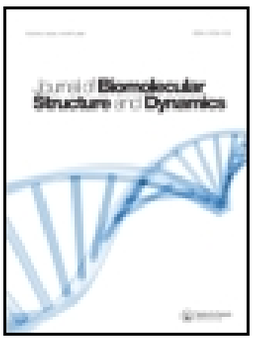

Taylor & Francis
Taylor & Francis Group

Journal of Biomolecular Structure and Dynamics

ISSN: 0739-1102 (Print) 1538-0254 (Online) Journal homepage: http://www.tandfonline.com/loi/tbsd20

Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition

Jianhua Jia, Zi Liu, Xuan Xiao, Bingxiang Liu & Kuo-Chen Chou

To cite this article: Jianhua Jia, Zi Liu, Xuan Xiao, Bingxiang Liu & Kuo-Chen Chou (2015): Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition, Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2015.1095116

To link to this article: http://dx.doi.org/10.1080/07391102.2015.1095116

Journal of Biomolecular Structure and Dynamics, 2015
http://dx.doi.org/10.1080/07391102.2015.1095116

Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition

Jianhua Jia[a], Zi Liu[a], Xuan Xiao[a,c]*, Bingxiang Liu[a] and Kuo-Chen Chou[b,c]

[a]*Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China;* [b]*Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia;* [c]*Gordon Life Science Institute, Boston, MA 02478, USA*

Communicated by Ramaswamy H. Sarma

(*Received 25 August 2015; accepted 14 September 2015*)

With the explosive growth of protein sequences entering into protein data banks in the post-genomic era, it is highly demanded to develop automated methods for rapidly and effectively identifying the protein–protein binding sites (PPBSs) based on the sequence information alone. To address this problem, we proposed a predictor called **iPPBS-PseAAC**, in which each amino acid residue site of the proteins concerned was treated as a 15-tuple peptide segment generated by sliding a window along the protein chains with its center aligned with the target residue. The working peptide segment is further formulated by a general form of pseudo amino acid composition via the following procedures: (1) it is converted into a numerical series via the physicochemical properties of amino acids; (2) the numerical series is subsequently converted into a 20-D feature vector by means of the stationary wavelet transform technique. Formed by many individual "Random Forest" classifiers, the operation engine to run prediction is a two-layer ensemble classifier, with the 1st-layer voting out the best training data-set from many bootstrap systems and the 2nd-layer voting out the most relevant one from seven physicochemical properties. Cross-validation tests indicate that the new predictor is very promising, meaning that many important key features, which are deeply hidden in complicated protein sequences, can be extracted via the wavelets transform approach, quite consistent with the facts that many important biological functions of proteins can be elucidated with their low-frequency internal motions. The web server of **iPPBS-PseAAC** is accessible at http://www.jci-bioinfo.cn/iPPBS-PseAAC, by which users can easily acquire their desired results without the need to follow the complicated mathematical equations involved.

Keywords: protein–protein binding sites; physicochemical property; stationary wavelet transform; pseudo amino acid composition; random forest; asymmetric bootstrap

1. Introduction

All cellular processes depend on precisely orchestrated interactions between proteins (Chou & Cai, 2006). A critical step in understanding the biological function of a protein is identification of the interface sites on which it interacts with other protein(s). Characterization of protein interactions is important for many problems covering from rational drug design to analysis of various biological networks (see, e.g. Fan, Xiao, & Min, 2014; Min, Xiao, & Chou, 2013; Xiao, Min, Lin, & Liu, 2015; Xiao, Min, & Wang, 2013a, 2013c; Zhong & Zhou, 2014; Zhou, 2015). The number of experimentally determined structures of protein–protein and protein–ligand complexes is still quite small, as reflected by the fact that the entries in UniprotKB/Swissprot (UniProt, 2013) is much larger than that in the Protein Data Bank (Berman et al., 2000). The limited availability of structures often restricts the identification of binding sites of proteins and their functional annotation. Furthermore, the chemical or biological experimental methods are expensive, time-consuming and labor-intensive. Therefore, as a complement to the experimental methods, it is highly demanded to develop computational methods for identifying the protein–protein binding sites (PPBSs) according to their sequences information alone (Gallet, Charloteaux, Thomas, & Brasseur, 2000; Valencia & Pazos, 2002).

Given a protein sequence, how can we identify which of its constituent amino acid residues are located in the binding site? Ofran and Rost (2003) and Yan, Dobbs, and Honavar (2004) have reported the following findings: (1) the residues involved in this kind of interactions usually tend to form clusters in sequences within four neighboring residues on either side; and (2) 97–98% of interface residues have at least one additional interface residue and 70–74% have at least four additional interface residues. Their analysis indicates that the neighboring residues of an actual interface residue have higher potential for being the interface residues,

*Corresponding author. Email: xxiao@gordonlifescience.org

suggesting that fragments of protein sequences (referred to as sub-sequences hereafter) may contain useful information or features for discriminating between interaction and non-interaction sites. Several approaches have been proposed for predicting protein–protein interaction sites from amino acid sequence. Kini and Evans (1996), based on their observations on the frequency of proline residues occurring near the interaction sites, proposed a method for predicting the potential PPBSs by detecting the presence of proline bracket. Shortly afterward, using the multiple sequence alignment to detect correlated changes of the interacting protein domains, Pazos, Helmer-Citterich, Ausiello, and Valencia (1997) offered a different method to predict the contacting residue pairs. In 2000, Gallet et al. (2000) introduced an approach to identify the interacting residues by analyzing the sequence hydrophobicity with the method developed by Eisenberg, Schwarz, Komaromy, and Wall (1984). In 2003, Ofran and Rost (2003) used sub-sequences of nine consecutive residues to develop a neural network-based method with a post-processing filter to predict interface residues. Subsequently, Yan et al. (2004) also used sub-sequence of nine residues to develop a two-stage classifier by combining support vector machine (SVM) and Bayesian network classifiers, achieving a higher accuracy. Two years later, Wang et al. (2006) also developed a predictor in this regard by using SVM with features extracted from spatial sequence and evolutionary scores based on a phylogenetic tree.

Since the three-dimensional (3D) structures are unknown for most of proteins, the sequence-based method plays an important role in protein binding site prediction. Unfortunately, several issues (Chen & Jeong, 2009; Sikic, Tomic, & Vlahovicek, 2009) exist that have made the sequence-based approach particularly difficult. The main problems are as follows: (i) the effective features common to all the binding sites are hard to extract because the biological properties responsible for protein–protein interacting are not fully understood; (ii) the prediction of binding sites is to deal with a highly imbalanced classification problem because the number of non-binding sites of a protein pair is substantially larger than that of binding ones, and hence prone to cause bias; (iii) there is no good benchmark data-set due to lack of a unique definition for the binding sites, as reflected by the fact that one definition of the binding sites is based on the distance between the carbon atoms concerned, but another on the change of the accessible surface area (ASA) value between the bounded and unbounded status.

The present study was initiated in an attempt to develop a new approach to predict the PPBSs in hope to help deal with the aforementioned problems.

As demonstrated in a series of recent publications (Ding, Deng, Yuan, & Liu, 2014; Jia, Liu, & Xiao, 2015; Liu, Fang, Liu, & Wang, 2015; Liu, Fang, Liu, Wang, & Chen, 2015; Liu, Fang, Wang, & Wang, 2015; Liu, Xu, Lan, Xu, & Zhou, 2014; Qiu, Xiao, & Lin, 2015; Xu, Wen, Wen, & Wu, 2014; Xu, Zhou, Liu, He, & Zou, 2015) in using Chou's 5-step rule (Chou, 2011), to develop a really useful sequence-based predictor for a biological system, we should make the following five procedures very clear: (1) how to construct or select a valid benchmark data-set to train and test the predictor; (2) how to formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (3) how to introduce or develop a powerful algorithm (or engine) to operate the prediction; (4) how to properly perform cross-validation tests to objectively evaluate its anticipated accuracy; (5) how to establish a user-friendly web-server that is accessible to the public. Below, we are to address the five procedures one-by-one.

2. Materials and methods

2.1. Benchmark data-set

Two benchmark data-sets were used for the current study. One is the "surface-residue" data-set and the other is "all-residue" data-set, as elaborated below.

The protein–protein interfaces are usually formed by those residues, which are exposed to the solvent after the two counterparts are separated from each other. Given a protein sample with L residues as expressed by

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L \tag{1}$$

where R_1 represents the 1st amino acid residue of the protein **P**, R2 the 2nd residue, and so forth. The residue $\mathrm{R}_i(i = 1, 2, \ldots, L)$ is deemed as a surface residue if it satisfies the following condition

$$\phi(\mathrm{R}_i) = \frac{\mathrm{ASA}(\mathrm{R}_i|\mathbf{P})}{\mathrm{ASA}(\mathrm{R}_i)} > 25\% \tag{2}$$

where $\mathrm{ASA}(\mathrm{R}_i|\mathbf{P})$ is the ASA of R_i when it is a part of protein **P**, $\mathrm{ASA}(\mathrm{R}_i)$ is the accessible surface area of the free R_i that is actually its maximal ASA as given in Table 1 (Ofran & Rost, 2003), and $\phi(\mathrm{R}_i)$ is the ratio of the two. Furthermore, the surface residue R_i is deemed as interfacial residue (Jones & Thornton, 1996) if

$$\mathrm{ASA}(\mathrm{R}_i|\mathbf{P}) - \mathrm{ASA}(\mathrm{R}_i|\mathbf{PP}) > 1\text{Å}^2 \tag{3}$$

where $\mathrm{ASA}(\mathrm{R}_i|\mathbf{PP})$ is the accessible surface area of R_i when it is a part of protein–protein complex.

For a given protein, we can use DSSP program (Kabsch & Sander, 1983) to find out all its surface residues based on Equation (2), and use PSAIA program (Mihel, Šikić, Tomić, Jeren, & Vlahovicek, 2008) to find all its interfacial residues based on Equation (3).

Table 1. Maximum ASA of different amino acids.[a]

AA	A	B	C	D	E	F	G	H	I	K	L	M
MaxASA	106	160	135	163	194	197	84	184	169	205	164	188
AA	N	P	Q	R	S	T	V	W	X	Y	Z	
MaxASA	157	136	198	248	130	142	142	227	180	222	196	

Note: B stands for D or N; Z for E or Q, and X for an undetermined amino acid.[a]Amino acids are represented by their one-letter codes.

