essential because a predictor with a poor self-consistency cannot be deemed as a good one (Table 1).

For the 10-fold cross-validation tests by Jackknifing, it has correctly identified 2022 out of 2121 halophilic proteins and 2311 out of 2400 non-halophilic proteins, yielding an accuracy of 95.3% for halophilic proteins and 96.3% for non-halophilic, respectively. The overall accuracy was 95.8%, which was 1.9% lower than that of self-consistency check. The SE value is 95.3%, suggesting the SVMs with PUFK can recognize about 95.3% of the halophilic proteins, while the SP value was 96.3%, suggesting the models can recognize about 96.3% of the non-halophilic proteins. The area under the ROC curve was 0.958 (larger than 0.9), suggesting the prediction was excellent. Our result shows individual amino acid composition in the protein sequences with no structural information can achieve this degree of accuracy in classification. The MCC value of 10-fold cross-validation was 0.917, that is, 0.033 worse than that got in the

Table 2
Prediction performances of different sequence length.

	Sequence length (L)			
	$100 \le L < 200$	$200 \le L < 500$	$500 \le L < 800$	$L \ge 800$
SE	82.3	86	90.5	97.2
SP	84.5	88	90.5	90.9
ACC	83.3	86.8	90.5	95.5
MCC	0.664	0.732	0.797	0.887

SE: sensitivity; SP: specificity; ACC: accuracy; MCC: Matthew's correlation coefficient.

Table 3
Distribution of sequences that did not contained the listed amino acids.

	Sequence length (%)				Protein type (%)		Total number
	$100 \le L < 200$	$200 \le L < 500$	$500 \le L < 800$	$L \ge 800$	HP	NP	
Cys	43.50	44.19	8.48	3.83	68.50	31.50	731
Trp	55.07	40.31	4.41	0.22	52.90	47.10	454
His	73.39	24.77	1.83	0.00	55.00	45.00	109
Lys	54.55	43.18	2.27	0.00	95.50	4.50	88
Asn	79.69	20.31	0.00	0.00	76.60	23.40	64

self-consistency check. This could be explained by the larger size of the training datasets during self-consistency check compared to 10-fold cross-validation.

Further, as a demonstration of practical application, we conducted the independent test for the SVMs with FUFK as the same parameters mentioned above. For the testing dataset, our method successfully identified 2032 out of 2350 halophilic proteins and 1369 out of 1565 non-halophilic proteins, and yielded an accuracy of 86.5% and 87.5%, respectively. The overall prediction accuracy was about 86.9%, that is, 8.9% worse than the performance obtained during 10-fold cross-validation. The MCC value was 0.713, which was 0.186 worse than that of 10-fold cross-validation. The area under the ROC curve was 0.870, meaning the prediction was good.

As we know, the more classes under the dataset coverage, the more difficult to get a high success rate. We addressed in our case is a 2-class problem. Accordingly, compared with the 20-class problem (Lin et al., 2013) and the 22-class problem (Chou et al., 2011), the 2-class problem is relatively easier to get such a high success rate.

In an earlier research (Ebrahimie et al., 2011) with a dataset that contained 258 HPs and 16 NPs, they analyzed the performances of different screening, clustering and decision tree algorithms for discriminating halophilic and non-halophilic proteins for the first time. The prediction accuracy was range from 94 to 100%, which was better than our results. However, they used more than 850 protein attributes, which was much more than our research (only 20 attributes). Besides, our dataset contained more sequences and we used more validation methods to check our classifiers.

