Mol Genet Genomics (2015) 290:1919–1931
DOI 10.1007/s00438-015-1044-4

ORIGINAL PAPER

Protein remote homology detection by combining Chou's distance-pair pseudo amino acid composition and principal component analysis

Bin Liu[1,2,3] · **Junjie Chen**[1] · **Xiaolong Wang**[1,2]

Received: 5 March 2015 / Accepted: 6 April 2015 / Published online: 21 April 2015

Abstract Protein remote homology detection is one of the important tasks in computational proteomics, which is important for basic research and practical application. Currently, the SVM-based discriminative methods have shown superior performance. However, the existing feature vectors still cannot suitably represent the protein sequences, and often lack an interpretable model for analysis of characteristic features. Previous studies showed that sequence-order effects and physicochemical properties are important for representing protein sequences. However, how to use these kinds of information for constructing predictors is still a challenging problem. In this study, in order to incorporate the sequence-order information and physicochemical properties into the prediction, a method called disPseAAC is proposed, in which the feature vector is constructed by combining the occurrences of amino acid pairs within the Chou's pseudo amino acid composition (PseAAC) approach. The predictive performance and computational cost are further improved by employing the principal component analysis strategy. Various experiments are conducted on a benchmark dataset. Experimental results show that disPseAAC achieves an ROC score of 0.922, outperforming some existing state-of-the-art methods. Furthermore, the learnt model can easily be analyzed in terms of discriminative features, and the computational cost of the proposed method is much lower than that of other profile-based methods.

Keywords Protein remote homology · Pseudo amino acid composition · Support vector machine · Principal component analysis

Communicated by S. Hohmann.

✉ Bin Liu
bliu@insun.hit.edu.cn

Junjie Chen
chenjunjie@hitsz.edu.cn

Xiaolong Wang
wangxl@insun.hit.edu.cn

[1] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town, Xili, Shenzhen 518055, Guangdong, People's Republic of China

[2] Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town, Xili, Shenzhen 518055, Guangdong, People's Republic of China

[3] Gordon Life Science Institute, Belmont, MA 02478, USA

Background

With the development of biology sequencing techniques, the protein sequence data show explosive growth, while the data of protein structure and function grow slowly. In structural biology field, many 3D protein structures, particularly for the structures of many transmembrane proteins, have not been determined by experimental methods such as NMR spectroscopy and X-ray crystallography (Zhou and Troy 2003, 2005a, b; Sharma et al. 2008; Bjorndahl et al. 2011; Zhou et al. 2015). For example, it is very difficult to make a qualified NMR study due to the membrane environment and the presence of a large number of hydrophobic amino acid residues in many transmembrane proteins and glycotransferase enzymes, as described in previous studies and recent review (Zhou and Troy 2005a; Zhou et al. 2015). Similarly, the preparation of a qualified X-ray crystal sample and data collection also faces difficulties because of the large flexible coil regions in the many proteins (Zhou et al.

2015). Thus, using protein sequences to effectively predict their structure and function is very important, which will promote protein-biomolecule interaction study. Therefore, it is necessary to use protein sequences to predict their functional and spatial structure (Lin et al. 2013a). As an important task in the field of protein structure and function prediction, the aim of protein remote homology detection is to detect the evolutionary homology in proteins with low similarities. During evolutionary process, structure is more conserved than sequence. Therefore, using protein sequences to effectively predict their structure and function is very important, which will promote protein binding study (Liu et al. 2009a, b, 2014b, 2015a; Song et al. 2014), rational drug design, and many other related fields and applications (Lin et al. 2013a; Zou et al. 2013). However, protein remote homology detection is still a challenging problem in bioinformatics, and therefore accurately and efficiently computational approaches are required. Some computational methods have been proposed to solve this task, which are mainly divided into four major categories: pairwise alignment methods, profile-based alignment methods, generative methods, and discriminative methods.

Early computational approaches for protein remote homology detection are pairwise alignment methods, also called sequence–sequence alignment methods. Pairwise alignment methods detect the family of query protein sequence based on the sequence similarity between query protein and other known structure proteins in a dataset. Commonly pairwise alignment algorithms include Needleman–Wunsch algorithm (Needleman and Wunsch 1970) of global alignment and Smith–Waterman algorithm (Smith and Waterman 1981) of local alignment. Later, their computational efficiency is extensively improved by the generative methods, such as BLAST (Altschul et al. 1990). However, when the protein sequence similarity is below 35 % at the amino acid level (Rost 1999), the alignment score usually falls into a twilight zone.

To overcome the disadvantages of pairwise alignment methods, the profile-based alignment methods are proposed. Profiles generated from multiple sequence alignments describe the occurrence frequencies of 20 standard amino acids in each position of a protein during evolutionary process. Profile is a smart representation of protein sequence, which contains the evolutional information. Therefore, the profile-based alignment methods show better performance than the pairwise alignment methods, such as PSI-BLAST (Altschul et al. 1997), IMPALA (Schäffer et al. 1999), and COMPASS (Sadreyev et al. 2009). In order to facilitate the researchers who are interested in detecting protein remote homology, some online servers are constructed, including FORTE (Tomii and Akiyama 2004), RANKPOOP (Noble et al. 2005), webPRC (Brandt and Heringa 2009), Phyre (Kelley and Sternberg 2009), GenThreader (Lobley et al. 2009), COMA (Margelevicius and Venclovas 2010), Bioshell (Gront et al. 2012), etc. Most of these web servers only require the protein sequences as the input.

Generative methods train a model to represent a protein family and then this model is used to evaluate a query protein sequence. For example, hidden Markov model (HMM) (Karplus et al. 1998) searches the protein database according to a model built by both positively labeled and unlabeled proteins. Because the generative methods can only use positive training samples to build the models, their performance is still too low to be used for real applications. Some generative methods incorporate more sensitive profiles and show better performance, for example, HHsearch (Såding 2005) employs the hidden Markov model to calculate a novel profile, and this method achieves the state-of-the-art performance in the field of protein remote homology detection.

In contrast to generative methods, the discriminative methods use the information of both positive and negative samples in given benchmark dataset. These approaches learn a distinction between two classes, and show better performance. The most widely used discriminative methods for remote homology detection are based on the support vector machines (SVMs) (Vapnik 1998). Given the positive and negative training samples, SVM builds a plane of maximum interval from the vector space to discriminate between the unseen samples. A key feature of SVM is that it requires fixed length feature vectors. Therefore, some researchers proposed various types of feature vectors for protein representation. Some methods are based on sequence composition information, including Fisher kernel (Jaakkola et al. 1999), SVM-Pairwise (Liao and Noble 2003), SVM-LA (Saigo et al. 2004), Monomer-dist (Lingner and Meinicke 2006), Motif kernel (Hur and Brutlag 2003), Mismatch (Leslie et al. 2004), LSA (Dong et al. 2006), SVM-DR (Liu et al. 2014c), PseAACIndex (Liu et al. 2013), etc. Because the sequence composition methods only use the sequence features without considering the evolutionary information or 3-dimension structure information, their performance is not satisfying. Later, the performance of discriminative approaches is further improved by incorporating protein profiles containing the evolutionary information, for example SW-PSSM (Rangwala and Karypis 2005), Profile kernel (Kuang et al. 2005), BioSVM (Muda et al. 2011), SVM-LSA (Dong et al. 2006), Top-n-gram (Liu et al. 2008), SVM-PDT (Liu et al. 2012), ACC (Liu et al. 2011), SVM-HMMSTR (Weston et al. 2004), SVM-RQA (Yang et al. 2008), Multiple Query (Joshi et al. 2013), Profile-based protein representation (Liu et al. 2014d), MRFalign (Ma et al. 2014), etc.

