BIOINFORMATICS ORIGINAL PAPER
Vol. 30 no. 4 2014, pages 472–479
doi:10.1093/bioinformatics/btt709

Sequence analysis
Advance Access publication December 5, 2013

Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection

Bin Liu[1,2,3,4,*], Deyuan Zhang[5], Ruifeng Xu[1,2], Jinghao Xu[1], Xiaolong Wang[1,2], Qingcai Chen[1,2], Qiwen Dong[6] and Kuo-Chen Chou[4,7]

[1]School of Computer Science and Technology and [2]Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China, [3]Shanghai Key Laboratory of Intelligent Information Processing, Shanghai 200433, China, [4]Gordon Life Science Institute, Belmont, MA 02478, USA, [5]School of Computer, Shenyang Aerospace University, Shenyang, Liaoning, China, [6]School of Computer Science, Fudan University, Shanghai 200433, China and [7]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

Associate Editor: John Hancock

Downloaded from http://bioinformatics.oxfordjournals.org/ at King Abdulaziz University on April 3, 2016

ABSTRACT

Motivation: Owing to its importance in both basic research (such as molecular evolution and protein attribute prediction) and practical application (such as timely modeling the 3D structures of proteins targeted for drug development), protein remote homology detection has attracted a great deal of interest. It is intriguing to note that the profile-based approach is promising and holds high potential in this regard. To further improve protein remote homology detection, a key step is how to find an optimal means to extract the evolutionary information into the profiles.

Results: Here, we propose a novel approach, the so-called profile-based protein representation, to extract the evolutionary information via the frequency profiles. The latter can be calculated from the multiple sequence alignments generated by PSI-BLAST. Three top performing sequence-based kernels (SVM-Ngram, SVM-pairwise and SVM-LA) were combined with the profile-based protein representation. Various tests were conducted on a SCOP benchmark dataset that contains 54 families and 23 superfamilies. The results showed that the new approach is promising, and can obviously improve the performance of the three kernels. Furthermore, our approach can also provide useful insights for studying the features of proteins in various families. It has not escaped our notice that the current approach can be easily combined with the existing sequence-based methods so as to improve their performance as well.

Availability and implementation: For users' convenience, the source code of generating the profile-based proteins and the multiple kernel learning was also provided at
http://bioinformatics.hitsz.edu.cn/main/~binliu/remote/

Contact: bliu@insun.hit.edu.cn or bliu@gordonlifescience.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 8, 2013; revised on November 25, 2013; accepted on November 29, 2013

1 INTRODUCTION

By March 2013, 89 003 experimentally determined protein structures were deposited in the Protein Data Bank (Berman *et al.*, 2007). However, this number is only about one-sixth of 539 616, the number of protein sequences held in the UniProtKB/Swiss-Prot database (Wu *et al.*, 2006). To timely use such vast amount of structure-unknown protein sequences for basic research and drug development, it is highly desired to determine their 3D structures and functions by means of homology approaches (Chou, 2004). Unfortunately, protein remote homology detection is still a challenging problem in bioinformatics.

The early methods in dealing with this problem were based on the pairwise sequence comparison approaches, such as BLAST (Altschul *et al.*, 1990) and Smith–Waterman local alignment algorithm (Smith and Waterman, 1981). However, in many cases, this kind of sequence alignment method failed to detect remote homologies due to the low sequence similarities. Later methods to challenge this problem were based on the generative models to induce a probability distribution over the protein family, and then to generate the unknown proteins as new members of the family from the stochastic model. For example, the hidden Markov model (HMM) (Karplus *et al.*, 1998) can be trained iteratively in a semi-supervised manner by using both positively labeled and unlabeled samples of a particular family to generate the positive set (Qian and Goldstein, 2004).

Recently, the discriminative methods, such as support vector machine (SVM) (Vapnik, 1998), were used to address this problem by focusing on the differences between protein families. The key of the SVM methods is the kernel function by which to compute the inner product between two samples in the feature space. The most straightforward approach to generate the kernels was based on the features extracted from protein sequences. SVM-Ngram (Dong *et al.*, 2006), SVM-pairwise (Liao and Noble, 2003) and SVM-LA (Saigo *et al.*, 2004) were three of the most successful sequence-based kernels. SVM-Ngram (Dong *et al.*, 2006) was based on the feature space that contains all short subsequence of length *N*. In SVM-pairwise (Liao and Noble, 2003), a protein sequence was represented as a vector of pairwise similarities to all protein

[*]To whom correspondence should be addressed.

© The Author 2013. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

sequences in the training set, and then inner product between these vector-space representations was taken as the kernel. SVM-LA (Saigo *et al*., 2004) measured the similarity between a pair of proteins by taking all the optimal local alignment scores with gaps between all possible subsequences into account. Besides these kernels, several other sequence-based kernels were also proposed, such as Mismatch (Leslie *et al*., 2004) and SVM-BALSA (Webb-Robertson *et al*., 2005). The profile-based kernels could further improve the performance by using the evolutional information extracted from the profiles. For example, Top-*n*-grams (Liu *et al*., 2008) extracted the profile-based patterns by considering the most frequent elements in the profiles; profile kernel (Kuang *et al*., 2005) extracted the short substrings according to the profile-based ungapped alignment scores; some profile-based methods improved the predictive performance by developing more sensitive profiles. HHsearch method (Söding, 2005) was based on a novel profile using the HMM. In COMPASS (Sadreyev *et al*., 2009), numerical profiles were generated to construct optimal profile–profile alignments and to estimate the statistical significance of the corresponding alignment scores.

In the meantime, some other features and techniques have been applied to this field to further improve the predictive performance. For instance, the kernel combination methodology (VBKC) (Damoulas and Girolami, 2008) used a single multi-class kernel machine to combine various kernels based on different feature spaces; SVM-physicochemical distance transformation (PDT) (Liu *et al*., 2012) combined the amino acid physicochemical properties and the profile features via PDT to incorporate the local sequence-order information of the entire protein sequences. Also, based on the similarities between protein sequences and natural languages, the natural language processing techniques were applied to this field. It was shown that the performance of building-block-based methods could be improved by using the latent semantic analysis (LSA) (Dong *et al*., 2006). Moreover, P_{ROT}E_{MBED} (Melvin *et al*., 2011) detected protein remote homology by embedding protein sequences into a low-dimensional semantic space.