3.3. Compared with other machine learning approaches

In this section, we compared the performances of SVMs with PUFK to other classifiers. The classifiers include the SVMs with the polynomial, linear and RBF kernels, RBF neural network, naïve Bayes, Bayes network, Random forest, J48 trees, decision stump, Logitboost and Adaboost algorithms. The performances of the classifiers are also shown in Table 1. In the self-consistency check, the accuracy of SVMs with PUFK module was about 14.8%, 8.4% and 6.2% higher than that of decision stump, naïve Bayes, Bayes network, respectively. Meanwhile, it also outperformed SVMs with other kernels in terms of overall accuracy by more than 2.0%, 3.3% and 5.2%, respectively. But the overall accuracy was 1.1% and 2.2% less than that of J4.8 tree and random forest. In the 10-fold cross-validation check, the SVMs with PUFK outperformed all the other machine learning methods in terms of overall accuracies ranging from 0.3% to 12.9%. In the independent testing, the performance of SVMs with PUFK in terms of overall accuracy was also the best. The accuracy of SVMs with PUFK was able to recognize 87.5% of the non-halophilic proteins, that is, nearly 34.5% higher than decision stump. The naïve Bayes could identify about 91.2% halophilic proteins (4.7% better than SVMs with PUFK) but just recognized the non-halophilic proteins with an accuracy of 65.8%. Interestingly, SVMs with RBF kernel could predict halophilic proteins with an accuracy of 90.5% (4.0% better than SVMs with PUFK) but moderately predicted the non-halophilic proteins with an accuracy of 78.5% (9.0% less than SVMs with PUFK). During the three validation check methods, the performances of SVMs with polynomial kernel ($E = 2$) were comparable with SVMs with PUFK, which were only 2.0%, 0.5% and 1.3% less, respectively. However, the SVMs with PUKF were faster during the data processing, it can save about 10% of the computation time.

As the parameters have a great impact on the performance of the classifiers, we used a uniform design method to optimize the SVMs. For the SVMs with linear, polynomial and PUKF, a slight improvement (about 0.1%) was achieved. For the SVMs with RBF kernel, the best performances reached an accuracy of 94.5%, with an improvement of 2.1% than the default parameters. This means the SVMs with RBF kernel was more sensitive to the running parameters. However, even if the parameters were optimized, the performance of SVMs with PUKF was also the best, with an over accuracy of 95.9%. Thus, in our hand, SVMs with PUFK outperformed other machine learning and statistical techniques for discriminating of halophilic and non-halophilic proteins.

3.4. Influence of protein size on discrimination

When using amino acid composition as the attributes, the prediction accuracy varied among sequences with different length. For example, when using sequences information to discriminate globular and outer membrane proteins (Gromiha, 2005), the prediction accuracy for proteins with 300 residues or less was 86%, while it achieved an accuracy of 100% for large-size proteins (more than 800 residues) in outer membrane proteins. In another study, the prediction accuracy for small protein (less than 200 residues) was only 79.0%, and for large proteins (with or more than 800 residues) 100% (Zhang and Fang, 2007). However, some enough explanations were not provided in their researches.

To study the influence of protein size on the predictive performance, we have further analyzed the prediction results based on the independent testing validation and showed the results in Table 2. There were 2350 HPs and 1565 NPs in the testing dataset and their sequences length distribution was shown in Fig. 2. As can be seen the average of the sequence length was 357.7, and most of the sequences length (about 81.2%) were between 100 and 500. We divided the testing 3915 sequences into four groups (Table 2) and they were 24.7, 56.5, 13.0 and 5.8% of the total sequences, respectively. For the large-size proteins (with or more than 800 residues), the method achieved an overall accuracy of 95.5%. For the proteins with residues between 500 and 800 residues, the method correctly picked up 484 out of the 535 proteins with an accuracy of 90.5%. For the proteins with residues between 200 and 500 residues, the method successfully identified 1920 out of the 2212 proteins and achieved overall accuracy of 86.8%. However, the accuracy for the small size proteins (fewer than 200 residues) was 3.6% less than the overall accuracy; it only correctly picked up 806 out of 968 proteins with an accuracy of about 83.3%.