A key step to improve the performance of discriminative methods, such as SVM-based methods is to explore a fast

and accurate representation of protein sequence. Although the profile-based methods achieve the state-of-the-art performance, they show a significant disadvantage concerning the interpretability of the resulting discriminative model. These methods do not provide an intuitive insight into the associated feature space for further analysis of relevant sequence features leant from the data. Therefore, these methods do not offer additional utility for the researchers who are interested in finding the characteristic features of proteins. Another disadvantage of profile-based methods is their high computational cost preventing the application of these methods to a large database. By contrast, the feature vectors generated by the sequence-based methods with low computational cost can be analyzed to reveal the characteristic of protein families. However, if a protein is represented by its amino acid composition alone, all the information of its sequence-order and sequence length is lost (Ding et al. 2014a). Previous studies show that the sequence-order information is relevant for discrimination (Lin et al. 2013b; Ding et al. 2014b), because the structure and function properties of proteins are determined by the amino acid order information in proteins. However, it is often very difficult to incorporate the sequence-order information into the computational predictors, because the proteins have different length, and the very high number of different sequence order combinations as noted by Chou (2001).

To deal with such a dilemma, the pseudo amino acid composition (Chou 2001, 2005) or Chou's PseAAC (Cao et al. 2013; Lin and Lapointe 2013; Du et al. 2014) was proposed. A recent review paper (Chou 2014) systematically introduced the rapid development of PseAAC and its wide applications in medicinal chemistry on how to use PseAAC to deal with various problems in protein/peptide areas, as well as how it has been extended to deal with many important problems in DNA/RNA areas as well. For a brief introduction about PseAAC, see a Wikipedia article at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition. It has rapidly penetrated into almost all the areas of computational proteomics, such as identifying cysteine S-nitrosylation sites in proteins (Xu et al. 2013), predicting bacterial virulent proteins (Nanni et al. 2012), identifying bacterial secreted proteins (Yu et al. 2010), predicting antifreeze proteins (Mondal and Pai 2014), predicting supersecondary structure (Zou et al. 2011), predicting protein subcellular location in various organisms (Zhang et al. 2008b, 2014a; Lin et al. 2009; Kandaswamy et al. 2010; Fan and Li 2012a; Mei 2012b; Chang et al. 2013; Huang and Yuan 2013a, b; Qin et al. 2013; Wan et al. 2013; Li et al. 2014; Zuo et al. 2014; Dehzangi et al. 2015), predicting membrane protein types (Chen and Li 2013; Huang and Yuan 2013b; Han et al. 2014), discriminating outer membrane proteins (Hayat and Khan 2012), identifying antibacterial peptides (Khosravian et al. 2013), identifying allergenic proteins (Mohabatkar et al. 2013), predicting metalloproteinase family (Mohammad Beigi et al. 2011), predicting protein structural class (Sahu and Panda 2010; Hayat and Iqbal 2014; Kong et al. 2014; Nanni et al. 2014; Zhang et al. 2014c), identifying GPCRs and their types (Zia Ur and Khan 2012; Xie et al. 2013), identifying protein quaternary structural attributes (Zhang et al. 2008a; Sun et al. 2012), predicting protein submitochondria locations (Nanni and Lumini 2008; Zeng et al. 2009; Fan and Li 2012b; Mei 2012a), identifying risk type of human papillomaviruses (Esmaeili et al. 2010), identifying cyclin proteins (Mohabatkar 2010), predicting GABA(A) receptor proteins (Mohabatkar et al. 2011), classifying amino acids (Georgiou et al. 2009), predicting the cofactors of oxidoreductases (Zhang and Fang 2008), predicting enzyme subfamily classes (Zhou et al. 2007), detecting remote homologous proteins (Liu et al. 2013, 2014d), analyzing genetic sequences (Georgiou et al. 2013), predicting anticancer peptides (Hajisharifi et al. 2014), predicting DNA binding proteins (Liu et al. 2015e), predicting protein binding sites (Liu et al. 2014a), predicting S-nitrosylation sites in proteins (Jia et al. 2014; Zhang et al. 2014b), among many others [see a long list of papers cited in the References section of a 2011 review article (Chou 2011)]. Also, PseAAC has been selected as one of the key topics for a special issue entitled "Molecular Science for Drug Development and Biomedicine" (Zhong and Zhou 2014). Because it has been widely and increasingly used, in addition to the web-server 'PseAAC' (Shen and Chou 2008) built in 2008, recently three powerful open access softwares, called 'PseAAC-Builder' (Du et al. 2012), 'propy' (Cao et al. 2013), and 'PseAAC-General' (Du et al. 2014), were established: the former two are for generating various modes of Chou's special PseAAC; while the third one for those of Chou's general PseAAC. Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA, RNA and nucleotides (Chen et al. 2012, 2013, 2014; Guo et al. 2014; Lin et al. 2014; Qiu et al. 2014; Liu et al. 2015b, c, d), as well as other biological samples [see, e.g., (Huang et al. 2012; Li et al. 2012; Jiang et al. 2013)].

Motivated by the successful application of computational approaches using PseAAC, in this study, we propose an intuitively interpretable feature space for protein sequences called distance-pair pseudo amino acid composition (disPseAAC) for protein remote homology detection. In this method, the PseAAC is improved by combining the occurrences of amino acid pairs within a given distance and various physicochemical property scores in the amino acid index (AAIndex) database (Kawashima et al. 2008). This approach is further improved by employing the principal component analysis (PCA) strategy (Pearson 1901), which cannot only improve the predictive performance, but can also reduce the computational cost. Experimental results show that the proposed disPseAAC achieves an ROC score of 0.922,

outperforming other compared methods, including SVM-DR (Liu et al. 2014c), PseAACIndex (Liu et al. 2013), Monomer-dist (Lingner and Meinicke 2006), Mismatch (Leslie et al. 2004), and SVM-Pairwise (Liao and Noble 2003).

Methods

Dataset description

A benchmark dataset (Liao and Noble 2003) is employed to evaluate the performance of different methods, which can be downloaded at http://noble.gs.washington.edu/proj/svm-pairwise/. Because this is a widely used benchmark dataset (Hur and Brutlag 2003; Liao and Noble 2003; Leslie et al. 2004; Saigo et al. 2004; Dong et al. 2006; Lingner and Meinicke 2006; Liu et al. 2013, 2014c), it can provide good comparability with previous methods. The benchmark is constructed based on SCOP version 1.53 (Andreeva et al. 2004) containing 54 families and 4352 proteins, which are extracted from the Astral database (Brenner et al. 2000) and no pairwise alignments is higher than an *E*-value of 10^{-25}. The target families with significant number of proteins are selected from the 1356 families so as to validate the performance of different method. Protein sequences within one SCOP family are treated as positive testing samples, and proteins outside the family but within the same superfamily are taken as positive training samples. Negative samples are selected from outside of the superfamily and are separated into training and testing sets.