As we can see from the aforementioned introduction, most of the top-performing methods were developed based on the features extracted from profiles. This is consistent with the fact that a profile is much richer than an individual sequence in encoding information. Also, biology is a natural science with historic dimension. All biological species have developed beginning from a limited number of ancestral species. It is true for protein sequence as well (Chou, 2004). Their evolution involves changes of single residues, insertions and deletions of several residues, gene doubling and gene fusion (Chou, 1995). With these changes accumulated for a long period, many similarities between initial and resultant amino acid sequences are gradually eliminated, but the corresponding proteins may still share many common features, such as having basically the same biological function (Loewenstein *et al*., 2009), folding topology, subcellular location and other attributes (Chou, 2013).

Accordingly, the key to improve the performance of these methods is to find a suitable approach to extract the evolutionary information from the profiles. In view of this, the current study was initiated in an attempt to propose a profile-based protein representation by extracting the evolutionary information from the frequency profiles.

2 MATERIALS AND METHODS

As shown by a series of publications (Chen *et al*., 2013; Liu *et al*., 2009; Xiao *et al*., 2013; Xu *et al*., 2013) and summarized in a comprehensive review (Chou, 2011), to develop a useful statistical prediction method or model for a biological system, one needs to engage the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; and (v) provide the downloadable source code or a web-server for the prediction method. Below, let us describe how to deal these procedures.

2.1 SCOP benchmark

Suppose $\mathbb{S}$ is a remote homology system investigated in the current study that contains 4352 protein sequences, which were taken from (Liao and Noble, 2003) at http://noble.gs.washington.edu/proj/svm-pairwise/. These proteins were selected from SCOP version 1.53 and extracted from the Astral database (Brenner *et al*., 2000). None of the 4352 proteins has sequence pairwise similarity to any other with higher than 10^{-25} in the E-value [for more about the E-value and its implication in reducing homology bias and redundancy, see (Brenner *et al*., 2000)]. These proteins were also used by others (Dong *et al*., 2006; Lingner and Meinicke, 2006; Saigo *et al*., 2004) to study remote homology detection. The 4352 proteins in $\mathbb{S}$ can be classified into 853 superfamilies and 1356 families; i.e.

$$\mathbb{S} = \mathbb{S}_1^{\mathrm{F}} \mathrm{U} \mathbb{S}_2^{\mathrm{F}} \mathrm{U} \cdots \mathrm{U} \mathbb{S}_{853}^{\mathrm{F}} = \mathbb{S}_1^{\mathrm{f}} \mathrm{U} \mathbb{S}_2^{\mathrm{f}} \mathrm{U} \cdots \mathrm{U} \mathbb{S}_{1356}^{\mathrm{f}} \tag{1}$$

where $\mathbb{S}_i^{\mathrm{F}}\ (i = 1, 2, \ldots, 853)$ is the *i*th superfamily, $\mathbb{S}_k^{\mathrm{f}}\ (k = 1, 2, \ldots, 1356)$ is the *k*th family and the symbol $\cup$ represents the 'union' in the set theory. For readers' convenience, the codes of the 4352 proteins and their sequence as well as the attributes of their families and superfamilies are given in Supplementary Material S1.

Because some families and superfamilies in $\mathbb{S}$ do not contain significant number of protein sequences, and also because the negative dataset for each protein family can be any proteins except those belonging to its own superfamily, it is not so straightforward but a little more complicated and subtle for how to select protein samples from $\mathbb{S}$ to define the training and testing datasets. To provide a clear description, let us consider a different manner to address this. As demonstrated by many previous studies on a series of important biological topics, such as enzyme-catalyzed reactions (Zhou and Deng, 1984), inhibition of human immunodeficiency virus-1 reverse transcriptase (Althaus *et al*., 1993), drug metabolism systems (Chou, 2010) and applying wenxiang diagram or graph (Chou *et al*., 2011) to study protein–protein interactions (Zhou, 2011; Zhou and Huang, 2013), using graphical approaches to study complicated problems can provide an intuitive picture and useful insights for in-depth studying and analyzing various complicated relations in these systems (Lin and Lapointe, 2013). In view of this, let us also use graphic approach to describe the feature and relation of the families and superfamilies in $\mathbb{S}$, as shown in Figure 1, where the open circles denote the families or superfamilies that have significant number of protein sequences and the gray circles denote those that do not.

Of the 1356 families in $\mathbb{S}$ (cf. Equation 1), 54 contain significant number of proteins (see the third row of Fig. 1) and form a target family set $\mathbb{S}_{\mathrm{target}}^{\mathrm{f}}$; i.e.

$$\mathbb{S}_{\mathrm{target}}^{\mathrm{f}} = \mathbb{S}_1^{\mathrm{f}} \mathrm{U} \mathbb{S}_2^{\mathrm{f}} \mathrm{U} \mathbb{S}_3^{\mathrm{f}} \mathrm{U} \cdots \mathrm{U} \mathbb{S}_{54}^{\mathrm{f}} \tag{2}$$

Of the 853 superfamilies in $\mathbb{S}$, 23 contain at least one target family (see the open circles in the second row of Fig. 1) and form a target superfamily set $\mathbb{S}_{\mathrm{Target}}^{\mathrm{F}}$; i.e.

$$\mathbb{S}_{\mathrm{target}}^{\mathrm{F}} = \mathbb{S}_1^{\mathrm{F}} \mathrm{U} \mathbb{S}_2^{\mathrm{F}} \mathrm{U} \cdots \mathrm{U} \mathbb{S}_{23}^{\mathrm{F}} \tag{3}$$

Downloaded from http://bioinformatics.oxfordjournals.org/ at King Abdulaziz University on April 3, 2016