If only considering the surface residues as done in Wang, Huang, and Jiang (2014) for the 99 polypeptide chains extracted by Deng, Guan, Dong, and Zhou (2009) from the 54 heterocomplexes in Protein Data Bank, we have obtained the results that can be formulated as follows:

$$\mathbb{S}_{\text{surf}} = \mathbb{S}^{+}_{\text{surf}} \cup \mathbb{S}^{-}_{\text{surf}} \tag{4}$$

where $\mathbb{S}_{\text{surf}}$is called the "surface-residue data-set" that contains a total of 13,771 surfaces residues, of which 2,828 are interfacial residues belonging to the positive subset $\mathbb{S}^{+}_{\text{surf}}$, while 10,943 are non-interfacial residues belonging the negative subset $\mathbb{S}^{-}_{\text{surf}}$, and $\cup$ is the symbol of union in the set theory.

If considering all the residues as done in Chen and Jeong (2009), however, the corresponding benchmark data-set can be expressed by

$$\mathbb{S}_{\text{all}} = \mathbb{S}^{+}_{\text{all}} \cup \mathbb{S}^{-}_{\text{all}} \tag{5}$$

where $\mathbb{S}_{\text{all}}$ is called the "all-residue data-set" that contains a total of 27,442 residues, of which 2828 are interfacial residues belonging to the positive subset $\mathbb{S}^{+}_{\text{all}}$, while 24,614 are non-interfacial residues belonging the negative subset $\mathbb{S}^{-}_{\text{all}}$.

For readers' convenience, given in S1 Data-set is a combination of the two benchmark data-sets, where those labeled in column 3 are all the residues determined by experiments, those in column 4 are of surface and non-surface residues, and those in column 5 are of interface and non-interface residues.

As pointed out in a comprehensive review (Chou & Shen, 2007a), there is no need to separate a benchmark data-set into a training data-set and a testing data-set for examining the quality of a prediction method if it is tested by the jackknife test or subsampling (K-fold) cross-validation test because the outcome thus obtained is actually from a combination of many different independent data-set tests.

2.2. *Flexible sliding window approach*

For a protein chain as formulated by Equation (1), the sliding window approach (Chou, 2001a) and flexible sliding window approach (Chou & Shen, 2007b) are often used to investigate its various post-translational modification (PTM) sites (see, e.g. Qiu, Xiao, & Lin, 2014; Qiu et al., 2015; Xu, Ding, & Wu, 2013; Xu, Shao, Wu, Deng, 2013; Xu, Wen, & Shao, 2014; Xu, Wen, Wen, et al., 2014) and HIV (human immunodeficiency virus) protease cleavage sites (Chou, 1996). Here, we also use it to study PPBSs. In the sliding window approach, a scaled window is denoted by $[-\xi, +\xi]$ (Chou, 2001a). Its width is $2\xi + 1$, where ξ is an integer. When sliding it along a protein chain **P** (Equation (1)), one can see through the window a series of consecutive peptide segments as formulated by

$$\mathbf{P}_{\xi}(\mathbb{R}_0) = \mathrm{R}_{-\xi}\mathrm{R}_{-(\xi-1)}\cdots\mathrm{R}_{-2}\mathrm{R}_{-1}\mathbb{R}_0\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+(\xi-1)}\mathrm{R}_{+\xi} \tag{6}$$

where $\mathrm{R}_{-\xi}$ represents the ξ-th upstream amino acid residue from the center, $\mathrm{R}_{+\xi}$ the ξ-th downstream amino acid residue, and so forth. The amino acid residue $\mathbb{R}_0$ at the center is the targeted residue. When its sequence position in **P** (cf. Equation (1)) is less than ξ or greater $L - \xi$, the corresponding $\mathbf{P}_{\xi}(\mathbb{R}_0)$ is defined, instead by **P** of Equation (1), but by the following dummy protein chain

$$\begin{aligned}\mathbf{P}(\text{dummy}) = {} & \mathbb{R}_{\xi}\cdots\mathbb{R}_2\mathbb{R}_1 \\ & \Updownarrow \mathrm{R}_1\mathrm{R}_2\cdots\mathrm{R}_{\xi}\cdots\mathrm{R}_i\cdots\mathrm{R}_{L-\xi+1}\cdots\mathrm{R}_{L-1}\mathrm{R}_L \\ & \Updownarrow \mathbb{R}_L\mathbb{R}_{L-1}\cdots\mathbb{R}_{L-\xi+1}\end{aligned} \tag{7}$$

where the symbol $\Updownarrow$ stands for a mirror, the dummy segment $\mathbb{R}_{\xi}\cdots\mathbb{R}_2\mathbb{R}_1$ stands for the image of $\mathrm{R}_1\mathrm{R}_2\cdots\mathrm{R}_{\xi}$ reflected by the mirror, and the dummy segment $\mathbb{R}_L\mathbb{R}_{L-1}\cdots\mathbb{R}_{L-\xi+1}$ for the mirror image of $\mathrm{R}_{L-\xi+1}\cdots\mathrm{R}_{L-1}\mathrm{R}_L$ (Figure 1). Accordingly, **P**(dummy) of Equation (7) is also called the mirror-extended chain of protein **P**.

Thus, for each of the L amino acid residues in protein **P** (Equation (1)), we have a working segment as defined by Equation (6). In the current study, the $(2\xi + 1)$-peptides $\mathbf{P}_{\xi}(\mathbb{R}_0)$ can be further classified into the following categories:

$$\mathbf{P}_{\xi}(\mathbb{R}_0) \in \begin{cases} \mathbf{P}^{+}_{\xi}(\mathbb{R}_0), & \text{if its center is a PPBS} \\ \mathbf{P}^{-}_{\xi}(\mathbb{R}_0), & \text{otherwise} \end{cases} \tag{8}$$

where $\in$ represents "a member of" in the set theory.

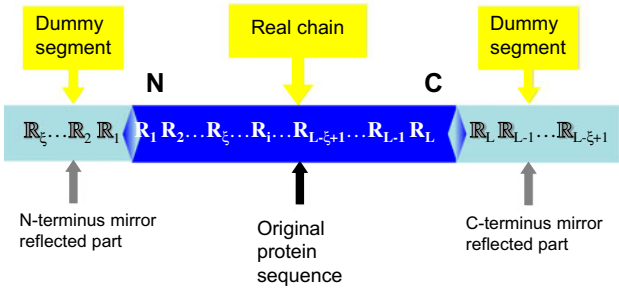

Figure 1. A schematic drawing to show how to use the extended chain of Equation (7) to define the working segments of Equation (6) for those sites when their sequence positions in the protein are less than ξ or greater $L - \xi$, where the left dummy segment stands for the mirror image of $\mathrm{R}_1\mathrm{R}_2 \cdots \mathrm{R}_\xi$ at N-terminus and the right dummy segment for that of $\mathrm{R}_{L-\xi+1} \cdots \mathrm{R}_{L-1}\mathrm{R}_L$ at the C-terminus.

2.3. Using pseudo amino acid composition to represent peptide chains

One of the most challenging problems in computational biology today is how to effectively formulate the sequence of a biological sample (such as protein, peptide, DNA, or RNA) with a discrete model or a vector that can considerably keep its sequence order information or capture its key features. The reasons are as follows. (1) If using the sequential model, i.e. the model in which all the samples are represented by their original sequences, it is hardly able to train a machine that can cover all the possible cases concerned, as elaborated in Chou (2011). (2) All the existing computational algorithms, such as optimization approach (Zhang & Chou, 1992), correlation-angle approach (Chou, 1993), covariance discriminant (Chou & Elrod, 2002), neural network (Feng, Cai, & Chou, 2005), K-nearest neighbor (KNN) (Shen, Yang, & Chou, 2006), OET-KNN (Shen & Chou, 2009b), SLLE algorithm (Wang, Yang, & Xu, 2005), SVM (Lin, Deng, Ding, & Chen, 2014; Xu et al., 2015), random forest (RF) (Lin, Fang, & Xiao, 2011), conditional random field (Xu, Ding, et al., 2013), nearest neighbor (Cai & Chou, 2003), Fuzzy KNN (Xiao, Min, & Wang, 2013b), and ML-KNN algorithm (Xiao, Wu, & Chou, 2011), can only handle vector but not sequence samples.

However, a vector defined in a discrete model may completely lose the sequence-order information. To cope with such a dilemma, the approach of pseudo amino acid composition (Chou, 2001c, 2005) or Chou's PseAAC (Cao, Xu, & Liang, 2013; Du, Wang, Xu, & Gao, 2012; Lin & Lapointe, 2013) was proposed. For a brief introduction about PseAAC, see a Wikipedia article at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition. Ever since it was introduced in 2001 (Chou, 2001c), the concept of PseAAC has been widely used to represent protein/peptide sequences in nearly all the areas of computational proteomics, such as in predicting protein subcellular location in various organisms and levels (Li et al., 2014; Mei, 2012; Nanni & Lumini, 2008; Shen, Yang, & Chou, 2007; Wang, Zhang, Zhang, & Li, 2015; Zhang, Zhang, Yang, Zhao, & Pan, 2008; Zuo et al., 2014), predicting protein structural class (Kong, Zhang, & Lv, 2014; Zhang, Zhao, & Kong, 2014), predicting membrane protein types (Chou & Cai, 2005; Han, Yu, & Anh, 2014), predicting antifreeze proteins (Mondal & Pai, 2014), predicting anticancer peptides (Hajisharifi, Piryaiee, Mohammad Beigi, Behbahani, & Mohabatkar, 2014), identifying bacterial virulent proteins (Nanni, Lumini, Gupta, & Garg, 2012), discriminating outer membrane proteins (Hayat & Khan, 2012), analyzing genetic sequence (Georgiou, Karakasidis, & Megaritis, 2013), identifying cyclin proteins (Mohabatkar, 2010), predicting GABA(A) receptor proteins (Mohabatkar, Mohammad Beigi, & Esmaeili, 2011), identifying antibacterial peptides (Khosravian, Faramarzi, Beigi, Behbahani, & Mohabatkar, 2013), identifying allergenic proteins (Mohabatkar, Beigi, Abdolahi, & Mohsenzadeh, 2013), predicting metalloproteinase family (Mohammad Beigi, Behjati, & Mohabatkar, 2011), identifying GPCRs and their types (Zia-ur-Rehman & Khan, 2012), identifying protein quaternary structural attributes (Chou & Cai, 2003; Sun et al., 2012), identifying risk type of human papillomaviruses (Esmaeili, Mohabatkar, & Mohsenzadeh, 2010), identifying various PTM (post-translational modification) sites in proteins (Jia, Lin, & Wang, 2014; Qiu, Xiao, & Lin, 2014; Qiu et al., 2015; Xu, Ding, et al., 2013; Xu, Shao, et al., 2013; Zhang, Zhao, Sun, & Ma, 2014), among many others (see a long list of references cited in Chen, Lin, and Chou [2015] published very recently). It has also been used in some disciplines of drug development and biomedicine (Zhong & Zhou, 2014) as well as drug target area (Chou, 2015). Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA/RNA sequences for studying various problems in computational genetics/genomics or performing genome analysis (see, e.g. (Chen, Feng, and Lin [2013]; Chen, Feng, and Lin [2014]; Chen, Lei, and Jin [2014]; Chen, Feng, Deng, and Lin [2014]; Chen, Feng, Ding, and Lin [2015]; Guo, Deng, Xu, and Ding [2014]; Lin et al. [2014]; Liu, Liu, Fang, and Wang [2015a, 2015b]; Liu, Liu, Wang, Chen, and Fang [2015]; Liu, Xiao, and Qiu [2015]; Qiu, Xiao, and Chou [2014] as well as a recent review Chen, Lin, et al. [2015]). Because it has been widely and increasingly used, five types of open access software, called "PseAAC" (Shen & Chou, 2008), "PseAAC-Builder" (Du et al., 2012), "propy" (Cao et al., 2013), "PseAAC-General" (Du, Gu, & Jiao, 2014), and "Pse-in-One" (Liu, Liu, Wang, et al., 2015) were established: the former three are for generating various modes of Chou's special PseAAC; the 4th one for those of Chou's general PseAAC; the 5th one can generate all the

pseudo components for protein/peptide as well as DNA/RNA sequences.