Amino acid indices

The Amino Acid Index (AAIndex) (Kawashima et al. 2008) is a database of numerical indices, which represent various physicochemical properties of amino acids or pairs of amino acids (http://www.genome.jp/aaindex/). The AAIndex1 is employed in this study to incorporate the physicochemical properties of amino acids. After removing the incomplete data and the indices with all zero values, 531 indices are selected for further analysis.

Distance-pair pseudo amino acid composition (disPseAAC)

It is often difficult to effectively incorporate the sequence-order information into a discrete model or a vector without losing important information, because the lengths of different proteins vary greatly. Furthermore, the number of different physicochemical property combinations is very high, which has posed a difficulty for including physicochemical property information for a computational predictor. To deal with such problem, a new approach called distance-pair pseudo amino acid composition (disPseAAC) is proposed to approximately combine the sequence-order information and physicochemical properties into one feature vector. Here, we will introduce the process of the proposed disPseAAC.

Suppose a protein sequence **P** with L amino acids, as formulated by

$$\mathbf{P} = R_1 R_2 R_3 \ldots R_L \tag{1}$$

where R_1 is the first residue; R_2 is the second residue in protein sequence **P**, etc. The disPseAAC feature vector can be formulated as a vector given by

$$\mathbf{P} = \left[\begin{array}{cccccc} \psi_1 & \psi_2 & \ldots & \psi_u & \ldots & \psi_\Omega \end{array} \right]^{\mathrm{T}} \tag{2}$$

where Ω represents the dimension of feature vector.

Firstly, in order to capture the sequence-order information of the residues in **P**, we find all the amino acid pairs at a given distance d, called the "distance-pairs", which can be represented as:

$$\mathrm{DP}\left(R_i, R_j, d\right) \tag{3}$$

where R_i or R_j represents one of the 20 standard amino acids in a protein sequence; d represents the distance between R_i and R_j. This distance-pair means that amino acid R_i occurs before amino acid R_j, and the distance counted by the number of amino acids between R_i and R_j along the protein sequence is d.

Secondly, in order to incorporate the physicochemical properties for **P**, The sequence-order information associated with the physicochemical properties can be efficiently reflected by the following equation:

$$\theta^j(m) = \frac{\sum_{i=1}^{L-j} \Theta^m(R_i, R_{i+j})}{L-j} \quad (0 \le j \le \lambda) \tag{4}$$

where m represents an amino acid index in AAindex1; R_i is the amino acid at protein sequence position i, and so forth; j is the distance between two amino acids along the protein sequence; λ is the maximum physicochemical distance; $\Theta^m(R_i, R_{i+j})$ can be calculated by the following equation:

$$\Theta^m(R_i, R_{i+j}) = \left(I^m(R_i) - I^m(R_{i+j})\right)^2 \tag{5}$$

where $I^m(R_i)$ and $I^m(R_{i+j})$ represent the normalized physicochemical property values of amino acid R_i and R_{i+j} in index m, which can be calculated by the following equation:

$$I^m(R_i) = \frac{\widehat{I}^m(R_i) - \frac{1}{20}\sum_{k=1}^{20} \widehat{I}^m(R_k)}{\sqrt{\frac{1}{20}\sum_{k=1}^{20} \left(\widehat{I}^m(R_k) - \frac{1}{20}\sum_{k=1}^{20} \widehat{I}^m(R_k)\right)^2}} \tag{6}$$

where $\hat{I}^m(R_i)$ represents the raw physicochemical property value of amino acid R_i in index m; R_k ($k = 1, 2, 3, 4,\ldots, 20$) represent the 20 standard amino acids.

Finally, the elements of Eq. (2) can be calculated by the following equation:

$$\psi_u = \begin{cases} \dfrac{f(n,d)}{1+\omega \sum_{i=1}^{\lambda} \sum_{n=1}^{N} \theta^j(m)} \\ \dfrac{\omega\theta^j(m)}{1+\omega \sum_{i=1}^{\lambda} \sum_{M=1}^{M} \theta^j(m)} \end{cases} \quad \begin{array}{l} (1 \le n \le N, 0 \le d \le D, \\ 1 \le m \le M, 1 \le j \le \lambda) \end{array} \tag{7}$$

where ω is the factor to adjust the weight between the distance-pairs and physicochemical properties; N represents the total number of different distance-paris; D is a maximum pairwise distance ($0 \le D \le$ length of the longest protein in the benchmark dataset); M is the total number of different physicochemical property values; $f(n,d)$ represents the frequency of a distance-pair with distance d, which can be calculated by:

$$f(n,d) = \begin{cases} f\left(\mathrm{DP}(R_i,R_j,0)\right) & 1 \le n \le 20 \\ f\left(\mathrm{DP}(R_i,R_j,1)\right) & 1 \le n \le 400 \\ f\left(\mathrm{DP}(R_i,R_j,2)\right) & 1 \le n \le 400 \\ \vdots & \vdots \\ f\left(\mathrm{DP}(R_i,R_j|d)\right) & 1 \le n \le 400 \end{cases} \quad (0 \le d \le D) \tag{8}$$

where n is the number of distance-pairs at distance d.

Therefore, the dimension of the feature vector of Eq. (2) is $20 + 400D + M \times \lambda$. In order to help the readers to understand how to generate the disPseAAC feature vector, an example is given in Fig. 1, which shows the process of converting a protein P into the final feature vector.

