Computational Biology and Chemistry 46 (2013) 16–22

Contents lists available at ScienceDirect

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/$ – see front matter © 2013 The Authors. Published by Elsevier Ltd. All rights reserved.
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