According to Chou (2011), PseAAC can be generally formulated as

$$\mathbf{P} = [\Psi_1 \Psi_2 \cdots \Psi_u \cdots \Psi_\Omega]^{\mathbf{T}} \tag{9}$$

where $\mathbf{T}$ is the transpose operator, while Ω an integer to reflect the vector's dimension. The value of Ω as well as the components $\Psi_u = (u = 1, 2, \cdots, \Omega)$ in Equation (9) will depend on how to extract the desired information from a peptide sequence. Below, we are to describe how to extract the useful information from the aforementioned benchmark data-sets (cf. Equations (4) and (5)) to define the working peptides via Equation (9). For the convenience of formulation below, we convert $(2\xi + 1)$-peptide in Equation (6) to

$$\mathbf{P}_\xi = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7 \cdots \mathrm{R}_{(2\xi+1)} \tag{10}$$

2.3.1. *Physicochemical properties*

Different types of amino acid in the above equation may have different physicochemical properties. In this study, we considered the following seven physicochemical properties: (1) hydrophobicity (Tanford, 1962) or $\Phi^{(1)}$; (2) hydrophicility (Hopp & Woods, 1981) or$\Phi^{(2)}$; (3) side-chain volume (Krigbaum & Knutton, 1973) or $\Phi^{(3)}$; (4) polarity (Grantham, 1974) or $\Phi^{(4)}$; (5) polarizability (Charton & Charton, 1982) or $\Phi^{(5)}$; (6) solvent-accessible surface area (SASA) (Rose, Geselowitz, Lesser, Lee, & Zehfus, 1985) or $\Phi^{(6)}$; and (7) side-chain net charge index (NCI) (Zhou, Tian, Li, Wu, & Li, 2006) or $\Phi^{(7)}$. Their numerical values are given in Table 2. Thus, the peptide segment $\mathbf{P}_\xi$ of Equation (10) can be encoded into seven different numerical series, as formulated by

$$\mathbf{P}_\xi = \begin{cases} \Phi_1^{(1)}\Phi_2^{(1)}\Phi_3^{(1)}\Phi_4^{(1)}\Phi_5^{(1)}\Phi_6^{(1)}\Phi_7^{(1)} \cdots \Phi_{2\xi+1}^{1} \\ \Phi_1^{(2)}\Phi_2^{(2)}\Phi_3^{(2)}\Phi_4^{(2)}\Phi_5^{(2)}\Phi_6^{(2)}\Phi_7^{(2)} \cdots \Phi_{2\xi+1}^{(2)} \\ \Phi_1^{(3)}\Phi_2^{(3)}\Phi_3^{(3)}\Phi_4^{(3)}\Phi_5^{(3)}\Phi_6^{(3)}\Phi_7^{(3)} \cdots \Phi_{2\xi+1}^{(3)} \\ \Phi_1^{(4)}\Phi_2^{(4)}\Phi_3^{(4)}\Phi_4^{(4)}\Phi_5^{(4)}\Phi_6^{(4)}\Phi_7^{(4)} \cdots \Phi_{2\xi+1}^{(4)} \\ \Phi_1^{(5)}\Phi_2^{(5)}\Phi_3^{(5)}\Phi_4^{(5)}\Phi_5^{(5)}\Phi_6^{(5)}\Phi_7^{(5)} \cdots \Phi_{2\xi+1}^{(5)} \\ \Phi_1^{(6)}\Phi_2^{(6)}\Phi_3^{(6)}\Phi_4^{(6)}\Phi_5^{(6)}\Phi_6^{(6)}\Phi_7^{(6)} \cdots \Phi_{2\xi+1}^{(6)} \\ \Phi_1^{(7)}\Phi_2^{(7)}\Phi_3^{(7)}\Phi_4^{(7)}\Phi_5^{(7)}\Phi_6^{(7)}\Phi_7^{(7)} \cdots \Phi_{2\xi+1}^{(7)} \end{cases} \tag{11}$$

where $\Phi_1^{(1)}$ is the hydrophobicity value of R_1 in Equation (9), $\Phi_2^{(2)}$ the hydrophilicity value of R_2, and so forth. Note that before substituting the physicochemical values of Table 2 into Equation (10), they all are subjected to the following standard conversion

$$\Phi_i^{(\varphi)} \Leftarrow \frac{\Phi_i^{\varphi} - \langle \Phi_i^{\varphi} \rangle}{\mathrm{SD}(\Phi_i^{\varphi})} (\varphi = 1, 2, \ldots, 7; i = 1, 2, \ldots, 2\xi + 1) \tag{12}$$

Table 2. The original values of the seven physicochemical properties for each amino acid.

	Physicochemical property (cf. Equation (11))[a]						
Amino acid code	$\Phi^{(1)}$ H1	$\Phi^{(2)}$ H2	$\Phi^{(3)}$ V	$\Phi^{(4)}$ P1	$\Phi^{(5)}$ P2	$\Phi^{(6)}$ SASA	$\Phi^{(7)}$ NCI
A	0.62	−0.5	27.5	8.1	0.046	1.181	0.007187
C	0.29	−1.0	44.6	5.5	0.128	1.461	−0.03661
D	−0.9	3.0	40.0	13.0	0.105	1.587	−0.02382
E	−0.74	3.0	62.0	12.3	0.151	1.862	0.006802
F	1.19	−2.5	115.5	5.2	0.29	2.228	0.037552
G	0.48	0.0	0.0	9.0	0.0	0.881	0.179052
H	−0.4	−0.5	79.0	10.4	0.23	2.025	−0.01069
I	1.38	−1.8	93.5	5.2	0.186	1.81	0.021631
K	−1.5	3.0	100.0	11.3	0.219	2.258	0.017708
L	1.06	−1.8	93.5	4.9	0.186	1.931	0.051672
M	0.64	−1.3	94.1	5.7	0.221	2.034	0.002683
N	−0.78	2.0	58.7	11.6	0.134	1.655	0.005392
P	0.12	0.0	41.9	8.0	0.131	1.468	0.239531
Q	−0.85	0.2	80.7	10.5	0.18	1.932	0.049211
R	−2.53	3.0	105.0	10.5	0.291	2.56	0.043587
S	−0.18	0.3	29.3	9.2	0.062	1.298	0.004627
T	−0.05	−0.4	51.3	8.6	0.108	1.525	0.003352
V	1.08	−1.5	71.5	5.9	0.14	1.645	0.057004
W	0.81	−3.4	145.5	5.4	0.409	2.663	0.037977
Y	0.26	−2.3	117.3	6.2	0.298	2.368	0.023599

[a]H1, hydrophobicity; H2, hydrophilicity; V, volume of side chains; P1, polarity; P2, polarizability; SASA, solvent accessible surface area; NCI, net charge index of side chains.

where the symbol $\langle\ \rangle$ means taking the average for the quantity therein over the 20 amino acid types, and SD means the corresponding standard deviation. The converted values via Equation (12) will have zero mean value over the 20 amino acid types, and will remain unchanged if they go thru the same standard conversion procedure again.

2.3.2. *Stationary wavelet transform approach*

The low-frequency internal motion is a very important feature of biomacromolecules (see, e.g. (Gordon, 2008; Madkan, Blank, Elson, Geddis, & Goodman, 2009; Martel, 1992) as well as a Wikipedia article at http://en.wikipedia.org/wiki/Low-frequency_collective_motion_in_proteins_and_DNA). Many marvelous biological functions in proteins and DNA and their profound dynamic mechanisms, such as switch between active and inactive states (Wang & Chou, 2009; Wang, Gong, Wei, & Li, 2009), cooperative effects (Chou, 1989a), allosteric transition (Chou, 1987; Schnell & Chou, 2008; Wang & Chou, 2010), intercalation of drugs into DNA (Chou & Mao, 1988), extra electron motion in DNA (Zhou, 1989), and assembly of microtubules (Chou, Zhang, & Maggiora, 1994), can be revealed by studying their low-frequency internal motions as summarized in a comprehensive review (Chou, 1988). Low-frequency Fourier spectrum was also used by Liu, Wang, and Chou (2005) to develop a sequence-based method for predicting membrane protein types. In view of this, it would be intriguing to introduce the stationary wavelet transform (SWT) into the current study.

The SWT (Shensa, 1992) is a wavelet transform algorithm designed to overcome the lack of shift-invariance of the discrete wavelet transform (DWT) (Mallat, 1989). Shift-invariance is achieved by removing the downsamplers and upsamplers in the DWT and upsampling (insert zero) the filter coefficients by a factor of 2^{j-1} in the jth level of the algorithm. The SWT is an inherently redundant scheme as the output of each level of SWT contains the same number of samples as the input; so for a decomposition of N levels, there is a redundancy of N in the wavelet coefficients. Shown in Figure 2 is the block diagram depicting the digital implementation of SWT. As we can see from the figure, the input peptide segment is decomposed recursively in the low-frequency part.