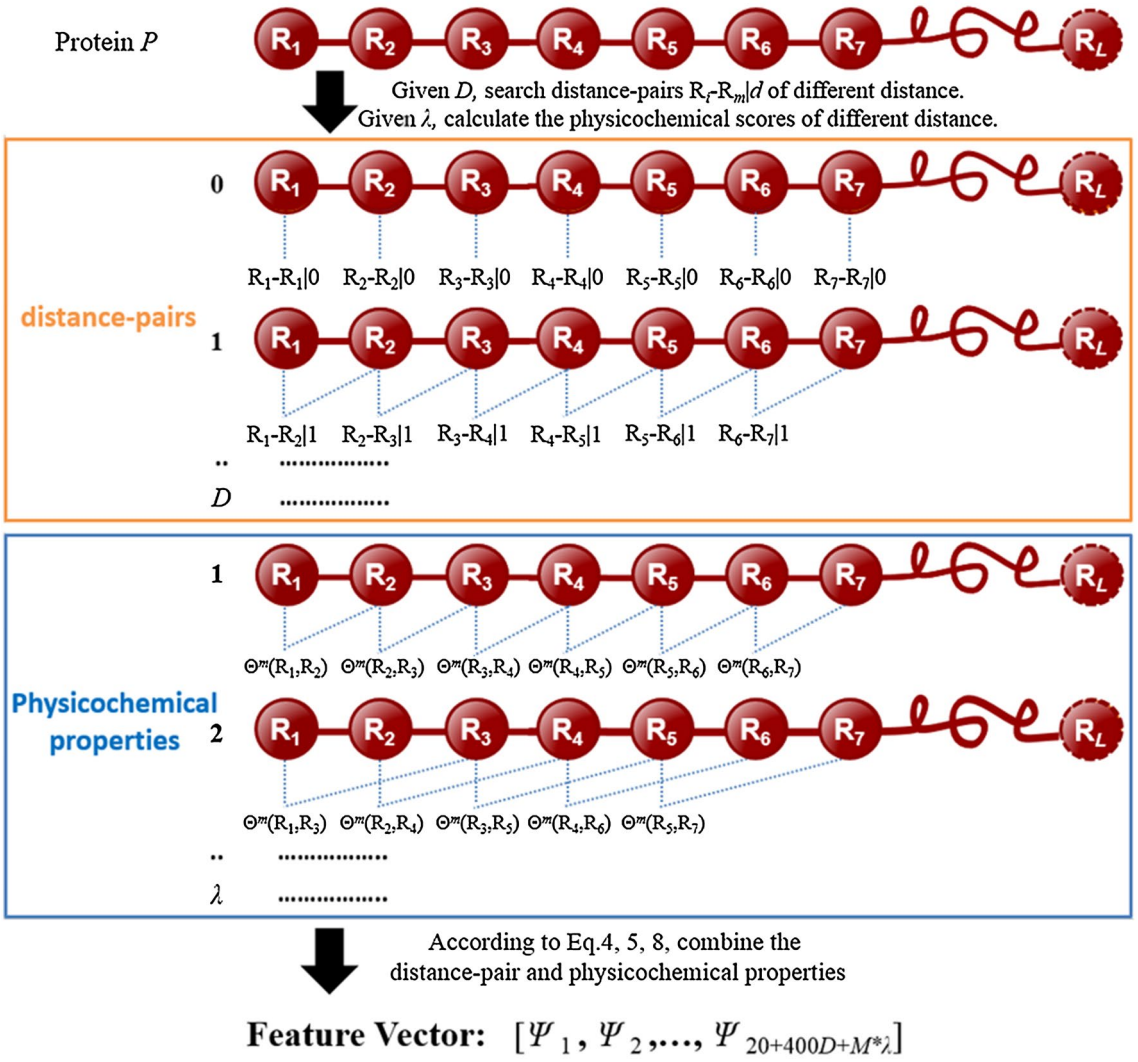

Fig. 1 The process of generating disPseAAC. *This figure* shows the process of converting a protein P into a feature vector of disPseAAC

Principal component analysis

Although we have removed the incomplete data and the indices with all zeros in AAIndex1 (Kawashima et al. 2008), there are still some redundant indices. In order to overcome this problem, we employ a feature extraction method, called principal component analysis (PCA) (Pearson 1901), to reduce the noise. PCA is a simple, non-parametric method for modern feature extraction, which can extract relevant information from confusing datasets. The aim of PCA is to select the important variables to represent the original data through a linear transformation. PCA transforms original data to a new coordinate system, in which the greatest variance on the first coordinate is called the first principal component, and the second greatest variance on the second coordinate is called the second principal component, etc. Therefore, the new feature space is a linear combination of the original feature space, and its dimension is much lower than that of the original one. The following describes the process of PCA for extracting important indices.

Firstly, we perform PCA on the distance-based feature vectors (Liu et al. 2012) to select the most discriminative physicochemical properties, which can be represented as u_1, u_2,…u_{531}. Their standard vectors can be represented as U_1, U_2,…U_{531}, which can be calculated by the following equation:

$$U_{i,j} = u_{i,j} - \bar{u}_j, \quad j = 1, 2, \ldots, 531 \tag{9}$$

where the $U_{i,j}$ represents the i-th feature of U_j; $u_{i,j}$ represents the i-th feature of u_j; $\bar{u}_j$ represents the average value of u_j. Therefore, the standard data matrix $\mathbf{U}$ can be represented as:

$$\mathbf{U} = \begin{bmatrix} U_{1,1} & U_{1,2} & \cdots & U_{1,531} \\ U_{2,1} & U_{1,2} & \cdots & U_{2,531} \\ \vdots & \vdots & \ddots & \vdots \\ U_{m,1} & U_{m,2} & \cdots & U_{m,531} \end{bmatrix} \tag{10}$$

where n is the number of samples in benchmark dataset. Then the eigenvalue and eigenvector of covariance matrix Cov($\mathbf{U}$) are calculated by

$$\mathrm{Cov}(\mathbf{U}) \bullet \mathbf{V} = \mathbf{V} \bullet \begin{bmatrix} \pi_1 & & & 0 \\ & \pi_2 & & \\ & & \ddots & \\ 0 & & & \pi_{531} \end{bmatrix} \tag{11}$$

where the π_i represents the i-th eigenvalue, as well as the weight of the i-th feature; $\mathbf{V}$ represents the eigenmatrix, while the feature of biggest eigenvalue is the first principal component, and index of second biggest eigenvalue is the second principal component, etc. The eigenvalues of physicochemical property features are sort in descending order so as to find the most important physicochemical indices:

$$\pi'_1 \geq \pi'_2 \geq \cdots \geq \pi'_{531} \tag{12}$$

where the π'_i represents the i-th sorted eigenvalue of indices. The principle of selecting indices is base on the cumulative weight ratio w:

$$w = \frac{\sum_{i=1}^{k} \pi'_i}{\sum_{i=1}^{531} \pi'_i} \tag{13}$$

In this study, the first k principal components are selected, whose corresponding cumulative weight ratio w is higher than 0.92 (please refer to the "Results" and "Discussion" section for more details).

Construction of SVM classifiers and classification

In the field of protein remote homology detection, the most widely used machine learning algorithm is support vector machine (SVM), which is proposed by Vapnik (1998). Given the positive and negative training samples, SVM builds a plane of maximum interval from the vector space to discriminate between the unseen protein sequences. A key feature of SVM is that it needs fixed length feature vectors as inputs. The proteins in the dataset are transformed into fixed-dimension feature vectors by the disPseAAC approach introduced above, and then the training feature vectors are input into SVM to construct the classifier, which will be used for predicting the unseen samples in the testing set. In order to fairly compare with other related methods, we employ the publicly available Gist SVM package version 2.3 (http://www.chibi.ubc.ca/gist/index.html) as the implementation of SVM algorithm. The SVM parameters are set as default values of the Gist Package except that the kernel function is set as either a quadratic or a radial basis function.

Evaluation methodology

Since graphic approaches can provide useful intuitive insights [see, e.g., (Chou and Forsen 1980; Zhou and Deng 1984; Chou 1989, 2010; Althaus et al. 1993; Zhou 2011; Zhou and Huang 2013)], a graphic comparison of the current predictor with their counterparts via the receiver operating characteristic (ROC) plot is preferred. However, there are 54 protein families in the benchmark dataset, and these results cannot be shown in one figure. Therefore, we apply the average receiver operating characteristics (ROC) score and average ROC50 score to measure the performance of

different methods for the 54 protein family datasets. The ROC score is the normalized area under a curve that plots true positives against false positives for different possible thresholds and the ROC50 score is the area under the ROC curve up to the first 50 false positives. The discriminative score obtained by the SVM classifier can be used to calculate the ROC score and ROC50 score.