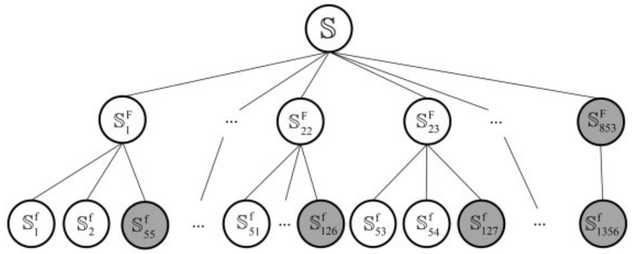

Fig. 1. A tree (or 3-row) graph to show the remote homology system on the SCOP benchmark. Only the open circles are in the target of the 23 superfamilies and 54 families, while the circles in gray are outside of the target. See the text for further explanation

Thus, we have

$$\mathbb{S}^{\mathrm{f}}_{\mathrm{target}} \subseteq \mathbb{S}^{\mathrm{F}}_{\mathrm{target}} \subseteq \mathbb{S} \tag{4}$$

meaning that $\mathbb{S}^{\mathrm{f}}_{\mathrm{target}}$ is the subset of $\mathbb{S}^{\mathrm{F}}_{\mathrm{Target}}$, and $\mathbb{S}^{\mathrm{F}}_{\mathrm{Target}}$ is the subset of $\mathbb{S}$; each of the three contains 857, 1508 and 4352 proteins, respectively.

Now, for each of the 54 families in the target family set $\mathbb{S}^{\mathrm{f}}_{\mathrm{target}}$, we can define a training dataset and testing dataset given by

$$\begin{cases} S_{\mathrm{train}}(k) = S^{+}_{\mathrm{train}}(k) \cup S^{-}_{\mathrm{train}}(k) \\ S_{\mathrm{test}}(k) = S^{+}_{\mathrm{test}}(k) \cup S^{-}_{\mathrm{test}}(k) \end{cases} (k = 1, 2, \ldots, 54) \tag{5}$$

where the positive training dataset $S^{+}_{\mathrm{train}}(k)$ contains at least 10 of its superfamily members, none of which belongs to the kth family, and the positive testing dataset $S^{+}_{\mathrm{test}}(k)$ contains at least five protein domains within the family. The proteins in the negative training and testing datasets, $S^{-}_{\mathrm{train}}(k)$ and $S^{-}_{\mathrm{test}}(k)$, were picked from $\mathbb{S}$ by excluding the superfamily of the kth family and randomly split between the two in the same ratio as the positive ones. The 54 training and testing datasets thus obtained are given in the Supplementary Materials S2 and S3, respectively.

2.2 Protein frequency profile

The frequency profile $\mathbb{M}$ for protein **P** with L amino acids can be represented by

$$\mathbb{M} = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,L} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,L} \\ \vdots & \vdots & \vdots & \vdots \\ m_{20,1} & m_{20,2} & \cdots & m_{20,L} \end{bmatrix} \tag{6}$$

where 20 is the total number of standard amino acids; $m_{i,j}$ $(0 \le m_{i,j} \le 1)$ is the target frequency, which reflects the probability of amino acid i $(i = 1,2,\ldots,20)$ occurring at the sequence position j $(j = 1,2,\ldots,L)$ in protein **P** during evolutionary processes. For each column in $\mathbb{M}$, the elements add up to 1.

The target frequency is calculated from the multiple sequence alignments generated by running PSI-BLAST (Altschul *et al.*, 1997) against the NCBI's NR with default parameters except that the number of iterations was not set at 1 but was set at 10 in the current study. The target frequency of amino acid i in sequence position j is calculated as:

$$m_{i,j} = \frac{(\alpha f_{ij} + \beta g_{ij})}{(\alpha + \beta)} \tag{7}$$

where f_{ij} is the observed frequency of amino acid i in column j; β is a free parameter set to a constant value of 10, which is initially used by PSI-BLAST; α is the number of different amino acids in column $j-1$; and g_{ij} is the pseudo-count for amino acid i in protein sequence position j, which can be calculated as:

$$g_{ij} = \sum_{k=1}^{20} \frac{f_{kj} q_{ik}}{p_k} \tag{8}$$

where p_k is the background frequency of amino acid k, and q_{ik} is the score of amino acid i being aligned to amino acid k in BLOSUM62 substitution matrix, which is the default score matrix of PSI-BLAST (Altschul *et al.*, 1997).

2.3 Profile-based protein representation

Although the methods by using amino acid sequence information achieve certain degree of success, only using sequence information cannot accurately detect protein remote homology. Recent studies demonstrated that the profile-based methods would show better performance because a profile is richer than an individual sequence as far as the encoding information is concerned. However, a profile is a 2D matrix, whereas a protein sequence is an amino acid string. Therefore, the 2D evolutionary profile information cannot be directly incorporated into the sequence-based methods for prediction. To deal with this problem, we propose an approach to convert the frequency profiles into a series of profile-based proteins. Thus, the existing sequence-based methods can be directly performed on these proteins for the prediction. The target frequencies in the frequency profiles reflect the probabilities of the corresponding amino acids appearing in the specific sequence positions. The higher the frequency is, the more likely the corresponding amino acid occurs. It is reasonable to use the nth most frequent amino acids in the frequency profiles to represent the protein sequences. Below is the description on how to convert frequency profiles into profile-based proteins.

Given the frequency profile $\mathbb{M}$ for protein **P** (Equation 6), we can sort the amino acids in each column according to a descending order. The frequency profile thus obtained by the sorting operation is called the sorted frequency profile and denoted by $\mathbb{M}'$. For each row in $\mathbb{M}'$, the amino acids are combined to produce the profile-based protein. By following this approach, the frequency profile $\mathbb{M}$ is converted into 20 profile-based proteins p_1, $\mathrm{p}_2, \ldots,$ p_{20} (Supplementary Fig. S1 in Supplementary Material S4), which contain the evolutionary information in the frequency profile. These 20 proteins have different importance. During evolutionary process, protein **P** is preferred to transform into p1, but not preferred to transform into p20. For reader's convenience, the source code for generating the profile-based proteins is accessible by clicking the link at http://bioinformatics.hitsz.edu.cn/main/∼binliu/remote/.

2.4 Sequence-based kernels

Three state-of-the-art sequence-based kernels [SVM-*N*gram (Dong *et al.*, 2006), SVM-pairwise (Liao and Noble, 2003) and SVM-LA (Saigo *et al.*, 2004)] were used to validate whether the proposed approach could improve their performance.