The concrete procedure of using the SWT to denote the $(2\xi+1)$-tuple peptides is as follows. For each of the $(2\xi+1)$-tuple peptides generated by sliding the scaled window $[-\xi,+\xi]$ along the protein chain concerned, the SWT was used to decompose it based on the amino acid values encoded by the seven physicochemical properties as given in Equation (11). Daubechies of number 1 (Db1) wavelet was selected because its wavelet possesses a lower vanish moment and easily generates non-zero

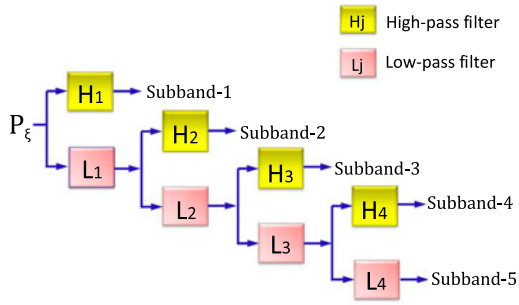

Figure 2. A schematic drawing to illustrate the procedure of multi-level SWT (stationary wavelets transform). See Equations (10)–(12) as well as the relevant text for further explanation. For more detailed explanation about SWT, see Nanson and Silverman (1995).

coefficients for the ensemble learning framework that will be introduced later.

In the preliminary study, we tested the sensitivity of the predicted outcome versus the value of parameter ξ from 4 to 10, and observed that when $\xi = 7$, i.e. the working segments are of 15-tuple peptides, the outcomes thus obtained were most promising, as shown in Figure 3. Accordingly, we only consider the case of $\xi = 7$ hereafter.

Using the SWT approach, we have generated 5 subbands (Figure 2), each of which has four coefficients: (1) α_i, the maximum of the wavelet coefficients in the subband $i(1,2,\ldots 5)$; (2) β_i, the corresponding mean of the wavelet coefficients; (3) γ_i, the corresponding minimum of the wavelet coefficients; (4) δ_i, the corresponding standard deviation of the wavelet coefficients. Therefore, for each working segment, we can get a feature vector that contains $\Omega = 5 \times 4 = 20$ components by using each of the seven physicochemical properties of Equation (11). In other words, we have seven different modes of PseAAC as given below:

$$\mathbf{P}^{(k)} = [\Psi_1^{(k)}\Psi_2^{(k)}\Psi_3^{(k)} \ \cdots\ \Psi_i^{(k)} \ \cdots\ \Psi_{20}^{(k)}]^{\mathbf{T}} \quad (k = 1, 2, \cdots, 7) \tag{13}$$

where

$$\Psi_u^{(k)} = \begin{cases} \alpha_u^{(k)} & \text{when} 1 \le u \le 5 \\ \beta_{u-5}^{(k)} & \text{when } 6 \le u \le 10 \\ \lambda_{u-10}^{(k)} & \text{when} 11 \le u \le 15 \\ \delta_{u-15}^{(k)} & \text{when} 11 \le u \le 20 \end{cases} \tag{14}$$

2.4. *Ensemble RF algorithm*

The RF algorithm is a powerful algorithm, which has been used in many areas of computational biology (see,

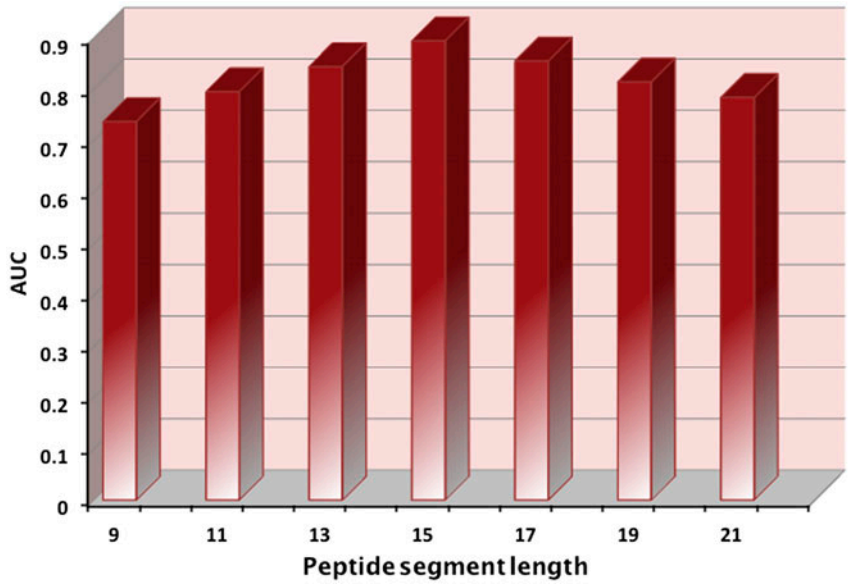

Figure 3. A histogram to show the results of AUC (Fawcett, 2005) obtained by using different values of ξ for the working peptides. As we can see, when $\xi = 7$, i.e. the working segments are of $(2\xi + 1) = 15$-tuple peptides (cf. Equation (10)), the outcomes thus obtained were most promising. For more explanation, see the text in the Section 2.2 and the legend of Figure 6 later.

e.g. Kandaswamy, Martinetz, Moller, Sridharan, & Pugalenthi, 2011; Lin et al., 2011; Pugalenthi, Kandaswamy, Vivekanandan, & Kolatkar, 2012). The detailed procedures and formulation of RF have been very clearly described in Breiman (2001), and hence there is no need to repeat here.

It should be pointed out, however, that the number of negative samples in the current case is much larger than that of positive ones, and most classifiers (including RF) are usually working properly for the benchmark data-sets consisting of balanced subsets. To deal with such a situation, an asymmetric bootstrap approach was adopted as elaborated in Jia, Xiao, Liu and Jiao (2011) and illustrated in Figure 4. As shown from the figure, in order to construct a balanced data-set to train each of the sub-classifiers, we randomly picked the negative training samples from $\mathbb{S}^-_{\text{all}}$ or $\mathbb{S}^-_{\text{surf}}$ making them have the same number of the corresponding positive samples in $\mathbb{S}^+_{\text{all}}$ or $\mathbb{S}^+_{\text{surf}}$, respectively

Also, as shown in Equation (13), a peptide segment concerned in the current study can be formulated with seven different PseAAC modes, each of which can be used to train the RF predictor. Accordingly, we have a total of seven individual predictors for identifying PPBS, as formulated by:

$$\text{PPBS individual predictor} = \mathbb{RF}(k) \quad k = 1, 2, \ldots, 7) \tag{15}$$

where $\mathbb{RF}(k)$ represents the RF predictor based on the k-th physicochemical property (cf. Equation (13)).

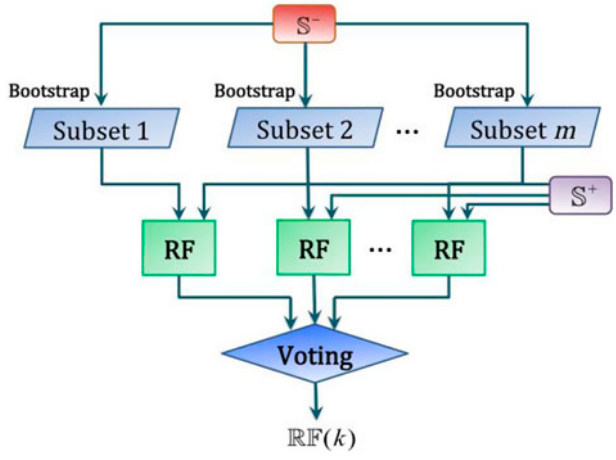

Figure 4. A flowchart to illustrate the 1st-layer ensemble classifier, a voting system by using the bootstrap approach to deal with the situation when the number of negative samples is overehelmingly larger than that of positive ones, as done in Jia et al. (2011). In the figure, RF denoted the RF classifier, $\mathbb{S}^+$ denotes either $\mathbb{S}^+_{\text{surf}}$ or $\mathbb{S}^+_{\text{all}}$, and $\mathbb{S}^-$ denotes either $\mathbb{S}^-_{\text{surf}}$ or $\mathbb{S}^-_{\text{all}}$ (cf. Equations (4)–(5)). See the text for more explanation.

Now, the problem is how to combine the results from the seven individual predictors to maximize the prediction quality. As indicated by a series of previous studies, using the ensemble classifier formed by fusing many individual classifiers can remarkably enhance the success rates in predicting protein subcellular localization (Chou & Shen, 2006, 2007c) and protein quaternary structural attribute (Shen & Chou, 2009a). Encouraged by the

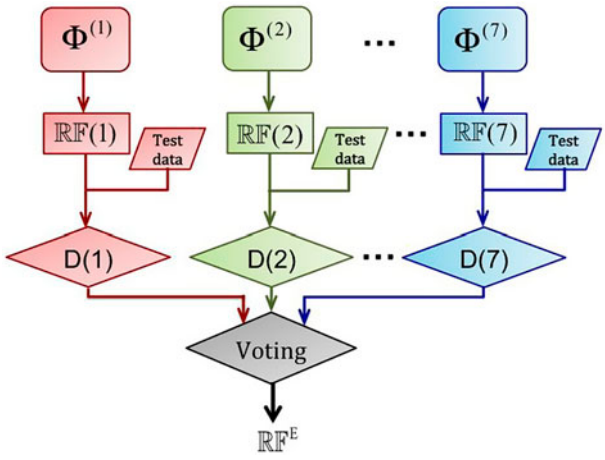

Figure 5. A flowchart to illustrate the 2nd-layer ensemble classifier that exploits all the different groups of features, where D(1) means the decision made by $\mathbb{RF}(1)$, D(2) means the decision made by $\mathbb{RF}(2)$, and so forth. See the text as well as Equations (11) and (15) for further explanation.

previous investigators' studies, here we are also developing an ensemble classifier by fusing the seven individual predictors $\mathbb{RF}(k)(k = 1, 2, \ldots, 7)$ through a voting system, as formulated by:

$$\mathbb{RF}^{\mathrm{E}} = \mathbb{RF}(1) \forall \cdots \forall \mathbb{RF}(7) = \forall_{k=1}^{7} \mathbb{RF}(k) \qquad (16)$$

where $\mathbb{RF}^{\mathrm{E}}$ stands for the ensemble classifier, and the symbol $\forall$ for the fusing operator. For the detailed procedures of how to fuse the results from the seven individual predictors to reach a final outcome via the voting system, see Equations (30)–(35) in Chou and Shen (2007a), where a crystal clear and elegant derivation was elaborated and hence there is no need to repeat here. To provide an intuitive picture, a flowchart is given in Figure 5 to illustrate how the seven individual RF predictors are fused into the ensemble classifier.

The final predictor thus obtained is called "**iPPBS-PseAAC**", where "i" stands for "identify", "PPBS" for "protein–protein binding site", and "PseAAC" for "pseudo amino acid composition" approach.

3. Result and discussion

As pointed out in the Introduction section, one of the important procedures in developing a predictor is how to properly and objectively evaluate its anticipated success rates (Chou, 2011). Toward this, we need to consider the following two aspects: one is what kind of metrics should be used to quantitatively measure the prediction accuracy; the other is what kind of test method should be adopted to derive the metrics values, as elaborated below.