Results

Performance of disPseAAC

There are four parameters in the proposed disPseAAC predictor, namely D, λ, M and ω, which would influence its predictive performance. In this study, D is the maximum pairwise distance; λ is the maximum physicochemical distance; M is the number of amino acid index; ω is the factor to adjust the weight between the distance-pairs and physicochemical properties. Generally speaking, the greater the D is, the more distance-pairs are incorporated. The greater the λ is, the more physicochemical distances are taken into account. However, if D or λ is too large, it would cause the "overfitting" or "high dimension disaster" problem. The number of M depends on the weight ratio w of selected indices [ef. Eq. (13)]. Accordingly, in the current study, their optimal values are determined within the ranges as defined below

$$\begin{cases} 1 \le D \le 9 & \Delta D = 2 \\ 1 \le \lambda \le 15 & \Delta\lambda = 2 \\ 0.01 \le \omega \le 0.09 & \Delta\omega = 0.02 \\ 0.80 \le w \le 1.0 & \Delta w = 0.02 \end{cases} \tag{14}$$

It can be seen from Eq. (14) that, to determine the optimal values for the four parameters, $5 \times 8 \times 5 \times 10 = 2000$ different combination cases need to be considered. The values of the four parameters are optimized based on the ROC scores calculated from all the 2000 combination cases, as given by

$$\begin{cases} D = 5 \\ \lambda = 9 \\ \omega = 0.03 \\ w = 0.92 \end{cases} \tag{15}$$

Comparison of closely related sequence-based methods

Six state-of-the-art sequence-based methods for protein remote homology detection are compare with the proposed method disPseAAC, including SVM-DR (Liu et al. 2014c), PseAACIndex (Liu et al. 2013), Monomer-dist (Lingner and Meinicke 2006), SVM-LA (Saigo et al. 2004), Mismatch (Leslie et al. 2004), and SVM-Pairwise (Liao and Noble 2003). SVM-DR (Liu et al. 2014c) is based on the distance-pairs, PseAACIndex (Liu et al. 2013) is based on the pseudo amino acid composition (PseAAC). Monomer-dist (Lingner and Meinicke 2006) constructs the feature vectors by the occurrences of short oligomers. The kernel of SVM-LA (Saigo et al. 2004) measures the similarity between a pair of proteins by taking into account all the optimal local alignment scores with gaps between all possible subsequences. Mismatch kernel (Leslie et al. 2004) is calculated based on occurrences of (k, m)-patterns in the data. In SVM-Pairwise (Liao and Noble 2003), each protein is represented as a vector of pairwise similarities to all proteins in the training set.

Table 1 Average ROC and ROC50 scores over 54 families for different methods

Methods	ROC	ROC50	Source
disPseAAC	0.922	0.721	This study
SVM-DR	0.919	0.715	Liu et al. (2014c)
PseAACIndex ($\lambda = 5$)	0.880	0.620	Liu et al. (2013)
Monomer-dist	0.919	0.508	Lingner and Meinicke (2006)
SVM-LA ($\beta = 0.5$)	0.925	0.649	Saigo et al. (2004)
Mismatch	0.872	0.400	Leslie et al. (2004)
SVM-pairwise	0.901	0.399	Liao and Noble (2003)

Table 1 shows the predictive results of the proposed method disPseAAC and other six related methods. The disPseAAC approach outperforms five sequence-based methods, including SVM-DR (Liu et al. 2014c), PseAACIndex (Liu et al. 2013), Monomer-dist (Lingner and Meinicke 2006), Mismatch (Leslie et al. 2004), and SVM-Pairwise (Liao and Noble 2003), and is highly comparable with SVM-LA (Saigo et al. 2004). The experimental results demonstrate that the disPseAAC approach is a useful computational method for protein remote homology detection.

Discussion

Correlations between discriminative features and protein family

In order to investigate the importance of the features and reveal the biological meaning of the feature space, we followed the study (Liu et al. 2012) to calculate the discriminant weight vector in the feature space. The sequence-specific weight obtained from the SVM training process is used to calculate the discriminant weight of each feature in order to reveal the importance of different features. Given the weight vectors of the training set with N samples

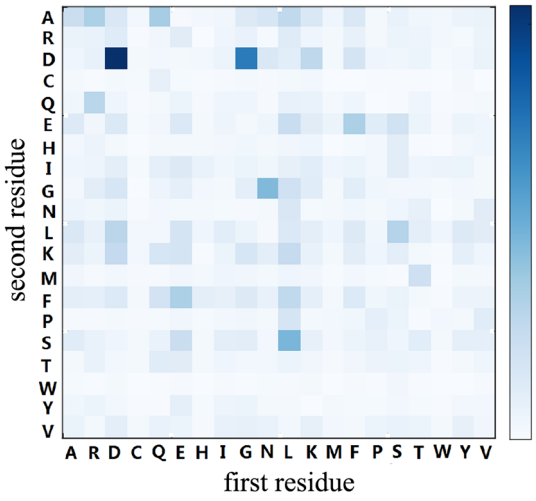

Fig. 2 The discriminative power (L_2-norm) of discriminant vectors for all the combinations of distance-pairs for protein family 1.41.1.5. The amino acids are identified by their *one-letter code*. The amino acids labeled by *x*-axis and *y*-axis in *this figure* indicate the first residue and the second residue in distance-pairs of disPseAAC, respectively. The *adjacent color bar* shows the mapping of L_2-norm values (color figure online)

obtained from the kernel-based training $\mathbf{A} = [a_1, a_2, a_3,\ldots, a_N]$, the discriminant weight vector $\mathbf{W}$ in the feature space is calculated by the following equation:

$$\mathbf{W} = \mathbf{A} \cdot \mathbf{M} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}^{\mathbf{T}} \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1j} \\ m_{21} & m_{22} & \cdots & m_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ m_{N1} & m_{N2} & \cdots & m_{Nj} \end{bmatrix} \tag{16}$$

where $\mathbf{M}$ represents the matrix of sequence representation; $\mathbf{A}$ represent the weight vectors of the training samples; N is the number of training samples; j is the dimension of the feature vector. The elements in $\mathbf{W}$ reflect the discriminative power of the features. The L_2-norm of the discriminant weight vector is employed to measure the importance of these features. Family 1.41.1.5 is selected as a target family for the feature analysis.

The L_2-norm values of 400 distance-pairs for the proposed methods are depicted in Fig. 2. According to this figure, interestingly, the top two most discriminative distance-pairs are (*D*, *D*), (*D*, *G*) (the two darkest spots in Fig. 2, indicating the importance of amino acid *D* and *G*). The top five most important indices for family1.41.1.5 are listed in Table 2. The experimental results show that these sequence patterns would reflect the features of this protein family.