From the average amino acid composition in Uniprot database, we found some amino acids such as Cys, Trp, and His are significantly lower than others. Thus, we checked these least residues in the 3915 testing sequences, analyzed the distribution of the

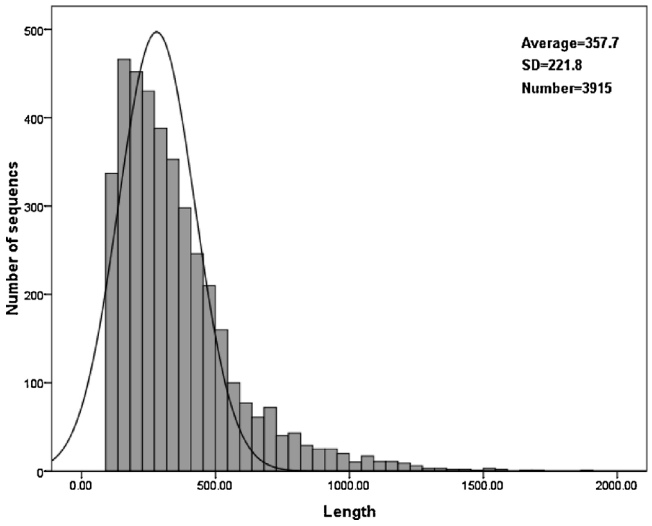

Fig. 2. Sequence length distribution of the testing dataset.

sequences that did not contain these residues and listed the result in Table 3. As clearly shown in Table 3, there are 731 sequences do not contain Cys, 454 sequences do not contain Trp and 109 sequences do not contain His. As mentioned above, small proteins are 24.7% of the total in the testing dataset, but many of the sequences (missing the Cys, Trp and His) are small proteins (from 43.5% to 73.39%). From this viewpoint, most of the proteins that did not contain the three residues are small (fewer than 200 residues). Interestingly, the composition of Lys and Asn in the proteins in Uniprot is moderate, which is 4.07% and 5.54%, respectively. However, the proteins that did not contain Lys and Asn are more than the left 15 kinds of residues. Similarly, most of them are small proteins; their percentages were 54.55 and 79.69, respectively. Among the five residues, Cys and Lys are among the significant amino acid as mentioned in Section 3.1. Thus, the small proteins, which have less information contents and missing the significant residues might be more difficult for the classifier to correctly discriminate them. This can be proved by the fact the overall prediction accuracy was slightly improved and reached 87.2% when removed the small proteins that did not contain Cys and Lys. As for the protein types (halophilic and non-halophilic), most of the proteins did not contain the five residues are halophilic proteins, their percentages ranging from 55 to 95.5. This might explain the overall prediction accuracy for HPs are 1% less than that of NPs.

As we know, there are many structural features that halostable proteins have, these features include amino acid composition (Satoshi et al., 2003), lower propensities for helix formation (Sandip et al., 2008), electrostatic interactions (Elcock and McCammon, 1998) and weak hydrophobic cores (Kastritis et al., 2007). Amino acid composition is just the one key factor. Based on this point, we consider it is reasonable to believe the algorithm based on SVMs with PUFK still has potential to improve, especially for the small-size proteins.

4. Conclusion

Support vector machines (SVMs), which have many desirable properties, have shown promising results on several biological pattern classification problems and have been a standard tool in bioinformatics (Ward et al., 2003; Kandaswamy et al., 2011). This can only be realized if a suitable kernel function is applied. Since the nature of the bio-data is usually unknown, it is very difficult to make, on beforehand, a proper choice out of the three commonly used kernels mentioned above. Therefore, one has to select the kernel which gives the best performances during the model building process; this will lead to a very time-consuming optimization procedure. The PUFK is introduced to circumvent this disadvantage (Uestuen et al., 2006). From the result of the three validation-check methods, it was conclude the PUFK is robust and has an equal or even stronger mapping power as compared to the standard kernel functions. This leads to an equal or better performance of SVMs. It is anticipated the power in discriminating halophilic and non-halophilic proteins as well as many other bio-macromolecular attributes will be further strengthened if the SVMs with PUFK and some other existing algorithms can be effectively complemented with each other.

Perhaps, in the future study, we will deal with more and proper attribute sets, such as pseudo-amino acid composition proposed by Chou (2009) with the datasets of higher quality to improve the performance of SVMs with PUFK. It should be possible from extracting more primary structure features and updated databases.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or preditors (Chou and Shen, 2009), we shall make efforts in our future work to provide a web-server for the method presented in this paper.