3.1. Success rate metrics and validation approach

For measuring the success rates in identifying PPBS, a set of four metrics are often used in the literature. They are: (1) overall accuracy or Acc, (2) Mathew's correlation coefficient or MCC, (3) sensitivity or Sn, and (4) specificity or Sp (see, e.g. Chen, Liu, & Yang, 2007). Unfortunately, the conventional formulations for the four metrics are not quite intuitive for most experimental scientists, particularly the one for MCC. Interestingly, by using the symbols and derivation as used in Chou (2001a) for studying signal peptides, the aforementioned four metrics can be formulated by a set of equations given below (Chen et al., 2013; Lin et al., 2014; Qiu, Xiao, & Chou, 2014):

$$\begin{cases} Sn = 1 - \frac{N_-^+}{N^+} & 0 \le Sn \le 1 \\ Sp = 1 - \frac{N_+^-}{N^-} & 0 \le Sp \le 1 \\ Acc = \Lambda = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le Acc \le 1 \\ \mathrm{MCC} = \frac{1 - (\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-})}{\sqrt{(1 + \frac{N_+^- - N_-^+}{N^+})(1 + \frac{N_-^+ - N_+^-}{N^-})}} & -1 \le \mathrm{MCC} \le 1 \end{cases} \qquad (17)$$

where N^+ represents the total number of PPBSs investigated, whereas N_-^+ the number of true PPBSs incorrectly predicted to be of non-PPBS; N^- the total number of the non-PPBSs investigated, whereas N_+^- the number of non-PPBSs incorrectly predicted to be of PPBS.

According to Equation (17), it is crystal clear to see the following. When $N_-^+ = 0$ meaning none of the true PPBSs are incorrectly predicted to be of non-PPBS, we have the sensitivity $\mathrm{Sn} = 1$. When $N_-^+ = N^+$ meaning that all the PPBSs are incorrectly predicted to be of non-PPBS, we have the sensitivity $\mathrm{Sn} = 0$. Likewise, when $N_+^- = 0$ meaning none of the non-PPBSs are incorrectly predicted to be of PPBS, we have the specificity $\mathrm{Sp} = 1$; whereas $N_+^- = N^-$ meaning that all the non-PPBSs are incorrectly predicted to be of PPBS, we have the specificity $\mathrm{Sp} = 0$. When $N_-^+ = N_+^- = 0$ meaning that none of PPBSs in the positive data-set and none of the non-PPBSs in the negative data-set are incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = 1$ and $\mathrm{MCC} = 1$; when $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the PPBSs in the positive data-set and all the non-PPBSs in the negative data-set are incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = 0$ and $\mathrm{MCC} = -1$; whereas when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, we have $\mathrm{Acc} = 0.5$ and $\mathrm{MCC} = 0$ meaning no better than random guess. As we can see from the above discussion, it would make the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient much more intuitive and easier-to-understand by using Equation (17), particularly for the meaning of MCC.

It should be pointed out, however, the set of metrics as defined in Equation (17) is valid only for the

single-label systems. For the multi-label systems whose emergence has become more frequent in system biology (Chou, Wu, & Xiao, 2012; Lin, Fang, & Xiao, 2013; Xiao et al., 2011) and system medicine (Xiao, Wang, Lin, & Jia, 2013), a completely different set of metrics as defined in Chou (2013) is needed.

With the evaluation metrics available, the next thing is what validation method should be used to generate the metrics values.

In statistical prediction, the following three cross-validation methods are often used to derive the metrics values for predictor: independent data-set test, subsampling (or *K*-fold cross-validation) test, and jackknife test (Chou & Zhang, 1995). Of the three methods, however, the jackknife test is deemed the least arbitrary that can always yield a unique outcome for a given benchmark data-set as elucidated in Chou (2011) and demonstrated by Equations (28)–(32) therein. Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g. Chou & Cai, 2005; Dehzangi et al., 2015; Hajisharifi et al., 2014; Khan, Hayat, & Khan, 2015; Kumar, Srivastava, Kumari, & Kumar, 2015; Mondal & Pai, 2014; Shen et al., 2007; Xiao et al., 2011). However, to reduce the computational time, in this study we adopted the 10-fold cross-validation, as done by most investigators with SVM and RFs algorithms as the prediction engine. In the 10-fold cross-validation test, all the samples in the benchmark data-set are divided into 10 approximately equal-sized subsets. And then each of the 10 subsets will be singled out one-by-one and tested by the predictor trained with the samples in the remaining subsets. The performance measures are then calculated as an average over the 10 different single-out subsets or divisions. In other words, during the process of 10-fold cross-validation, both the training data-set and testing data-set are actually open, and each subset will be in turn moved between the two. The 10-fold cross-validation test can exclude the "memory" effect, just like conducting 10 different independent data-set tests.

3.2. *Comparison with the existing methods*

Listed in Table 3 are the values of the four metrics (cf. Equation (17)) obtained by the current **iPPBS-PseAAC** predictor using the 10-fold cross-validation on the surface-residue benchmark data-set $\mathbb{S}_{surf}$ (Equation (4)) and the all-residue benchmark data-set $\mathbb{S}_{all}$ (Equation (5)), respectively. See S1 Data-set for the details of the two benchmark data-sets. For facilitating comparison, the corresponding results obtained by the existing methods (Chen & Jeong, 2009; Deng et al., 2009) are also given there.

As we can see from the table, the new predictor **iPPBS-PseAAC** proposed in this paper remarkably outperformed its counterparts, particularly in Acc and MCC; the former stands for the overall accuracy, and the latter for the stability. At the first glance, although the value of Sn by Deng et al.'s method (Deng et al., 2009) is higher than that of the current predictor when tested by the surface-residue benchmark data-set, its corresponding Sp value is more than 30% lower than that of the latter, indicating the method (Deng et al., 2009) is very unstable with extremely high noise.

Because graphic approaches can provide useful intuitive insights (see, e.g. Althaus et al., 1993; Chou, 1989b, 2010; Chou & Forsen, 1980; Wu, Xiao, & Chou, 2010; Zhou, 2011), here we also provide a graphic comparison of the current predictor with their counterparts via the receiver operating characteristic (ROC) plot (Fawcett, 2005), as shown in Figure 6. According to ROC (Fawcett, 2005), the larger the area under the curve (AUC), the better the corresponding predictor is. As we can see from the figure, the area under the ROC curve of the new predictor is remarkably greater than those of their counterparts fully consistent with the AUC values listed on Table 3, once again indicating a clear improvement of the new predictor in comparison with the existing ones.

All the above facts have shown that **iPPBS-PseAAC** is really a very promising predictor for identifying PPBSs. Or at the very least, it can play a complementary

Table 3. Comparison of the **iPPBS-PseAAC** with the other existing methods via the 10-fold cross-validation on the surface-residue benchmark data-set (Equation (4)) and the all-residue benchmark data-set (Equation (5)).

Benchmark data-set	Method	Acc (%)	MCC	Sn (%)	Sp (%)	AUC
Surface-residue	Deng[a]	N/A	.3456	76.77	63.16	.7976
	Chen[b]	75.09	.4248	43.81	92.12	.8004
	iPPBS-PseAAC[c]	**84.90**	**.5862**	60.56	93.49	**.8828**
All-residue	Deng[a]	N/A	.3763	76.33	78.61	.8465
	Chen[b]	73.77	.3286	24.95	96.52	.8001
	iPPBS-PseAAC[c]	**84.88**	**.4554**	38.00	96.62	**.8709**

Note: Text in bold inicates the predictor proposed in this paper and its results.
[a]Results reported by Deng et al. (2009).
[b]Results reported by Chen and Jeong (2009).
[c]Results obtained by the current predictor using the same cross-validation method on the same benchmark data-set.

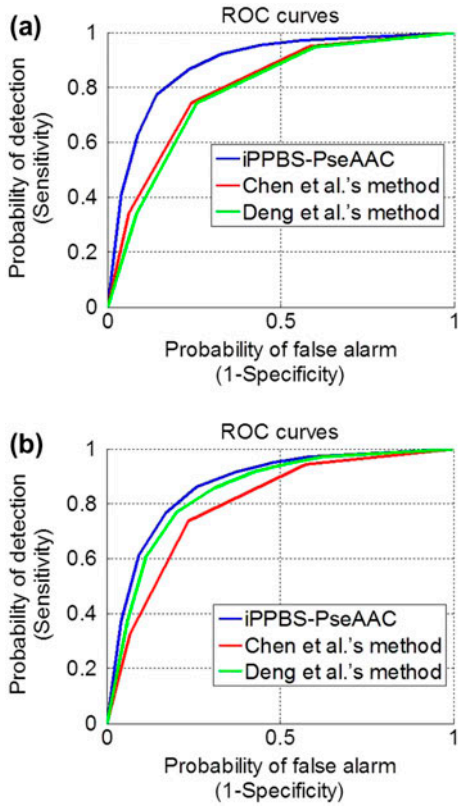

Figure 6. The ROC curves to show the 10-fold cross-validation by **iPPBS-PseAAC**, Deng et al.'s method (Deng et al., 2009), and Chen et al.'s method (Chen & Jeong, 2009) on (a) surface-residue benchmark data-set, and (b) the all-residue benchmark data-set. As shown on the figure, the area under the ROC curve for **iPPBS-PseAAC** is obviously larger than those of their counterparts, indicating a clear improvement of the new predictor in comparison with the existing ones.

role to the existing prediction methods in this area. Particularly, none of the existing predictors has provided a web server. In contrast to this, a user-friendly and publically accessible web server has been established for **iPPBS-PseAAC** at http://www.jci-bioinfo.cn/iPPBS-PseAAC, which is no doubt very useful for the majority of experimental scientist in this or related areas without the need to follow the complicated mathematical equations.

Why could the proposed method be so powerful? This is because many key features, which are deeply hidden in complicated protein sequences, can be extracted via the wavelets transform approach. Just like in dealing with the extremely complicated internal motions of proteins, it is the key to grasp the low-frequency collective motion (Gordon, 2008; Madkan et al., 2009) for in-depth understanding or revealing the dynamic mechanisms of their various important biological functions (Chou, 1988), such as cooperative effects (Chou, 1989a), allosteric transition (Chou, 1987; Schnell & Chou, 2008), assembly of microtubules (Chou et al., 1994), and switch between active and inactive states (Wang & Chou, 2009). Furthermore, a dual ensemble technique was used in this study: one for dealing the unbalanced training data-set via a bootstrap voting system (Figure 4), and one for selecting the most relevant one from seven classes of different physicochemical properties (Figure 5).

3.3. *Web server and user guide*

As emphasized in a recent review (Chou, 2015), an open accessible web-server is very important for the impact of a prediction method. To enhance the value of its practical applications, the web-server for iPPBI-PseAAC has been established at http://www.jci-bioinfo.cn/iPPBS-PseAAC. Furthermore, to maximize the convenience for the majority of experimental scientists, a step-to-step guide is provided below.