In SVM-Ngram (Dong *et al.*, 2006) method, the Ngrams were the set of all possible subsequences of amino acids of a fixed length. A protein sequence was mapped to a feature vector by the occurrence frequency of each Ngram. The value of N was set at 3 as suggested by the authors (Dong *et al.*, 2006), and therefore the dimension of the vector is 8000. At the heart of the SVM is a kernel function that acts as a similarity score between pairs of vectors. The kernel was normalized so that each vector had length 1 in the feature space:

$$K(X, Y) = \frac{X \cdot Y}{\sqrt{(X \cdot X)(Y \cdot Y)}} \tag{9}$$

where X and Y are two proteins in the dataset. This normalized step was also used by SVM-pairwise (Liao and Noble, 2003) and SVM-LA (Saigo *et al.*, 2004). The normalized kernel thus obtained was then transformed into a radial basis kernel.

In the SVM-pairwise (Liao and Noble, 2003) method, the feature vector was a list of pairwise sequence similarity scores, computed with respect to all of the sequences in the training set. The radial basis function

Downloaded from http://bioinformatics.oxfordjournals.org/ at King Abdulaziz University on April 3, 2016

was used as the kernel. The rest steps were the same as the ones used in SVM-Ngram (Dong *et al.*, 2006).

In the SVM-LA (Saigo *et al.*, 2004), the kernel was calculated by summing up scores obtained from the local alignments with gaps between the two sequences, computed by Smith–Waterman dynamic programming algorithm. Such kernel might not be a positive definite kernel and the authors (Saigo *et al.*, 2004) provided two solutions for this problem. Owing to its performance and simplicity, we implemented one of the two methods, namely, the LA-ekm kernel. The parameters of LA-ekm kernel took the optimal values ($\beta = 0.5$, $d = -11$, $e = -4$).

2.5 Multiple kernel learning

The kernel described in the previous section can be used by kernel methods to train the SVM classifier. Each kernel contains different discriminative information, and therefore combining the kernels automatically is a promising way to improve the performance. In machine learning field, this approach is called multiple kernel learning (MKL) (Cortes *et al.*, 2010; Varma and Babu, 2009), which has attracted a lot of attention recently. The MKL technique aimed to combine different kernels to improve the performance, and showed the state-of-the-art results on image classification field (Varma and Babu, 2009). In this article, we focused on the weighted linear combination of kernels. The weight of each kernel can be optimized based on different criterion, which can be categorized by two groups. One group is the one-stage kernel learning methods, which optimize the weight and the SVM objective function simultaneously (Varma and Babu, 2009). These methods suffer from the high training complexity. The other group is two-stage kernel learning methods, which optimize the weight by using a criterion first and then train the SVM classifier using the kernel combined by the learned weight of each kernel. Compared with one-stage learning methods, the two-stage kernel learning methods showed better performance with reduced training cost. Therefore, in this study, we adopted the two-stage kernel learning method. Specifically, the kernel target alignment (KTA) objective function was used to optimize the weight of each kernel, which showed theoretical guarantees and can improve the performance in practice (Cortes *et al.*, 2010; Varma and Babu, 2009).

Given m training samples $x_1, x_2, \ldots, x_m$ and their corresponding labels $y_1, y_2, \ldots, y_m$, the ideal kernel matrix can be formulated as $K_y = y^T y$, where y is the vector of labels $[y_1, y_2, \ldots, y_m]$. For the given n kernels $K_{1,}\ K_2, \ldots K_n$, the aim is to learn the weight of each kernel. To avoid the kernel scaling problem, we center kernel K_k and the corresponding ideal kernel K_y in feature space by the following equation:

$$K_{ck}(x_i, x_j) = K_k(x_i, x_j) - \frac{1}{m}\sum_{i=1}^{m} K_k(x_i, x_j) - \frac{1}{m}\sum_{j=1}^{m} K_k(x_i, x_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} K_k(x_i, x_j) \tag{10}$$

where K_{ck} is normalized by:

$$K'_k(x_i, x_j) = \frac{m K_{ck}(x_i, x_j)}{\sum_{i=1}^{m} K_{ck}(x_i, x_j)} \tag{11}$$

Following the above steps, each kernel is normalized, and then these n kernels are linearly combined by the following equation:

$$K_{\mathrm{Comb}} = \sum_{k=1}^{n} w_k K'_k \tag{12}$$

where w_k ($0 \le w_k \le 1$, $\sum_{k=1}^{n} w_k = 1$) is the weight of kernel K'_k. The weight is learned by KTA objective function, which maximizes the alignment between K_{Comb} and the centered ideal kernel K_{cy}.

$$\rho(K_{\mathrm{Comb}}, K_{cy}) = \frac{K_{\mathrm{Comb}} \cdot K_{cy}}{\sqrt{(K_{\mathrm{Comb}} \cdot K_{\mathrm{Comb}}) \times (K_{cy} \cdot K_{cy})}} \tag{13}$$

This leads to a quadratic program problem and can be solved quite efficiently. For implementation details, please refer to (Cortes *et al.*, 2010).

In this study, three kernels (K_p, K_{p1} and K_{p2}) for each selected sequence-based method were linearly combined by using the above KTA approach to further improve the performance. For reader's convenience, the source code of the MKL is accessible by clicking the link at http://bioinformatics.hitsz.edu.cn/main/~binliu/remote/.

2.6 SVM

SVM is a class of supervised learning algorithms first introduced by Vapnik (1998). SVM-based machine learning algorithm has been successfully used to investigate various problems in molecular biology, such as identifying DNA recombination spots (Chen *et al.*, 2013), membrane protein types (Cai *et al.*, 2003) and heat-shock protein functions (Feng *et al.*, 2013), among many others. In this study, the publicly available Gist SVM package (http://www.chibi.ubc.ca/gist/) was used.

2.7 Evaluation methodology

Because the test sets have many more negative than positive samples, simply measuring error-rates will not give a good evaluation of performance. For the cases in which the positive and negative samples are not evenly distributed, the best way to evaluate the trade-off between the specificity and sensitivity is to use a receiver operating characteristic (ROC) score (Gribskov and Robinson, 1996). An ROC score is the normalized area under a curve that plots true positives against false positives for different classification thresholds. A score of 1 means perfect separation of positive samples from negative ones, whereas a score of 0 means that none of the sequences selected by the algorithm is positive. Another performance measure is ROC50 score, which is the area under the ROC curve up to the first 50 false positives.