Computational efficiency

With the explosive growth of protein sequence data, high computational cost methods are not suitable for protein

Table 2 The top five most discriminative indices in disPseAAC for protein family 1.41.1.5

AA	Normalized frequency of N-terminal helix	Alpha-CH chemical shifts	Retention coefficient in TFA	Normalized frequency of extended structure	Free energy of solution in water (kcal/mole)
A	1.29	4.349	7.3	0.288	−0.368
R	0.44	4.396	−3.6	0.362	−1.03
N	0.81	4.755	−5.7	0.229	0
D	2.02	4.765	−2.9	0.271	2.06
C	0.66	4.686	−9.2	0.533	4.53
Q	1.22	4.373	−0.3	0.327	0.731
E	2.44	4.295	−7.1	0.262	1.77
G	0.76	3.972	−1.2	0.312	−0.525
H	0.73	4.63	−2.1	0.2	0
I	0.67	4.224	6.6	0.411	0.791
L	0.58	4.385	20	0.4	1.07
K	0.66	4.358	−3.7	0.265	0
M	0.71	4.513	5.6	0.375	0.656
F	0.61	4.663	19.2	0.318	1.06
P	2.01	4.471	5.1	0.34	−2.24
S	0.74	4.498	−4.1	0.354	−0.524
T	1.08	4.346	0.8	0.388	0
W	1.47	4.702	16.3	0.231	1.6
Y	0.68	4.604	5.9	0.429	4.91
V	0.61	4.184	3.5	0.495	0.401

remote homology detection. In this field, the SVM-based methods can achieve the-state-of-art performance. The computational cost of the feature vector construction is the main bottleneck preventing their application to large databases. For the methods based on protein sequences alignment, SVM-LA (Saigo et al. 2004) requires a local alignment step, whose computational cost is high, and SW-PSSM (Rangwala and Karypis 2005) needs to search the query protein sequence against a non-redundant database in order to build the profiles. For the methods based on protein sequences composition, the computational cost of the proposed disPseAAC and Mismatch (Leslie et al. 2004) is much lower, because they do not require any time consuming alignment step. As reported by Hochreiter et al. (2007), on a dataset with 20,000 protein sequences, the running time of SVM-LA and SW-PSSM is 550 and 620 h, respectively, while Mismatch method only requires 380 s for the same task, which is much faster than SVM-LA and SW-PSSM.

In order to further clarify the efficiency of the disPseAAC, its time complexity is given. In this approach, the protein sequence feature can be calculated by Eq. (7). The distance-pair feature can be calculated by Eq. (8) with a time complexity of $O(D \times l)$, where D is maximum pairwise distance and l is the length of the sequence. Given the optimal value 5 of d, the time complexity of distance-pair feature is $O(l)$. The physicochemical property features can be calculated by Eq. (4) with a time complexity of $O(M \times \lambda \times l)$, where M is the number of amino acid indices; λ is the maximum physicochemical distance; l is the length of the sequence. Given the optimal values of M and λ, the time complexity of generating distance-pairs feature is $O(l)$. Therefore, the total time complexity of disPseAAC is $O(l)$. All the 4352 protein sequences in the SCOP1.53 dataset can be converted into fixed length vectors via disPseAAC within 100 s. This test is performed on a personal computer with CPU of 3.2 GHz and memory of 8 GB. The time complexity of SVM-DR (Liu et al. 2014c) is also $O(l)$, but it requires 171 s for the computation of the same task.

In this study, we proposed a sequence-based method called disPseAAC for protein remote homology detection, in which the feature vectors are constructed based on the occurrences of distance-pairs within a given distance and various physicochemical property scores in the AAIndex1 database. By using this approach, sequence-order information and physicochemical properties are incorporated into the disPseAAC predictor. Experimental results show that disPseAAC outperforms some sequence-based methods in this field. These results are not surprising, because many studies show that the sequence-order information and physicochemical properties of proteins are important for improving the predictive performance. The predictive precision and computational efficiency of disPseAAC are further improved by using the PCA strategy. Another important advantage of our approach arises from the explicit feature space representation: the possibility to calculate the discriminant weight vector in feature space, which allows the users to analyze the learnt model for identifying the most discriminative features. Some interesting patterns are discovered for the target protein family.

Acknowledgments This work was supported by the National Natural Science Foundation of China (No. 61300112 and 61272383), the Scientific Research Innovation Foundation in Harbin Institute of Technology (Project No. HIT.NSRIF.2013103), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, the Natural Science Foundation of Guangdong Province (2014A030313695), and Shenzhen Municipal Science and Technology Innovation Council (Grant No. CXZZ20140904154910774).

Conflict of interest The authors declare that they have no competing interests.

Ethical standard This article does not contain any studies with human participants or animals performed by any of the authors.

References

Althaus IW, Chou JJ, Gonzales AJ, Deibel MR, Chou KC, Kezdy FJ, Romero DL, Palmer JR, Thomas RC (1993) Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32:6548–6554

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32:D226–D229

Bjorndahl TC, Zhou GP, Liu X, Perez-Pineiro R, Semenchenko V, Saleem F, Acharya S, Bujold A, Sobsey CA, Wishart DS (2011) Detailed biophysical characterization of the acid-induced PrPc to PrPβ conversion process. Biochemistry 50:1162–1173

Brandt BW, Heringa J (2009) WebPRC: the profile comparer for alignment-based searching of public domain databases. Nucleic Acids Res 37:W48–W52

Brenner SE, Koehl P, Levitt M (2000) The ASTRAL compendium for sequence and structure analysis. Nucleic Acids Res 28:254–256

Cao DS, Xu QS, Liang YZ (2013) Propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29:960–962

Chang TH, Wu LC, Lee TY, Chen SP, Huang HD, Horng JT (2013) EuLoc: a web-server for accurately predict protein subcellular localization in eukaryotes by incorporating various features of sequence segments into the general form of Chou's PseAAC. J Comput Aided Mol Des 27:91–103

Chen YK, Li KB (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. J Theor Biol 318:1–12

Chen W, Lin H, Feng PM, Ding C, Zuo YC, Chou KC (2012) iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PLoS One 7:e47843

Chen W, Feng PM, Lin H, Chou KC (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41:e68

Chen W, Lei TY, Jin DC, Lin H, Chou KC (2014) PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. Anal Biochem 456:53–60

Chou KC (1989) Graphic rules in steady and non-steady state enzyme kinetics. J Biol Chem 264:12074–12079

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins Struct Func Genet 43:246–255 **(Erratum: ibid., 2001, vol 44, 60)**

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19

Chou KC (2010) Graphic rule for drug metabolism systems. Curr Drug Metab 11:369–378

Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). J Theor Biol 273:236–247

Chou KC (2014) Impacts of bioinformatics to medicinal chemistry. Med Chem (Shariqah, United Arab Emirates)

Chou KC, Forsen S (1980) Graphical rules for enzyme-catalyzed rate laws. Biochemistry 187:829–835

Dehzangi A, Heffernan R, Sharma A, Lyons J, Paliwal K, Sattar A (2015) Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. J Theor Biol 364:284–294

Ding H, Deng EZ, Yuan LF, Liu L, Lin H, Chen W, Chou KC (2014a) iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. Biomed Res Int 2014:286419

Ding H, Lin H, Chen W, Li ZQ, Guo FB, Huang J, Rao N (2014b) Prediction of protein structural classes based on feature selection technique. Interdiscip Sci 6:235–240

Dong QW, Wang XL, Lin L (2006) Application of latent semantic analysis to protein remote homology detection. Bioinformatics 22:285–290

Du P, Wang X, Xu C, Gao Y (2012) PseAAC-builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Anal Biochem 425:117–119

Du P, Gu S, Jiao Y (2014) PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. Int J Mol Sci 15:3495–3506

Esmaeili M, Mohabatkar H, Mohsenzadeh S (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J Theor Biol 263:203–209