Step 1. Opening the web-server at http://www.jci-bioinfo.cn/iPPBD-PseAAC, you will see the top page of **iPPBS-PseAAC** on your computer screen, as shown in Figure 7. Click on the *Read Me* button to see a brief introduction about the **PPBS-PseAAC** predictor.

Step 2. Either type or copy/paste the query protein sequences into the input box at the center of Figure 7. The input sequence should be in the FASTA format. A sequence in FASTA format consists of a single initial line beginning with the symbol, >, in the first column, followed by lines of sequence data in which amino acids are represented using single-letter codes. Except for the mandatory symbol >, all the other characters in the single initial line are optional and only used for the purpose

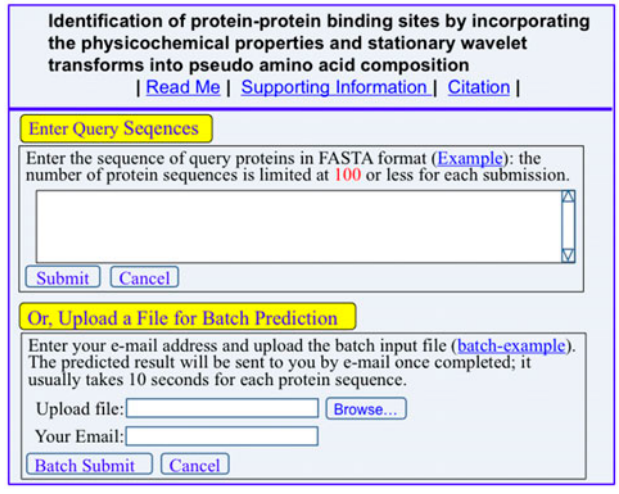

Figure 7. A semi-screenshot of the top page for the web server **iPPBS-PseAAC** at http://www.jci-bioinfo.cn/iPPBS-PseAAC.

of identification and description. The sequence ends if another line starting with the symbol > appears; this indicates the start of another sequence. For the examples of sequences in FASTA format, click the *Example* button right above the input box.

Step 3. Click on the *Submit* button to see the predicted result. For example, if you use the two query protein sequences in the *Example* window as the input, after 20 s or so, you will see the following on the screen of your computer: (1) Sequence-1 contains 63 amino acid residues, of which 19 are highlighted with red, meaning belonging to binding site. (2) Sequence-2 contains 224 residues, of which 11 are highlighted with red, belonging binding site. All these predicted results are fully consistent with experimental observations except for residues 44 and 63 in sequence-1 and residues 52 in sequence-2 that are overpredicted.

Step 4. As shown on the lower panel of Figure 7, you may also choose the batch prediction by entering your email address and your desired batch input file (in FASTA format of course) via the "Browse" button. To see the sample of batch input file, click on the button Batch-example.

Step 5. Click on the *Citation* button to find the relevant papers that document the detailed development and algorithm of **iPPBS-PseAAC**.

Step 6. Click the *Supporting Information* button to download the benchmark data-set used in this study.

4. Conclusion

In the new PPBS predictor, each of the protein residue sites investigated is treated as a 15-tuple peptide generated by sliding the scaled window [−7,+7] (Chou, 2001b) along a protein chain with its center aligned with the amino acid residue concerned. The working peptide segment is further formulated by a general form of PseAAC via the following procedures: (1) it is converted into a numerical series via the physicochemical properties of amino acids; (2) the numerical series is subsequently converted into a 20-D feature vector by means of the SWT technique.

The operation engine to run the PPBS prediction is a dual ensemble formed by two voting systems with one for finding the best training data-set and the other for finding the most relevant physicochemical property.

It was demonstrated via cross-validations that the new predictor established with the above procedures is very powerful and promising. We anticipate that **iPPBS-PseAAC** predictor will become a very useful high throughput tool for identifying PPBSs, or at the very least, a complementary tool to the existing prediction methods in this area.

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.org/10.1080/07391102.2015.1095116.

Acknowledgments

The authors wish to thank the two anonymous reviewers for their constructive comments, which are very helpful for strengthening the presentation of this study.

Disclosure statement

The authors declare no conflict of interest.

Funding

This work was partially supported by the National Nature Science Foundation of China [grant number 61261027], [grant number 61262038], [grant number 31260273], [grant number 61202313], [grant number 31560316]; the Natural Science Foundation of Jiangxi Province, China [grant number 20122BAB211033], [grant number 20122BAB201044], [grant number 20132BAB201053]; the Scientific Research plan of the Department of Education of JiangXi Province [GJJ14640]; The Young Teacher Development Plan of Visiting Scholars Program in the University of Jiangxi Province. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Althaus, I. W., Gonzales, A. J., Diebel, M. R., Kezdy, F. J., Aristoff, P. A., Tarpley, W. G., & Reusser, F. (1993). Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-88204E. *Biochemistry, 32*, 6548–6554. doi:10.1021/bi00077a008

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research, 28*, 235–242. doi:10.1093/nar/28.1.235

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. doi:10.1023/A:1010933404324

Cai, Y. D. & Chou, K. C. (2003). Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. *Biochemical and Biophysical Research Communications, 305*, 407–411. doi:10.1016/S0006-291X(03)00775-7

Cao, D. S., Xu, Q. S., & Liang, Y. Z. (2013). propy: A tool to generate various modes of Chou's PseAAC. *Bioinformatics, 29*, 960–962. doi:10.1093/bioinformatics/btt072

Charton, M., & Charton, B. I. (1982). The structural dependence of amino acid hydrophobicity parameters. *Journal of Theoretical Biology, 99*, 629–644. doi:10.1016/0022-5193(82)90191-6

Chen, J., Liu, H., & Yang, J. (2007). Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids, 33*, 423–428. doi:10.1007/s00726-006-0485-9

Chen, W., Feng, P. M., Deng, E. Z., & Lin, H. (2014). iTIS-PseTNC: A sequence-based predictor for identifying translation initiation site in human genes using pseudo

trinucleotide composition. *Analytical Biochemistry, 462*, 76–83. doi:10.1016/j.ab.2014.06.022

Chen, W., Feng, P. M., & Lin, H. (2013). iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Research, 41*, e68. doi:10.1093/nar/gks1450

Chen, W., Feng, P., Ding, H., & Lin, H. (2015). iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. *Analytical Biochemistry, 490*, 26–33.

Chen, W., Feng, P. M., & Lin, H. (2014). iSS-PseDNC: Identifying splicing sites using pseudo dinucleotide composition. *Biomed Research International (BMRI), 2014*, 623149. doi:10.1155/2014/623149

Chen, W., Lei, T. Y., & Jin, D. C. (2014). PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition. *Analytical Biochemistry, 456*, 53–60. doi:10.1016/j.ab.2014.04.001

Chen, W., Lin, H., & Chou, K. C. (2015). Pseudo nucleotide composition or PseKNC: An effective formulation for analyzing genomic sequences. *Molecular BioSystems, 11*, 2620–2634.

Chen, X. W., & Jeong, J. C. (2009). Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics, 25*, 585–591. doi:10.1093/bioinformatics/btp039

Chou, J. J. (1993). Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach. *Journal of Protein Chemistry, 12*, 291–302. doi:10.1007/BF01028191

Chou, K. C. (1987). The biological functions of low-frequency vibrations (phonons). VI. A possible dynamic mechanism of allosteric transition in antibody molecules. *Biopolymers, 26*, 285–295. doi:10.1002/bip.360260209

Chou, K. C. (1988). Low-frequency collective motion in biomacromolecules and its biological functions. *Biophysical Chemistry, 30*, 3–48. doi:10.1016/0301-4622(88)85002-6

Chou, K. C. (1989a). Low-frequency resonance and cooperativity of hemoglobin. *Trends in Biochemical Sciences, 14*, 212–213. doi:10.1016/0968-0004(89)90026-1

Chou, K. C. (1989b). Graphic rules in steady and non-steady enzyme kinetics. *Journal of Biological Chemistry, 264*, 12074–12079. Accession Number: 2745429

Chou, K. C. (1996). Prediction of human immunodeficiency virus protease cleavage sites in proteins. *Analytical Biochemistry, 233*, 1–14. doi:10.1006/abio.1996.0001

Chou, K. C. (2001a). Using subsite coupling to predict signal peptides. *Protein Engineering Design and Selection, 14*, 75–79. doi:10.1093/protein/14.2.75

Chou, K. C. (2001b). Prediction of signal peptides using scaled window. *Peptides, 22*, 1973–1979. doi:10.1016/S0196-9781(01)00540-X

Chou, K. C. (2001c). Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins: Structure, Function, and Genetics,* (Erratum: ibid., 2001, Vol. 44, 60) *43*, 246–255. doi:10.1002/prot.1035

Chou, K. C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics, 21*, 10–19. doi:10.1093/bioinformatics/bth466

Chou, K. C. (2010). Graphic rule for drug metabolism systems. *Current Drug Metabolism, 11*, 369–378. doi:10.2174/138920010791514261

Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology, 273*, 236–247. doi:10.1016/j.jtbi.2010.12.024

Chou, K. C. (2013). Some remarks on predicting multi-label attributes in molecular biosystems. *Molecular BioSystems, 9*, 1092–1100. doi:10.1039/C3MB25555G

Chou, K. C. (2015). Impacts of bioinformatics to medicinal chemistry. *Medicinal Chemistry, 11*, 218–234. doi:10.2174/1573406411666141229162834

Chou, K. C., & Cai, Y. D. (2003). Predicting protein quaternary structure by pseudo amino acid composition. *Proteins: Structure, Function, and Bioinformatics, 53*, 282–289. doi:10.1002/prot.10500

Chou, K. C., & Cai, Y. D. (2005). Prediction of membrane protein types by incorporating amphipathic effects. *Journal of Chemical Information and Modeling, 45*, 407–413. doi:10.1021/ci049686v

Chou, K. C., & Cai, Y. D. (2006). Predicting protein−protein interactions from sequences in a hybridization space. *Journal of Proteome Research, 5*, 316–322. doi:10.1021/pr050331g

Chou, K. C., & Elrod, D. W. (2002). Bioinformatical analysis of G-protein-coupled receptors. *Journal of Proteome Research, 1*, 429–433. doi:10.1021/pr025527k

Chou, K. C., & Forsén, S. (1980). Graphical rules for enzyme-catalysed rate laws. *Biochemical Journal, 187*, 829–835. doi:10.1042/bj1870829

Chou, K. C., & Mao, B. (1988). Collective motion in DNA and its role in drug intercalation. *Biopolymers, 27*, 1795–1815. doi:10.1002/bip.360271109