3 RESULTS AND DISCUSSION

3.1 Profile-based protein representation can improve the performance of methods based on sequence composition

The frequency profile of a protein **P** can be converted into 20 profile-based proteins (p1, p2, ..., p20) by using the proposed approach (see Section 2 for details). These 20 proteins have different importance. p1 is the most important protein, as it is the combination of the top frequent amino acids in frequency profile, whereas p20 is the profile-based protein to which protein **P** is the most unlikely to convert because it is the combination of the amino acids with lowest frequencies in frequency profile. If all the 20 profile-based proteins are used in the prediction, the computational cost is relatively high. In this study, only the top n most important profile-based proteins (p1, ..., pn) were used in the prediction. To select the value of n, the following experiment was conducted. The frequencies of 20 standard amino acids in each column of a frequency profiles add up to 1. Therefore, the average frequency is 0.05 ($1/20 = 0.05$). If an amino acid with frequency >0.05, it is likely to occur during evolutionary process; otherwise, it is not likely to occur. The percentage of the amino acids with frequencies >0.05 in each profile-based protein on the SCOP benchmark was calculated, and the results are shown in Figure 2. As we can see from the figure, such amino acids are abundant in profile-based proteins p1, p2 and p3 (99.99%, 99.60% and 98.13%, respectively), but for the other 17 profile-based proteins, the percentage decreases

Downloaded from http://bioinformatics.oxfordjournals.org/ at King Abdulaziz University on April 3, 2016

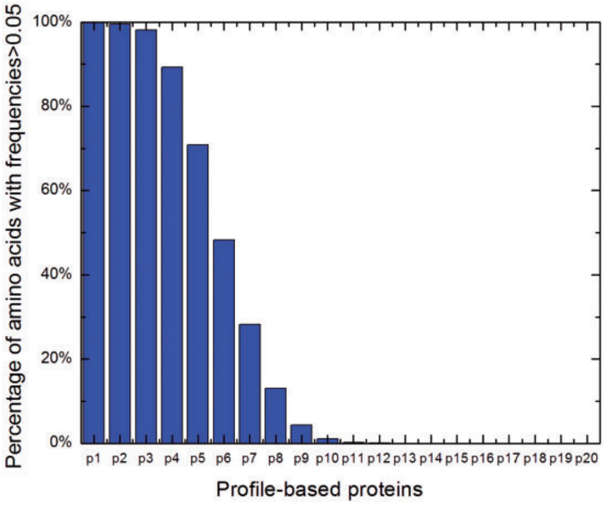

Fig. 2. Illustration to show the feature of frequency profile. Percentage of amino acids with frequencies > 0.05 in the 20 profile-based proteins derived from SCOP benchmark

significantly (from 89.28 to 0%). Therefore, in this study, only the top three profile-based proteins were used in the prediction. These profile-based proteins were combined with three state-of-the-art methods based on sequence composition, including SVM-Ngram (Dong *et al.*, 2006), SVM-pairwise (Liao and Noble, 2003) and SVM-LA (Saigo *et al.*, 2004), and the results are shown in Supplementary Table S1 of Supplementary Material S4. For each of the three methods, the best performance was achieved for the top important protein p1. Compared with the methods performed on the raw protein sequence **P**, the performance of the proposed methods can be improved by 3.7~7.5% and 9.6~13.7% in terms of average ROC and ROC50 scores, respectively, indicating that the proposed profile-based protein representation is useful for protein remote homology detection. The performance of the methods performed on p2 is similar as that of the methods performed on the raw protein **P**. The predictive results of the methods performed on p3 were the lowest. These results are consistent with the different importance of the three profile-based proteins p1, p2 and p3.

3.2 Comparison with closely related methods

Besides the current method, there are some other methods for predicting protein remote homologies based on profiles, such as SVM-Top-n-gram-combine-LSA (Liu *et al.*, 2008), SVM-PDT-Profile (Liu *et al.*, 2012), Profile (Kuang *et al.*, 2005), BioSVM-2L (Muda *et al.*, 2011) and HHSearch (Söding, 2005). SVM-Top-n-gram-combine-LSA (Liu *et al.*, 2008) extracted the building blocks of proteins from the frequency profiles, which could be treated as the 'words' of protein language. The LSA (Dong *et al.*, 2006) was applied to further improve the performance of this method. SVM-PDT-Profile (Liu *et al.*, 2012) combined the amino acid physicochemical properties in the Amino Acid Index (AAIndex) (Kawashima *et al.*, 2008) with the frequency profiles for the prediction. The feature vector of Profile method (Kuang *et al.*, 2005) was constructed by the short subsequences whose PSSM-based ungapped alignment score was above a predefined threshold. BioSVM-2L constructed two-layer SVM classifiers with profile-based kernels (Muda *et al.*, 2011). All the above three methods were based on SVM, and the difference among them was in the extracted features. HHSearch (Söding, 2005) was one of the best protein remote homology detection methods, which used a novel profile-based HMM. The results obtained by these four methods on the SCOP benchmark are listed in Supplementary Table S1 of Supplementary Material S4, from which we can see that the current method outperforms SVM-Top-n-gram-combine-LSA (Liu *et al.*, 2008), SVM-PDT-Profile (Liu *et al.*, 2012) and BioSVM-2L (Muda *et al.*, 2011) and is highly comparable with Profile (Kuang *et al.*, 2005) and HHSearch (Söding, 2005), indicating that the profile-based protein representation is a promising approach to extract the evolutionary information from frequency profiles for protein remote homology detection.