Fan GL, Li QZ (2012a) Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition. J Theor Biol 304:88–95

Fan GL, Li QZ (2012b) Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition. Amino Acids 43:545–555

Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (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

Georgiou DN, Karakasidis TE, Megaritis AC (2013) A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory. Open Bioinform J 7:41–48; open access at http://www.benthamscience.com/open/tobioij/articles/V007/SI0025TOBIOIJ/0041TOBIOIJ.pdf

Gront D, Blaszczyk M, Wojciechowski P, Kolinski A (2012) BioShell threader: protein homology detection based on sequence profiles and secondary structure profiles. Nucleic Acids Res 40:W257–W262

Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, Chen W, Chou KC (2014) iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30:1522–1529

Hajisharifi Z, Piryaiee M, Mohammad Beigi M, Behbahani M, Mohabatkar H (2014) Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. J Theor Biol 341:34–40

Han GS, Yu ZG, Anh V (2014) A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC. J Theor Biol 344:31–39

Hayat M, Iqbal N (2014) Discriminating protein structure classes by incorporating pseudo average chemical shift to Chou's general PseAAC and support vector machine. Comput Methods Programs Biomed 116:184–192

Hayat M, Khan A (2012) Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC. Protein Pept Lett 19:411–421

Hochreiter S, Heusel M, Obermayer K (2007) Fast model-based protein homology detection without alignment. Bioinformatics 23:1728–1736

Huang C, Yuan J (2013a) Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites. Biosystems 113:50–57

Huang C, Yuan JQ (2013b) Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions. J Theor Biol 335:205–212

Huang T, Wang J, Cai YD, Yu H, Chou KC (2012) Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma. PLoS One 7:e34460

Hur AB, Brutlag D (2003) Remote homology detection: a motif based approach. Bioinformatics 19:i26–i33

Jaakkola T, Diekhans M, Haussler D (1999) Using the Fisher Kernel method to detect remote protein homologies. In: Proceedings of the 7th international conference on intelligent systems for molecular biology, pp 149–158

Jia C, Lin X, Wang Z (2014) Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition. Int J Mol Sci 15:10410–10423

Jiang Y, Huang T, Chen L, Gao YF, Cai Y, Chou KC (2013) Signal propagation in protein interaction network during colorectal cancer progression. Biomed Res Int 2013:287019

Joshi AG, Raghavender US, Sowdhamini R (2013) Improved performance of sequence search algorithms in remote homology detection. F1000 Res 2:93

Kandaswamy KK, Pugalenthi G, Moller S, Hartmann E, Kalies KU, Suganthan PN, Martinetz T (2010) Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition. Protein Pept Lett 17:1473–1479

Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14:846–856

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36:D202–D205

Kelley LA, Sternberg MJ (2009) Protein structure prediction on the web: a case study using the phyre server. Nat Protoc 4:363–371

Khosravian M, Faramarzi FK, Beigi MM, Behbahani M, Mohabatkar H (2013) Predicting antibacterial peptides by the concept of

Chou's pseudo-amino acid composition and machine learning methods. Protein Pept Lett 20:180–186
Kong L, Zhang L, Lv J (2014) Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition. J Theor Biol 344:12–18
Kuang R, Ie E, Wang K, Wang K, Siddiqi M (2005) Profile-based direct kernels for remote homology detection and motif extraction. J Bioinform Comput Biol 3:527–550
Leslie CS, Eskin E, Cohen A, Weston J, Noble WS (2004) Mismatch string kernels for discriminative protein classification. Bioinformatics 20:467–476
Li BQ, Huang T, Liu L, Cai YD, Chou KC (2012) Identification of colorectal cancer related genes with mRMR and shortest path in protein–protein interaction network. PLoS One 7:e33393
Li L, Yu S, Xiao W, Li Y, Li M, Huang L, Zheng X, Zhou S, Yang H (2014) Prediction of bacterial protein subcellular localization by incorporating various features into Chou's PseAAC and a backward feature selection approach. Biochimie 104:100–107
Liao L, Noble WS (2003) Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. J Comput Biol 10:857–868
Lin SX, Lapointe J (2013) Theoretical and experimental biology in one—a symposium in honour of Professor Kuo-Chen Chou's 50th anniversary and Professor Richard Giegé's 40th anniversary of their scientific careers. J Biomed Sci Eng (JBiSE) 6:435–442
Lin H, Wang H, Ding H, Chen YL, Li QZ (2009) prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. Acta Biotheor 57:321–330
Lin C, Zou Y, Qin J, Liu X, Jiang Y, Ke C, Zou Q (2013a) Hierarchical classification of protein folds using a novel ensemble classifier. PLoS One 8:e56499
Lin H, Chen W, Ding H (2013b) AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes. PLoS One 8:e75726
Lin H, Deng EZ, Ding H, Chen W, Chou KC (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res 42:12961–12972
Lingner T, Meinicke P (2006) Remote homology detection based on oligomer distances. Bioinformatics 22:2224–2231
Liu B, Wang X, Lin L, Dong Q, Wang X (2008) A discriminative method for protein remote homology detection and fold recognition combining top-*n*-grams and latent semantic analysis. BMC Bioinform 9:510
Liu B, Wang X, Lin L, Dong Q, Wang X (2009a) Exploiting three kinds of interface propensities to identify protein binding sites. Comput Biol Chem 33:303–311
Liu B, Wang X, Lin L, Tang B, Dong Q, Wang X (2009b) Prediction of protein binding sites in protein structures using hidden Markov support vector machine. BMC Bioinform 10:381
Liu X, Zhao L, Dong Q (2011) Protein remote homology detection based on auto-cross covariance transformation. Comput Biol Med 41:640–647
Liu B, Wang X, Chen Q, Dong Q, Lan X (2012) Using amino acid physicochemical distance transformation for fast protein remote homology detection. PLoS One 7:e46633
Liu B, Wang X, Zou Q, Dong Q, Chen Q (2013) Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation. Mol Inform 32:775–782
Liu B, Liu B, Liu F, Wang X (2014a) Protein binding site prediction by combining hidden Markov support vector machine and profile-based propensities. Sci World J 2014:464093
Liu B, Xu J, Lan X, Xu R, Zhou J, Wang X, Chou K-C (2014b) iDNA-Protldis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. PLoS One 9:e106691
Liu B, Xu J, Zou Q, Xu R, Wang X, Chen Q (2014c) Using distances between top-*n*-gram and residue pairs for protein remote homology detection. BMC Bioinform 15:S3
Liu B, Zhang D, Xu R, Xu J, Wang X, Chen Q, Dong Q, Chou K-C (2014d) Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics 30:472–479
Liu B, Fang L, Chen J, Liu F, Wang X (2015a) miRNA-dis: microRNA precursor identification based on distance structure status pairs. Mol BioSyst 11:1194–1204
Liu B, Fang L, Liu F, Wang X, Chen J, Chou K-C (2015b) Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS One 10:e0121501
Liu B, Fang L, Liu F, Wang X, Chou K-C (2015c) iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. J Biomol Struct Dyn. doi:10.1080/07391102.07392015.01014422
Liu B, Liu F, Fang L, Wang X, Chou K-C (2015d) repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics 31:1307–1309. doi:10.1093/bioinformatics/btu1820
Liu B, Xu J, Fan S, Xu R, Zhou J, Wang X (2015e) PseDNA-Pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation. Mol Inform 34:8–17
Lobley A, Sadowski MJ, Jones DT (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily fiscrimination. Bioinformatics 25:1761–1767
Ma J, Wang S, Wang Z, Xu J (2014) MRFalign: protein homology detection through alignment of Markov random fields. Res Comput Mol Biol 8394:173–174
Margelevicius M, Venclovas MLC (2010) COMA server for protein distant homology search. Bioinformatics 26:1905–1906
Mei S (2012a) Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization. J Theor Biol 293:121–130
Mei S (2012b) Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. J Theor Biol 310:80–87
Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept Lett 17:1207–1214
Mohabatkar H, Mohammad Beigi M, Esmaeili A (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, Beigi MM, Abdolahi K, Mohsenzadeh S (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, Behjati M, Mohabatkar H (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
Mondal S, Pai PP (2014) Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. J Theor Biol 356:30–35
Muda HM, Saad P, Othman RM (2011) Remote protein homology detection and fold recognition using two-layer support vector machine classifiers. Comput Biol Med 41:687–699