Chou, K. C., & Shen, H. B. (2006). Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *Journal of Proteome Research, 5*, 1888–1897. doi:10.1021/pr060167c

Chou, K. C., & Shen, H. B. (2007a). Euk-mPLoc: A fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *Journal of Proteome Research, 6*, 1728–1734. doi:10.1021/pr060635

Chou, K. C., & Shen, H. B. (2007b). Recent progress in protein subcellular location prediction. *Analytical Biochemistry, 370*, 1–16. doi:10.1016/j.ab.2007.07.006

Chou, K. C., & Shen, H. B. (2007c). Signal-CF: A subsite-coupled and window-fusing approach for predicting signal peptides. *Biochemical and Biophysical Research Communications, 357*, 633–640. doi:10.1016/j.bbrc.2007.03.162

Chou, K. C., Wu, Z. C., & Xiao, X. (2012). iLoc-Hum: Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Molecular Biosystems, 8*, 629–641. doi:10.1039/C1MB05420A

Chou, K. C., & Zhang, C. T. (1995). Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology, 30*, 275–349. doi:10.3109/10409239509083488

Chou, K. C., Zhang, C. T., & Maggiora, G. M. (1994). Solitary wave dynamics as a mechanism for explaining the internal motion during microtubule growth. *Biopolymers, 34*, 143–153. doi:10.1002/bip.360340114

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. *Journal of Theoretical Biology, 364*, 284–294. doi:10.1039/C1MB05420A

Deng, L., Guan, J., Dong, Q., & Zhou, S. (2009). Prediction of protein–protein interaction sites using an ensemble method. *BMC Bioinformatics, 10*, 426. doi:10.1186/1471-2105-10-426

Ding, H., Deng, E. Z., Yuan, L. F., & Liu, L. (2014). iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed Research International (BMRI), 2014*, 286419. doi:10.1155/2014/286419

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. *International Journal of Molecular Sciences, 15*, 3495–3506. doi:10.3390/ijms15033495

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. *Analytical Biochemistry, 425*, 117–119. doi:10.1016/j.ab.2012.03.015

Eisenberg, D., Schwarz, E., Komaromy, M., & Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *Journal of Molecular Biology, 179*, 125–142. doi:10.1016/0022-2836(84)90309-7

Esmaeili, M., Mohabatkar, H., & Mohsenzadeh, S. (2010). Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *Journal of Theoretical Biology, 263*, 203–209. doi:10.1016/j.jtbi.2009.11.016

Fan, Y. N., Xiao, X., & Min, J. L. (2014). iNR-Drug: Predicting the interaction of drugs with nuclear receptors in cellular networking. *International Journal of Molecular Sciences, 15*, 4915–4937. doi:10.3390/ijms15034915

Fawcett, J. A. (2005). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861–874. doi:10.1016/j.patrec.2005.10.010

Feng, K. Y., Cai, Y. D., & Chou, K. C. (2005). Boosting classifier for predicting protein domain structural class. *Biochemical & Biophysical Research Communications (BBRC), 334*, 213–217.

Gallet, X., Charloteaux, B., Thomas, A., & Brasseur, R. (2000). A fast method to predict protein interaction sites from sequences. *Journal of Molecular Biology, 302*, 917–926. doi:10.1006/jmbi.2000.4092

Georgiou, D. N., Karakasidis, T. E., & Megaritis, A. C. (2013). A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory. *The Open Bioinformatics Journal, 7*, 41–48. doi:10.2174/1875036201307010041

Gordon, G. (2008). Extrinsic electromagnetic fields, low frequency (phonon) vibrations, and control of cell function: A non-linear resonance system. *Journal of Biomedical Science and Engineering, 1*, 152–156. doi:10.4236/jbise.2008.13025

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science, 185*, 862–864. doi:10.1126/science.185.4154.862

Guo, S. H., Deng, E. Z., Xu, L. Q., & Ding, H. (2014). iNuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics, 30*, 1522–1529. doi:10.1093/bioinformatics/btu083

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. *Journal of Theoretical Biology, 341*, 34–40. doi:10.1016/j.jtbi.2013.08.037

Han, G. S., Yu, Z. G., & 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. *Journal of Theoretical Biology, 344*, 31–39. doi:10.1016/j.jtbi.2013.11.017

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 & Peptide Letters, 19*, 411–421. doi:10.2174/092986612799789387

Hopp, T. P., & Woods, K. R. (1981). Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences, 78*, 3824–3828. doi:10.1073/pnas.78.6.3824

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. *International Journal of Molecular Sciences, 15*, 10410–10423. doi:10.3390/ijms150610410

Jia, J., Liu, Z., & Xiao, X. (2015). iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. *Journal of Theoretical Biology, 377*, 47–56. doi:10.1016/j.jtbi.2015.04.011

Jia, J., Xiao, X., Liu, B., & Jiao, L. (2011). Bagging-based spectral clustering ensemble selection. *Pattern Recognition Letters, 32*, 1456–1467. doi:10.1016/j.patrec.2011.04.008

Jones, S., & Thornton, J. M. (1996). Principles of protein–protein interactions. *Proceedings of the National Academy of Sciences, 93*, 13–20. doi:10.1073/pnas.93.1.13

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers, 22*, 2577–2637. doi:10.1002/bip.360221211

Kandaswamy, K. K., Martinetz, T., Moller, S., Sridharan, S., & Pugalenthi (2011). AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *Journal of Theoretical Biology, 270*, 56–62. doi:10.1016/j.jtbi.2010.10.037

Khan, Z. U., Hayat, M., & Khan, M. A. (2015). Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. *Journal of Theoretical Biology, 365*, 197–203. doi:10.1016/j.jtbi.2014.10.014

Khosravian, M., Faramarzi, F. K., Beigi, M. M., Behbahani, M., & Mohabatkar, H. (2013). Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. *Protein & Peptide Letters, 20*, 180–186. doi:10.2174/092986613804725307

Kini, R. M., & Evans, H. J. (1996). Prediction of potential protein–protein interaction sites from amino acid sequence. *FEBS Letters, 385*, 81–86. doi:10.1016/0014-5793(96)00327-4

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. *Journal of Theoretical Biology, 344*, 12–18. doi:10.1016/j.jtbi.2013.11.021

Krigbaum, W. R., & Knutton, S. P. (1973). Prediction of the amount of secondary structure in a globular protein from its aminoacid composition. *Proceedings of the National Academy of Sciences, 70*, 2809–2813. doi:10.1073/pnas.70.10.2809

Kumar, R., Srivastava, A., Kumari, B., & Kumar, M. (2015). Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. *Journal of Theoretical Biology, 365*, 96–103. doi:10.1016/j.jtbi.2014.10.008

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. doi:10.1016/j.biochi.2014.06.001

Lin, H., Deng, E. Z., Ding, H., & Chen, W. (2014). iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Research, 42*, 12961–12972. doi:10.1093/nar/gku1019

Lin, S. X., & Lapointe, J. (2013). Theoretical and experimental biology in one. *Journal of Biomedical Science and Engineering, 6*, 435–442. doi:10.4236/jbise.2013.64054

Lin, W. Z., Fang, J. A., & Xiao, X. (2011). iDNA-Prot: Identification of DNA Binding Proteins using Random Forest with Grey Model. *PLoS ONE, 6*, e24756. doi:10.1371/journal.pone.0024756

Lin, W. Z., Fang, J. A., & Xiao, X. (2013). iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. *Molecular BioSystems, 9*, 634–644. doi:10.1039/C3MB25466F

Liu, B., Fang, L., Liu, F., & Wang, X. (2015). iMiRNA-PseDPC: MicroRNA precursor identification with a pseudo distance-pair composition approach. *Journal of Biomolecular Structure & Dynamics (JBSD)*. doi:10.1080/07391102.2015.1014422

Liu, B., Fang, L., Liu, F., Wang, X., & Chen, J. (2015). Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS ONE, 10*, e0121501. doi:10.1371/journal.pone.0121501

Liu, B., Fang, L., Wang, S., & Wang, X. (2015). Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. *Journal of Theoretical Biology, 385*, 153–159.

Liu, B., Liu, F., Fang, L., & Wang, X. (2015a). 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/btu820

Liu, B., Liu, F., Fang, L., & Wang, X. (2015b). repRNA: A web server for generating various feature vectors of RNA sequences. *Molecular Genetics and Genomics*. doi:10.1007/s00438-015-1078-7

Liu, B., Liu, F., Wang, X., Chen, J., & Fang, L. (2015). Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Research, 43*, W65–W71. doi:10.1093/nar/gkv458

Liu, B., Xu, J., Lan, X., Xu, R., & Zhou, J. (2014). iDNA-Prot|dis: 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. doi:10.1371/journal.pone.0106691

Liu, H., Wang, M., Chou, K. C. (2005). Low-frequency Fourier spectrum for predicting membrane protein types. *Biochemical and Biophysical Research Communications, 336*, 737-739. Accession Number: 16140260

Liu, Z., Xiao, X., Qiu, W. R. (2015). iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. *Analytical Biochemistry, 474*, 69–77. (also Data in Brief, 2015, 4, 87-89). doi:10.1016/j.ab.2014.12.009

Madkan, A., Blank, M., Elson, E., Geddis, M. S., & Goodman, R. (2009). Steps to the clinic with ELF EMF. *Natural Science, 1*, 157–165. doi:10.4236/ns.2009.13020

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*, 674–693. doi:10.1109/34.192463

Martel, P. (1992). Biophysical aspects of neutron scattering from vibrational modes of proteins. *Progress in Biophysics and Molecular Biology, 57*, 129–179. doi:10.1016/0079-6107(92)90023-Y

Mei, S. (2012). Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. *Journal of Theoretical Biology, 310*, 80–87. doi:10.1016/j.jtbi.2012.06.028

Mihel, J., Šikić, M., Tomić, S., Jeren, B., & Vlahovicek, K. (2008). PSAIA – Protein structure and interaction analyzer. *BMC Structural Biology, 8*, 21. doi:10.1186/1472-6807-8-21

Min, J. L., Xiao, X., & Chou, K. C. (2013). iEzy-Drug: A web server for identifying the interaction between enzymes and drugs in cellular networking. *BioMed Research International (BMRI), 2013*, 701317. doi:10.1155/2013/701317

Mohabatkar, H. (2010). Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein & Peptide Letters, 17*, 1207–1214. doi:10.2174/092986610792231564

Mohabatkar, H., Beigi, M. M., 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. *Medicinal Chemistry, 9*, 133–137. doi:10.2174/1573406411309010133

Mohabatkar, H., Mohammad Beigi, M., & Esmaeili, A. (2011). Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *Journal of Theoretical Biology, 281*, 18–23. doi:10.1016/j.jtbi.2011.04.017