3.3 Combining different methods via MKL

As mentioned above, the approaches based on the top two profile-based proteins p1, p2 and the raw protein **P** are among the top performing methods. It is interesting to investigate whether these methods can be combined to further improve the performance. In this study, the MKL framework was used to combine these methods. The KTA method was used to automatically optimize the weight of each kernel on the training set, and then these kernels are combined with weights into a single kernel for the SVM-based prediction. The results are shown in Supplementary Table S2 of Supplementary Material S4 as well as Supplementary Materials S5–S7. The MKL approach can improve the performance of SVM-Ngram (Dong *et al.*, 2006), but only has minor impact on the SVM-pairwise (Liao and Noble, 2003) and SVM-LA (Saigo *et al.*, 2004). To uncover the reason, the weight of each kernel was analyzed. For each kernel, the average weight on all the 54 protein families is shown in Supplementary Table S2 of Supplementary Material S4. For these three methods, the p1-based kernel was weighted most heavily. For SVM-pairwise (Liao and Noble, 2003) and SVM-LA (Saigo *et al.*, 2004), the weight values of their corresponding **P** and p2 kernels are <0.1, indicating these kernels only have minor impact on the final results, and hence the performance improvement is modest. VBKC (Damoulas and Girolami, 2008) is another method based on the MKL, which combined four string kernels: SVM-pairwise (Liao and Noble, 2003), SVM-LA (Saigo *et al.*, 2004), SVM-MM (Leslie *et al.*, 2004) and SVM-Mono (Lingner and Meinicke, 2006). Our proposed SVM-pairwise-KTA and SVM-LA-KTA outperform VBKC (Damoulas and Girolami, 2008) by 1.2 ~ 2.2% and 29.9 ~ 31.3% according to the average ROC and ROC50 scores, respectively. The obvious performance improvement is mainly due to the proposed profile-based protein representation and MKL approach.

3.4 Correlations between discriminative power of Ngrams and protein families

The SVM-Ngram (Dong *et al.*, 2006) method is based on the explicit feature space representation, which provides the possibility to measure the correlations between Ngrams and protein families. The sequence-specific weight learnt from the SVM

Downloaded from http://bioinformatics.oxfordjournals.org/ at King Abdulaziz University on April 3, 2016

training process can be used to calculate the discriminant weight for each Ngram, which indicates the importance of the corresponding Ngram. By following Lingner and Meinicke's approach (Lingner and Meinicke, 2008), given the weight vector of a set of M sequences obtained from the kernel-based training process $\alpha = [\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_M]$, the discriminant weight vector w in the feature space can be calculated by the following equation:

$$w = \boldsymbol{F} * \alpha \tag{14}$$

where $\boldsymbol{F}$ is the matrix of sequence representatives. The magnitude of the element in w represents the discriminative power of the corresponding feature.

In most protein families, kernel p1 is weighted more heavily than kernel **P** and kernel p1. Two such protein families (SCOP ID: 2.1.1.4 and 3.2.1.5) were selected from the SCOP benchmark for further study, and the results are shown in Supplementary Tables S3 and S4 of Supplementary Material S4, respectively. For each kernel, the top 10 most discriminative Ngram features calculated by Equation 14 are shown in the tables too. For protein family 2.1.1.4, kernel **P** and kernel p1 share some common most discriminative Ngrams, such as 'mtm', 'yty', 'mtf' and 'wwf', indicating these Ngrams remain stable during evolutionary process and therefore these Ngrams would be the important sequence patterns for maintaining the structure and function of this protein family (Supplementary Table S3 of Supplementary Material S4). However, there are a few common most discriminative Ngrams between kernel **P** and kernel p1 in protein family 3.2.1.5. The top 10 most discriminative Ngrams of kernel p1 are all different from those in kernel **P** (Supplementary Table S4 of Supplementary Material S4). These Ngrams would contribute to the higher discriminative power of kernel p1 for this protein family.

Although in most cases, kernel p1 was weighted most heavily, some exceptions were observed. For example, for protein family 7.3.6.1, kernel p2 is the most discriminative kernel with weight value of nearly 1, while the other two kernels only have little contribution to the MKL (Supplementary Table S5 of Supplementary Material S4). The top 10 most discriminative Ngrams for each kernel were investigated, and the results are shown in Supplementary Table S5 of Supplementary Material S4, from which some interesting patterns can be observed. The Ngrams containing amino acids 'n' and 'f' tended to show strong discriminative power in both kernel **P** and kernel p1, whereas amino acid 'a' was abundant in the top discriminative Ngrams in kernel p2, indicating the Ngrams with amino acid 'a' could better describe the prosperities of protein family 7.3.6.1 in the evolutionary process.

3.5 Application of the proposed remote homology detection methods for studying the 3D structure of Nck5a

In addition to provide useful insights for evolution study, protein remote homology detection is useful for drug development as well. As is well known, many drug-targeted proteins are still without X-ray or nuclear magnetic resonance structure. Pharmaceutical scientists have to resort to the homology modeling technique or structural bioinformatics tools (Chou, 2004) to timely develop their 3D structures, so as to be able to conduct molecular docking study (Chou *et al*., 2003; Wang *et al*., 2009), one of the key steps in structure-based drug design. However, a reliable template, or a structure-known protein homologous to the target protein, is the necessary prerequisite in this regard (Chou, 2004). Unfortunately, many target proteins did not have significant sequence similarity with any structure-known proteins, and hence it was hard to find a proper template to develop their 3D structures. Actually, many of them did have structure-known homologous proteins, but the problem was how to detect them. For example, the sequence similarity between Nck5a and CyclinA was <20% (Chou *et al*., 1999) and hence their homologous relationship could not be detected by the simple sequence alignment technique (Mohabatkar, 2010). Now let us see what will happen if the current remote homology detection technique is applied.

To realize this, a dataset was constructed based on SCOP, from which 11 proteins were selected as the positive samples in the cyclin family (SCOP ID: a.74.1.1), while 3605 negative samples were selected from the SCOP version 1.67 by excluding all the proteins within the cyclin-like superfamily. None of these proteins shares >95% sequence similarity. Trained with such 11 positive proteins and 3605 negative proteins, the proposed best performing method SVM-LA (p1) was used to predict Nck5a. It was found that Nck5a is homologous to CyclinA, fully consistent with the experimental results obtained by the site-directed mutagenesis studies (Tang *et al*., 1997). Actually, Chou *et al*. (1999) did use CyclinA as a template to construct the 3D structure of activation domain of Nck5a, one of the important parts of tau protein kinase II, an important therapeutic target against Alzheimer's disease. Furthermore, based on the computed structure thus obtained, the molecular truncation experiments (Zhang *et al*., 2002) were conducted with an outcome that confirmed and validated the structure computed by using such a remote homologous protein as a template. Therefore, it is anticipated that the proposed method for detecting remote homology proteins will certainly enhance the power of homology modeling, and hence have impacts on drug development as well.