Acknowledgements

This work was supported by the Cultivation Project of Huaqiao University for the China National Funds for Distinguished Young Scientists (No. JB-GJ1006) and the Program for New Century Excellent Talents in Universities of Fujian Province (No. 07176C02).

References

Antunes, A., Alam, I., Bajic, V.B., et al., 2011. Genome sequence of *Halorhabdus tiamatea*, the first archaeon isolated from a deep-sea anoxic brine lake. Journal of Bacteriology 193, 4553–4554.

Bae, E., Bannen, R.M., Phillips, G.N.J., 2008. Bioinformatic method for protein thermal stabilization by structural entropy optimization. Proceedings of the National Academy of Sciences of the United States of America 105, 9594–9597.

Bardavid, R.E., Oren, A., 2012. The amino acid composition of proteins from anaerobic halophilic bacteria of the order Halanaerobiales. Extremophiles 16, 567–572.

Betts, M.J., Russell, R.B., 2003. Amino acid properties and consequences of substitutions. In: Bioinformatics for Geneticists. John Wiley & Sons, Ltd., pp. 289–316.

Chen, W., Feng, P.M., Lin, H., et al., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Research 41, e68.

Chen, Y.K., Li, K.B., 2013. Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. Journal of Theoretical Biology 318, 1–12.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Structure, Function, and Bioinformatics 43, 246–255.

Chou, K.C., Shen, H.B., 2009. Review: recent advances in developing web-servers for predicting protein attributes. Natural Science 2, 63–92.

Chou, K.C., Wu, Z.C., Xiao, X., 2011. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS One 6, e18258.

Chou, K.C., Zhang, C.T., 1995. Review: prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology 30, 275–349.

Chou, K.C., 2009. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Current Proteomics 6, 262–274.

Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). Journal of Theoretical Biology 273, 236–247.

Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nature Protocols 3, 153–162.

Chou, P.Y., Fasman, G.D., 1974. Conformational parameters for amino acids in helical, beta-sheet and random coil regions calculated from proteins. Biochemistry 13, 211–222.

Dalluge, R., Oschmann, J., Birkenmeier, O., et al., 2007. A tetrapeptide fragment-based design method results in highly stable artificial proteins. Proteins: Structure, Function, and Bioinformatics 68, 839–849.

Delgado-García, M., Valdivia-Urdiales, B., Aguilar-González, C.N., et al., 2012. Halophilic hydrolases as a new tool for the biotechnological industries. Journal of the Science of Food and Agriculture 92, 2575–2580.

Ding, Y.R., Cai, Y.J., Zhang, G.X., et al., 2004. The influence of dipeptide composition on protein thermostability. FEBS Letters 569, 284–288.