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. *Journal of Structural and Functional Genomics, 12*, 191–197. doi:10.1007/s10969-011-9120-4

Mondal, S., & Pai, P. P. (2014). Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. *Journal of Theoretical Biology, 356*, 30–35. doi:10.1016/j.jtbi.2014.04.006

Nanni, L., & Lumini, A. (2008). Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids, 34*, 653–660. doi:10.1007/s00726-007-0018-1

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 Transactions on Computational Biology and Bioinformatics, 9*, 467–475. doi:10.1109/TCBB.2011.117

Nason, G. P., & Silverman, B. W. (1995). The stationary wavelet transform and some statistical application. *Lecture Notes in Statistics, 103*, 281–299. doi:10.1007/978-1-4612-2544-7_17

Ofran, Y., & Rost, B. (2003). Predicted protein–protein interaction sites from local sequence information. *FEBS Letters, 544*, 236–239. doi:10.1016/S0014-5793(03)00456-3

Pazos, F., Helmer-Citterich, M., Ausiello, G., & Valencia, A. (1997). Correlated mutations contain information about protein–protein interaction. *Journal of Molecular Biology, 271*, 511–523. doi:10.1006/jmbi.1997.1198

Pugalenthi, G., Kandaswamy, K. K., Vivekanandan, S., & Kolatkar, P. (2012). RSARF: Prediction of residue solvent accessibility from protein sequence using random forest method. *Protein & Peptide Letters, 19*, 50–56. doi:10.2174/092986612798472875

Qiu, W. R., Xiao, X., & Chou, K. C. (2014). iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. *International Journal of Molecular Sciences, 15*, 1746–1766. doi:10.3390/ijms15021746

Qiu, W. R., Xiao, X., & Lin, W. Z. (2014). iMethyl-PseAAC: Identification of protein methylation sites via a pseudo amino acid composition approach. *Biomed Res Int (BMRI), 2014*, 947416. doi:10.1155/2014/947416

Qiu, W. R., Xiao, X., Lin, W. Z. (2015). iUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model *Journal of Biomolecular Structure and Dynamics, 33*, 1731-1742. doi:10.1080/07391102.2014.968875

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., & Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science, 229*, 834–838. doi:10.1126/science.4023714

Schnell, J. R., & Chou, J. J. (2008). Structure and mechanism of the M2 proton channel of influenza A virus. *Nature, 451*, 591–595. doi:10.1038/nature06531

Shen, H. B., & Chou, K. C. (2008). PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. *Analytical Biochemistry, 373*, 386–388. doi:10.1016/j.ab.2007.10.012

Shen, H. B., & Chou, K. C. (2009a). QuatIdent: A web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. *Journal of Proteome Research, 8*, 1577–1584. doi:10.1021/pr800957q

Shen, H. B., & Chou, K. C. (2009b). A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Analytical Biochemistry, 394*, 269–274. doi:10.1016/j.ab.2009.07.046

Shen, H. B., Yang, J., & Chou, K. C. (2006). Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. *Journal of Theoretical Biology, 240*, 9–13. doi:10.1016/j.jtbi.2005.08.016

Shen, H. B., Yang, J., & Chou, K. C. (2007). Euk-PLoc: An ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids, 33*, 57–67. doi:10.1007/s00726-006-0478-8

Shensa, M. (1992). The discrete wavelet transform: Wedding the a trous and Mallat algorithms. *IEEE Transactions on Signal Processing, 40*, 2464–2482. doi:10.1109/78.157290

Sikic, M., Tomic, S., & Vlahovicek, K. (2009). Prediction of protein–protein interaction sites in sequences and 3D structures by random forests. *PLoS Computational Biology, 5*, e1000278. doi:10.1371/journal.pcbi.1000278

Sun, X. Y., Shi, S. P., Qiu, J. D., Suo, S. B., Huang, S. Y., & Liang, R. P. (2012). Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Molecular BioSystems, 8*, 3178–3184. doi:10.1039/c2mb25280e

Tanford, C. (1962). Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *Journal of the American Chemical Society, 84*, 4240–4247. doi:10.1021/ja00881a009

UniProt, C. (2013). Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research, 41*, D43–D47. doi:10.1093/nar/gks1068

Valencia, A., & Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology, 12*, 368–373. doi:10.1016/S0959-440X(02)00333-0

Wang, B., Chen, P., Huang, D. S., Li, J. J., Lok, T. M., & Lyu, M. R. (2006). Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Letters, 580*, 380–384. doi:10.1109/TNB.2014.2316997

Wang, B., Huang, D. S., & Jiang, C. (2014). A new strategy for protein interface identification using manifold learning method. *IEEE Transactions on Nanobioscience, 13*, 118–123. doi:10.1109/TNB.2014.2316997

Wang, J. F., & Chou, K. C. (2009). Insight into the molecular switch mechanism of human Rab5a from molecular dynamics simulations. *Biochemical and Biophysical Research Communications, 390*, 608–612. doi:10.1016/j.bbrc.2009.10.014

Wang, J. F., & Chou, K. C. (2010). Insights from studying the mutation-induced allostery in the M2 proton channel by molecular dynamics. *Protein Engineering Design and Selection, 23*, 663–666. doi:10.1093/protein/gzq040

Wang, J. F., Gong, K., Wei, D. Q., & Li, Y. X. (2009). Molecular dynamics studies on the interactions of PTP1B with inhibitors: From the first phosphate-binding site to the second one. *Protein Engineering Design and Selection, 22*, 349–355. doi:10.1093/protein/gzp012

Wang, M., Yang, J., & Xu, Z. J. (2005). SLLE for predicting membrane protein types. *Journal of Theoretical Biology, 232*, 7–15. doi:10.1016/j.jtbi.2004.07.023

Wang, X., Zhang, W., Zhang, Q., & Li, G. Z. (2015). MultiP-SChlo: Multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier. *Bioinformatics, 31*, 2639–2645. doi:10.1093/bioinformatics/btv212

Wu, Z. C., Xiao, X., & Chou, K. C. (2010). 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. *Journal of Theoretical Biology, 267*, 29–34. doi:10.1016/j.jtbi.2010.08.007

Xiao, X., Min, J. L., Lin, W. Z., & Liu, Z. (2015). iDrug-Target: Predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. *Journal of Biomolecular Structure & Dynamics (JBSD), 33*, 2221–2233. doi:10.1080/07391102.2014.998710

Xiao, X., Min, J. L., & Wang, P. (2013a). iCDI-PseFpt: Identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. *Journal of Theoretical Biology, 337*, 71–79. doi:10.1016/j.jtbi.2013.08.013

Xiao, X., Min, J. L., & Wang, P. (2013b). iGPCR-Drug: A web server for predicting interaction between GPCRs and drugs in cellular networking. *PLoS ONE, 8*, e72234. doi:10.1371/journal.pone.0072234

Xiao, X., Min, J. L., & Wang, P. (2013c). Predict drug–protein interaction in cellular networking. *Current Topics in Medicinal Chemistry, 13*, 1707–1712. doi:10.2174/15680266113139990121

Xiao, X., Wang, P., Lin, W. Z., & Jia, J. H. (2013). iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Analytical Biochemistry, 436*, 168–177. doi:10.1016/j.ab.2013.01.019

Xiao, X., Wu, Z. C., & Chou, K. C. (2011). iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *Journal of Theoretical Biology, 284*, 42–51. doi:10.1016/j.jtbi.2011.06.005

Xu, Y., Ding, J., & Wu, L. Y. (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. doi:10.1371/journal.pone.0055844

Xu, Y., Shao, X. J., Wu, L. Y., & Deng, N. Y. (2013). iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ, 1*, e171. doi:10.7717/peerj.171

Xu, Y., Wen, X., & Shao, X. J. (2014). iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *International Journal of Molecular Sciences, 15*, 7594–7610. doi:10.3390/ijms15057594

Xu, Y., Wen, X., Wen, L. S., & Wu, L. Y. (2014). iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS ONE, 9*, e105018. doi:10.1371/journal.pone.0105018

Xu, R., Zhou, J., Liu, B., He, Y. A., & Zou, Q. (2015). Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. *Journal of Biomolecular Structure & Dynamics (JBSD), 33*, 1720–1730. doi:10.1080/07391102.2014.968624

Yan, C., Dobbs, D., & Honavar, V. (2004). A two-stage classifier for identification of protein–protein interface residues. *Bioinformatics, 20*, i371–i378. doi:10.1093/bioinformatics/bth920

Zhang, C. T., & Chou, K. C. (1992). An optimization approach to predicting protein structural class from amino acid composition. *Protein Science, 1*, 401–408. doi:10.1002/pro.5560010312

Zhang, J., Zhao, X., Sun, P., & Ma, Z. (2014). PSNO: Predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC. *International Journal of Molecular Sciences, 15*, 11204–11219. doi:10.3390/ijms150711204

Zhang, L., Zhao, X., & Kong, L. (2014). Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition. *Journal of Theoretical Biology, 355*, 105–110. doi:10.1016/j.jtbi.2014.04.008

Zhang, S. W., Zhang, Y. L., Yang, H. F., Zhao, C. H., & Pan, Q. (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. doi:10.1007/s00726-007-0010-9

Zhong, W. Z., & Zhou, S. F. (2014). Molecular science for drug development and biomedicine. *International Journal of Molecular Sciences, 15*, 20072–20078. doi:10.3390/ijms151120072

Zhou, G. P. (1989). Biological functions of soliton and extra electron motion in DNA structure. *Physica Scripta, 40*, 698–701. doi:10.1088/0031-8949/40/5/021

Zhou, G. P. (2011). The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein–protein interaction mechanism. *Journal of Theoretical Biology, 284*, 142–148. doi:10.1016/j.jtbi.2011.06.006

Zhou, G. P. (2015). Current progress in structural bioinformatics of protein–biomolecule interactions. *Medicinal Chemistry, 11*, 216–217. doi:10.2174/1573406411666141229162618

Zhou, P., Tian, F., Li, B., Wu, S., & Li, Z. (2006). Genetic algorithm-based virtual screening of combinative mode for peptide/protein. *Acta Chimica Sinica-Chinese Edition, 64*, 691. doi:10.3321/j.issn:0567-7351.2006.07.018

Zia-ur-Rehman, & 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 & Peptide Letters, 19*, 890–903. doi:10.2174/092986612801619589

Zuo, Y. C., Peng, Y., Liu, L., Chen, W., Yang, L., & Fan, G. L. (2014). Predicting peroxidase subcellular location by hybridizing different descriptors of Chou' pseudo amino acid patterns. *Analytical Biochemistry, 458*, 14–19. doi:10.1016/j.ab.2014.04.032