4 CONCLUSION

Discriminative methods based on SVM are the most effective and accurate methods for protein remote homology detection. The performance of the SVM-based methods depends on the kernel function, which measures the similarity between the samples in any pair. Varieties of kernels based on sequence composition have been proposed. However, these methods often fail to accurately predict the proteins sharing low sequence similarity. Recently, methods using the evolutionary information extracted from profiles achieved great success, such as Profile (Kuang *et al*., 2005), SW-PSSM (Rangwala and Karypis, 2005), SVM-Top-Ngram (Liu *et al*., 2008) and SVM-ACC (Liu *et al*., 2011). A key step to improve the performance of these methods is in how to find a suitable approach to incorporate the evolutionary information extracted from the profiles for prediction. In this article, we proposed a method that can convert the frequency profile into a series of profile-based proteins. Three state-of-the-art sequence-based kernels, i.e. SVM-Ngram (Dong *et al*., 2006), SVM-pairwise (Liao and Noble, 2003) and

Downloaded from http://bioinformatics.oxfordjournals.org/ at King Abdulaziz University on April 3, 2016

SVM-LA (Saigo *et al*., 2004), were selected for demonstration on a well-known benchmark. It was shown that the methods based on the profile-based proteins p1 and p2 achieved the best performance, outperforming the original three string kernels by 3.7 ~ 7.5% and 9.6 ~ 13.7%, respectively, according to the average ROC and ROC50 scores. These results are fully consistent with our previous findings that the top two most frequent amino acids show stronger discriminative power than the other low frequent amino acids in the frequency profiles (Liu *et al*., 2008), further confirming that the proposed profile-based protein representation is a promising approach in extracting the evolutionary information from frequency profiles for protein remote homology detection.

It has not escaped our notice that the current approach can be easily combined with sequence-based methods, and hence, with the development of the sequence-based kernels, the currently proposed method can be further improved accordingly. It is instructive to point out that since the concept of pseudo amino acid composition, or Chou's PseAAC (Lin and Lapointe, 2013), was introduced in 2001 (Chou, 2001), it has been successfully used to predict various attributes of proteins (e.g. Chen and Li, 2013; Chou, 2005; Georgiou *et al*., 2009; Huang and Yuan, 2013; Khosravian *et al*., 2013; Mohabatkar, 2010; Mohabatkar *et al*., 2011, 2013; Mohammad Beigi *et al*., 2011; Nanni *et al*., 2012; Sahu and Panda, 2010; Zhang *et al*., 2008; Zhou *et al*., 2007; Liu *et al*., 2013). Accordingly, the potential would be high to develop a powerful method for protein remote homology detection by combing PseAAC with profile-based protein representation. In the original PseAAC, it only uses three indices, including the hydrophobicity index, hydrophilicity index and side-chain mass index. Because protein remote homology detection is a more difficult problem, proteins in the dataset only share low sequence similarity. Only these three indices would not be enough to capture the different properties of various proteins. Therefore, our further research will focus on incorporating new amino acid indices into PseAAC and applying it to protein remote homology detection.

Funding: National Natural Science Foundation of China [numbers 61300112, 61370165, 61173075, 61272383 and 61370010), the Natural Science Foundation of Guangdong Province [numbers S2012040007390 and S2013010014475], the Scientific Research Innovation Foundation in Harbin Institute of Technology (project number HIT.NSRIF.2013103), the Shanghai Key Laboratory of Intelligent Information Processing, China [grant number IIPL-2012-002].

Conflict of Interest: none declared

REFERENCES

Althaus,I.W. *et al*. (1993) Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. *J. Biol. Chem*., **268**, 6119–6124.

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Altschul,S.F. *et al*. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Berman,H. *et al*. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic. Acids. Res*., **35**, D301–D303.

Brenner,S.E. *et al*. (2000) The ASTRAL compendium for sequence and structure analysis. *Nucleic Acids Res*., **28**, 254–256.

Cai,Y.D. *et al*. (2003) Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J*., **84**, 3257–3263.

Chen,W. *et al*. (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res*., **41**. http://dx.doi.org/doi:10.1093/nar/gks1450.

Chen,Y.K. and 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. (1995) The convergence-divergence duality in lectin domains of the selectin family and its implications. *FEBS Lett*., **363**, 123–126.

Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins*, **43**, 246–255.

Chou,K.C. (2004) Review: structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem*., **11**, 2105–2134.

Chou,K.C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **21**, 10–19.

Chou,K.C. (2010) Graphic rule for drug metabolism systems. *Curr. Drug Metab*., **11**, 369–378.

Chou,K.C. (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review). *J. Theor. Biol*., **273**, 236–247.

Chou,K.C. (2013) Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst*., **9**, 1092–1100.

Chou,K.C. *et al*. (1999) A model of the complex between cyclin-dependent kinase 5 (Cdk5) and the activation domain of neuronal Cdk5 activator. *Biochem. Biophys. Res. Commun*., **259**, 420–428.

Chou,K.C. *et al*. (2003) Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. *Biochem. Biophys. Res. Commun*., **308**, 148–151.

Chou,K.C. *et al*. (2011) Wenxiang: a web-server for drawing wenxiang diagrams. *Natural Science*, **3**, 862–865. http://dx.doi.org/10.4236/ns.2011.310111.

Cortes,C. *et al*. (2010) Two-stage learning kernel algorithms. In: *Proceedings of the 27th International Conference on Machine Learning*. pp. 239–246.

Damoulas,T. and Girolami,M.A. (2008) Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. *Bioinformatics*, **24**, 1264–1270.

Dong,Q.W. *et al*. (2006) Application of latent semantic analysis to protein remote homology detection. *Bioinformatics*, **22**, 285–290.

Feng,P.M. *et al*. (2013) iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem*., **442**, 118–125.

Georgiou,D.N. *et al*. (2009) Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *J. Theor. Biol*., **257**, 17–26.