Ebrahimie, E., Ebrahimi, M., Rahpayma, N.S., et al., 2011. Protein attributes contribute to halo-stability, bioinformatics approach. Saline Systems 7, 1.
Elcock, A.H., McCammon, J.A., 1998. Electrostatic contributions to the stability of halophilic proteins. Journal of Molecular Biology 280, 731–748.
Gromiha, M.M., 2005. Motifs in outer membrane protein sequences: applications for discrimination. Biophysical Chemistry 117, 65–71.
Hallam, S.J., Konstantinidis, K.T., Putnam, N., et al., 2006. Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. Proceedings of the National Academy of Sciences of the United States of America 103, 18296–18301.
Hayat, M., Khan, A., 2011. Predicting membrane protein types by fusing composite protein sequence features into pseudo amino acid composition. Journal of Theoretical Biology 271, 10–17.
Hendrickson, E.L., Kaul, R., Zhou, Y., et al., 2004. Complete genome sequence of the genetically tractable hydrogenotrophic methanogen *Methanococcus maripaludis*. Journal of Bacteriology 186, 6956–6969.
Huang, L.T., Gromiha, M.M., 2009. Reliable prediction of protein thermostability change upon double mutation from amino acid sequence. Bioinformatics 25, 2181–2187.
Inamdar, N.M., Ehrlich, K.C., Ehrlich, M., et al., 2004. Data mining in bioinformatics using Weka. Bioinformatics 20, 2479–2481.
Jahandideh, S., Srinivasasainagendra, V., Zhi, D., 2012. Comprehensive comparative analysis and identification of RNA-binding protein domains: multi-class classification and feature selection. Journal of Theoretical Biology 312, 65–75.
Kandaswamy, K.K., Pugalenthi, G., Moller, S., et al., 2010. Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition. Protein and Peptide Letters 17, 1473–1479.
Kandaswamy, K.K., Pugalenthi, G., Hazrati, M.K., et al., 2011. BLProt: prediction of bioluminescent proteins based on support vector machine and relief feature selection. BMC Bioinformatics 12, 345.
Kastritis, P.L., Papandreou, N.C., Hamodrakas, S.J., 2007. Haloadaptation: insights from comparative modeling studies of halophilic archaeal DHFRs. International Journal of Biological Macromolecules 41, 447–453.
Kennedy, S.P., Ng, W.V., Salzberg, S.L., et al., 2001. Under-standing the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. Genome Research 11, 1641–1650.
Li, Y., Zhang, J., Tai, D., et al., 2012. PROTS: a fragment based protein thermo-stability potential. Proteins 80, 81–92.
Lin, W.Z., Fang, J.A., Xiao, X., et al., 2013. iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Molecular Biosystems 9, 634–644.
Madern, D., Ebel, C., Zaccai, G., 2000. Halophilic adaptation of enzymes. Extremophiles 4, 91–98.
Madern, D., Pfister, C., Zaccai, G., 1995. A single acidic amino acid mutation enhances the halophilic behaviour of malate dehydrogenase from *Haloarcula marismortui*. European Journal of Biochemistry 230, 1088–1095.
Mevarech, M., Frolow, F., Gloss, L.M., 2000. Halophilic enzymes: proteins with a grain of salt. Biophysical Chemistry 86, 155–164.
Mohabatkar, H., 2010. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein and Peptide Letters 17, 1207–1214.
Mongodin, E.F., Nelson, K.E., Daugherty, S., et al., 2005. The genome of *Salinibacter ruber*: convergence and gene exchange among hyperhalophilic bacteria and archaea. Proceedings of the National Academy of Sciences of the United States of America 102, 18147–18152.
Mozo-Villarías, A., Cedano, J., Querol, E., 2003. A simple electrostatic criterion for predicting the thermal stability of proteins. Protein Engineering 16, 279–286.
Sahu, S.S., Panda, G., 2010. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Computational Biology and Chemistry 34, 320–327.
Sandip, P., Sumit, K.B., Sabyasachi, D., et al., 2008. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. Genome Biology 9, R70.
Satoshi, F., Kazuaki, Y., Mamoru, W., et al., 2003. Unique amino acid composition of proteins in halophilic bacteria. Journal of Molecular Biology 327, 347–357.
Tokunaga, H., Arakawa, T., Tokunaga, M., 2008. Engineering of halophilic enzymes: two acidic amino acid residues at the carboxy-terminal region confer halophilic characteristics to *Halomonas* and *Pseudomonas* nucleoside diphosphate kinases. Protein Science 17, 1603–1610.
Uestuen, B., Melssen, W.J., Buydens, L.M.C., 2006. Facilitating the application of support vector regression by using a universal Pearson-function based kernel. Chemometrics and Intelligent Laboratory Systems 81, 29–40.
Ward, J.J., McGuffin, L.J., Buxtion, B.F., et al., 2003. Secondary structure prediction with support vector machines. Bioinformatics 19, 1650–1655.
Xu, Y., Ding, J., Wu, L.Y., et al., 2013. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS One 8, e55844.
Zhang, G.Y., Fang, B.S., 2007. LogitBoost classifier for discriminating thermophilic and mesophilic proteins. Journal of Biotechnology 127, 417–424.