Nanni L, Lumini A (2008) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34:653–660

Nanni L, Lumini A, Gupta D, Garg A (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

Nanni L, Brahnam S, Lumini A (2014) Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. J Theor Biol 360C:109–116

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Noble WS, Kuang R, Leslie C, Weston J (2005) Identifying remote protein homologs by network propagation. FEBS J 272:5119–5128

Pearson K (1901) On lines and planes of closest fit to systems of points in space. Phil Mag 2:559–572

Qin YF, Zheng L, Huang J (2013) Locating apoptosis proteins by incorporating the signal peptide cleavage sites into the general form of Chou's pseudo amino acid composition. Int J Quantum Chem 113:1660–1667

Qiu WR, Xiao X, Chou KC (2014) iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int J Mol Sci 15:1746–1766

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

Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12:85–94

Såding J (2005) Protein homology detection by HMM–HMM comparison. Bioinformatics 21:951–960

Sadreyev RI, Tang M, Kim BH, Grishin NV (2009) COMPASS server for homology detection: improved statistical accuracy, speed and functionality. Nucleic Acids Res 37:W90–W94

Sahu SS, 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, Vert JP, Ueda N, Akutsu T (2004) Protein homology detection using string alignment kernels. Bioinformatics 20:1682–1689

Schäffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF (1999) Impala: matching a protein sequence against a collection of Psi-Blast-constructed position-specific score matrices. Bioinformatics 15:1000–1011

Sharma AK, Zhou GP, Kupferman J, Surks HK, Christensen EN, Chou JJ, Mendelsohn ME, Rigby AC (2008) Probing the interaction between the coiled coil leucine zipper of cGMP-dependent protein kinase Iα and the C terminus of the myosin binding subunit of the myosin light chain phosphatase. J Biol Chem 283:32860–32869

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

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Song L, Li D, Zeng X, Wu Y, Guo L, Zou Q (2014) nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. BMC Bioinformatics 15:298

Sun XY, Shi SP, Qiu JD, Suo SB, Huang SY, Liang RP (2012) Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. Mol BioSyst 8:3178–3184

Tomii K, Akiyama Y (2004) FORTE: a profile–profile comparison tool for protein fold recognition. Bioinformatics 20:594–595

Vapnik VN (1998) Statistical Learning Theory. Wiley-Interscience

Wan S, Mak MW, Kung SY (2013) GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition. J Theor Biol 323:40–48

Weston J, Elisseeff A, Zhou D, Leslie CS, Noble WS (2004) Protein ranking: from local to global structure in the protein similarity network. Proc Natl Acad Sci USA 101:6559–6563

Xie HL, Fu L, Nie XD (2013) Using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC. Protein Eng Des Sel 26:735–742

Xu Y, Ding J, Wu LY, Chou KC (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

Yang Y, Tantoso E, Li KB (2008) Remote protein homology detection using recurrence quantification analysis and amino acid physicochemical properties. J Theor Biol 252:145–154

Yu L, Guo Y, Li Y, Li G, Li M, Luo J, Xiong W, Qin W (2010) SecretP: identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. J Theor Biol 267:1–6

Zeng YH, Guo YZ, Xiao RQ, Yang L, Yu LZ, Li ML (2009) Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J Theor Biol 259:366–372

Zhang GY, Fang BS (2008) Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. J Theor Biol 253:310–315

Zhang SW, Chen W, Yang F, Pan Q (2008a) Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach. Amino Acids 35:591–598

Zhang SW, Zhang YL, Yang HF, Zhao CH, Pan Q (2008b) 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

Zhang J, Sun P, Zhao X, Ma Z (2014a) PECM: prediction of extracellular matrix proteins using the concept of Chou's pseudo amino acid composition. J Theor Biol 363:412–418

Zhang J, Zhao X, Sun P, Ma Z (2014b) PSNO: predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC. Int J Mol Sci 15:11204–11219

Zhang L, Zhao X, Kong L (2014c) Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition. J Theor Biol 355:105–110

Zhong WZ, Zhou SF (2014) Molecular science for drug development and biomedicine. Int J Mol Sci 15:20072–20078

Zhou GP (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 GP, Deng MH (1984) An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochem J 222:169

Zhou GP, Huang RB (2013) The pH-triggered conversion of the PrPc to PrPsc. Curr Top Med Chem 13:1152–1163

Zhou GP, Troy FA (2003) Characterization by NMR and molecular modeling of the binding of polyisoprenols and polyisoprenyl recognition sequence peptides: 3D structure of the complexes reveals sites of specific interactions. Glycobiology 13:51–71

Zhou GP, Troy FA (2005a) Invited review: NMR studies on how the binding complex of polyisoprenol recognition sequence peptides and polyisoprenols can modulate membrane structure. Curr Protein Pept 6:399–411

Zhou GP, Troy FA (2005b) NMR study of the preferred membrane orientation of polyisoprenols (dolichol) and the impact of their complex with polyisoprenyl recognition sequence peptides on membrane structure. Glycobiology 15:347–359

Zhou XB, Chen C, Li ZC, Zou XY (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 GP, Huang RB, Troy FA (2015) 3D structural conformation and functional domains of polysialyltransferase ST8Sia IV required for polysialylation of neural cell adhesion molecules. Protein Pept Lett 22:137–148

Zia Ur R, Khan A (2012) Identifying GPCRs and their types with Chou's pseudo amino acid composition: an approach from multi-scale energy representation and position specific scoring matrix. Protein Pept Lett 19:890–903

Zou D, He Z, He J, Xia Y (2011) Supersecondary structure prediction using Chou's pseudo amino acid composition. J Comput Chem 32:271–278

Zou Q, Li X, Jiang Y, Zhao Y, Wang G (2013) BinMemPredict: a web server and software for predicting membrane protein types. Curr Proteomics 10:2–9

Zuo YC, Peng Y, Liu L, Chen W, Yang L, Fan GL (2014) Predicting peroxidase subcellular location by hybridizing different descriptors of Chou's pseudo amino acid patterns. Anal Biochem 458:14–19