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (Roc) analysis to evaluate sequence matching. *Comput. Chem*., **20**, 25–33.

Huang,C. and Yuan,J.Q. (2013) A multilabel model based on Chou's pseudo-amino acid composition for identifying membrane proteins with both single and multiple functional types. *J. Membr. Biol*., **246**, 327–334.

Karplus,K. *et al*. (1998) Hidden markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Kawashima,S. *et al*. (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*., **36**, D202–D205.

Khosravian,M. *et al*. (2013) Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. *Protein Pept. Lett*., **20**, 180–186.

Kuang,R. *et al*. (2005) Profile-based string kernels for remote homology detection and motif extraction. *J. Bioinform. Comput. Biol*., **3**, 527–550.

Leslie,C.S. *et al*. (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467–476.

Liao,L. and Noble,W.S. (2003) Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput Biol*., **10**, 857–868.

Lin,S.X. and Lapointe,J. (2013) Theoretical and experimental biology in one. *J. Biomed. Sci. Eng*, **6**, 435–442. http://dx.doi.org/10.4236/jbise.2013.64054.

Lingner,T. and Meinicke,P. (2006) Remote homology detection based on oligomer distances. *Bioinformatics*, **22**, 2224–2231.

Lingner,T. and Meinicke,P. (2008) Word correlation matrices for protein sequence analysis and remote homology detection. *BMC Bioinformatics*, **9**, 259.

Liu,B. *et al*. (2008) A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis. *BMC Bioinformatics*, **9**, 510.

Downloaded from http://bioinformatics.oxfordjournals.org/ at King Abdulaziz University on April 3, 2016

Liu,B. *et al*. (2009) Prediction of protein binding sites in protein structures using hidden Markov support machine. *BMC Bioinformaitcs*, **10**, 381.

Liu,B. *et al*. (2012) Using amino acid physicochemical distance transformation for fast protein remote homology detection. *PLoS One*, **7**, e46633.

Liu,B. *et al*. (2013) Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation. *Molecular Informatics*, **32**, 775–782.

Liu,X. *et al*. (2011) Protein remote homology detection based on auto-cross covariance transformation. *Comput. Biol. Med*., **41**, 640–647.

Loewenstein,Y. *et al*. (2009) Protein function annotation by homology-based inference. *Genome Biol*., **10**, 207.

Melvin,I. *et al*. (2011) Detecting remote evolutionary relationships among proteins by large-scale semantic embedding. *PLoS Comput. Biol*., **7**, e1001047.

Mohabatkar,H. (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein Pept. Lett*., **17**, 1207–1214.

Mohabatkar,H. *et al*. (2011) Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol*., **281**, 18–23.

Mohabatkar,H. *et al*. (2013) Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. *Med. Chem*., **9**, 133–137.

Mohammad Beigi,M. *et al*. (2011) Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. *J. Struct. Funct. Genomics*, **12**, 191–197.

Muda,H.M. *et al*. (2011) Remote protein homology detection and fold recognition using two-layer support vector machine classifiers. *Comput. Biol. Med*., **41**, 687–699.

Nanni,L. *et al*. (2012) Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of chou's pseudo amino acid composition and on evolutionary information. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **9**, 467–475.

Qian,B. and Goldstein,R.A. (2004) Performance of an iterated T-HMM for homology detection. *Bioinformatics*, **20**, 2175–2180.

Rangwala,H. and Karypis,G. (2005) Profile-based direct kernels for remote homology detection and fold recognition. *Bioinformatics*, **21**, 4239–4247.

Sadreyev,R.I. *et al*. (2009) COMPASS server for homology detection: improved statistical accuracy, speed and functionality. *Nucleic Acids Res*., **37**, W90–W94.

Sahu,S.S. and Panda,G. (2010) A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Comput. Biol. Chem*., **34**, 320–327.

Saigo,H. *et al*. (2004) Protein homology detection using string alignment kernels. *Bioinformatics*, **20**, 1682–1689.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.

Söding,J. (2005) Protein homology detection by HMM–HMM comparison. *Bioinformatics*, **21**, 951–960.

Tang,D. *et al*. (1997) Cyclin-dependent kinase 5 (Cdk5) activation domain of neuronal Cdk5 activator. Evidence of the existence of cyclin fold in neuronal Cdk5a activator. *J. Biol. Chem*., **272**, 12318–12327.

Vapnik,V.N. (1998) *Statistical Learning Theory*. Wiley-Interscience, New York.

Varma,M. and Babu,B.R. (2009) More generality in efficient multiple kernel learning. In: *Proceedings of the 26th International Conference on Machine Learning*. pp. 1065–1072.

Wang,J.F. *et al*. (2009) Insights from investigating the interactions of adamantane-based drugs with the M2 proton channel from the H1N1 swine virus. *Biochem. Biophys. Res. Commun*., **388**, 413–417.

Webb-Robertson,B.-J. *et al*. (2005) SVM-BALSA: remote homology detection based on Bayesian sequence alignment. *Comput. Biol. Chem*., **29**, 440–443.

Wu,C.H. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids. Res*., **34**, D187–D191.

Xiao,X. *et al*. (2013) iCDI-PseFpt: identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. *J. Theor. Biol*., **337C**, 71–79.

Xu,Y. *et al*. (2013) iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ*, **1**, e171. https://peerj.com/articles/171.pdf.

Zhang,J. *et al*. (2002) Identification of the N-terminal functional domains of Cdk5 by molecular truncation and computer modeling. *Proteins*, **48**, 447–453.

Zhang,S.W. *et al*. (2008) Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. *Amino Acids*, **34**, 565–572.

Zhou,G.P. and Deng,M.H. (1984) An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. *Biochem. J*., **222**, 169–176.

Zhou,G.P. (2011) The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. *J. Theor. Biol*., **284**, 142–148.

Zhou,X.B. *et al*. (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol*., **248**, 546–551.

Zhou,G.P. and Huang,R.B. (2013) The pH-Triggered Conversion of the PrP(c) to PrP(sc.). *Curr Top Med Chem*., **13**, 1152–1163.

Downloaded from http://bioinformatics.oxfordjournals.org/ at King Abdulaziz University on April 3, 2016