iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC

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

[a] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China
[b] School of Computer Science, University of Birmingham, Edgbaston Birmingham B15 2TT, UK
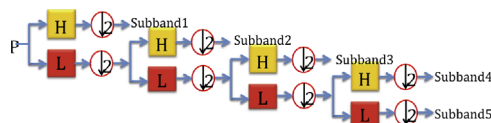[c] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia
[d] Gordon Life Science Institute, Boston, MA 02478, USA

HIGHLIGHTS

- Protein–protein interactions.
- Discrete wavelet transform approach.
- Ensemble classifier formed by fusing seven individual random forest.
- Web-server predictor.

GRAPHICAL ABSTRACT

H is high pass filter   L is low pass filter

↓2 is the operator of downsampling with 2

ARTICLE INFO

ABSTRACT

A cell contains thousands of proteins. Many important functions of cell are carried out through the proteins therein. Proteins rarely function alone. Most of their functions essential to life are associated with various types of protein–protein interactions (PPIs). Therefore, knowledge of PPIs is fundamental for both basic research and drug development. With the avalanche of proteins sequences generated in the postgenomic age, it is highly desired to develop computational methods for timely acquiring this kind of knowledge. Here, a new predictor, called "iPPI-Emsl", is developed. In the predictor, a protein sample is formulated by incorporating the following two types of information into the general form of PseAAC (pseudo amino acid composition): (1) the physicochemical properties derived from the constituent amino acids of a protein; and (2) the wavelet transforms derived from the numerical series along a protein chain. The operation engine to run the predictor is an ensemble classifier formed by fusing seven individual random forest engines via a voting system. It is demonstrated with the benchmark dataset from *Saccharomyces cerevisiae* as well as the dataset from *Helicobacter pylori* that the new predictor achieves remarkably higher success rates than any of the existing predictors in this area. The new predictor' web-server has been established at http://www.jci-bioinfo.cn/iPPI-Esml. For the convenience of most experimental scientists, we have further provided a step-by-step guide, by which users can easily get their desired results without the need to follow the complicated mathematics involved during its development.

* Corresponding authors at: Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China.
  E-mail addresses: jjh163yx@163.com (J. Jia), liuzi189836@163.com (Z. Liu), xxiao@gordonlifescience.org (X. Xiao), lbx1966@163.com (B. Liu),
kcchou@gordonlifescience.org (K.-C. Chou).
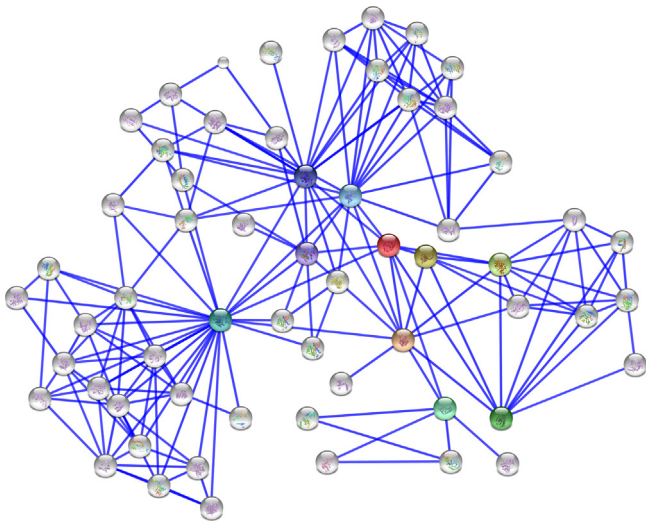
Fig. 1. A complicated protein–protein interaction network is expressed by the vertex–arc graph, where the vertex is used to represent each of the proteins in the network system while the arc to indicate their relation. If there is an arc between two proteins, they are in interaction with each other; otherwise, they are not. For more about using the graphic approach to deal with complicated biological systems, see (Chou, 1989a, 2010).

1. Introduction

Proteins play a vital role in nearly all biology functions such as composing cellular structure and promoting chemical reactions. Proteins in a cell rarely function in isolation but are actively and selectively interacting with each other (Fig. 1). Most of their functions essential to life are associated with different types of protein–protein interactions (PPIs). For instance, proteins are modified and degraded by enzyme proteins; many marvelous biological functions in proteins, such as allosteric regulation, are realized via the interactions between the protein subunits (Chou, 1989b; Perutz, 1942); signal transmission between cells is via binding of protein messengers to protein receptors (Chou, 2005a); proteins are directed to the correct compartments of cells through binding to other proteins; structural connections between cells are established via PPIs; the molecular mechanism of muscle contraction as well as the opening and closing of ion-channels are also closely associated with PPIs (Huang et al., 2008; OuYang et al., 2013; Schnell and Chou, 2008).

On the other hand, PPIs are also related to various disease states. For instance, if a cell suddenly lost some normal or necessary PPIs, the deprived cell would become "blind" and "deaf", completely paralytic finally leading to perish. Also, if many abnormal or unwanted PPIs suddenly occurred in a cell, the "unfortunate" cell would completely lose control, leading to network confuse and a terrible disaster.

Accordingly, it is vitally important to characterize PPIs and understand their interaction network, an important subject in the discipline called "protein network" or a frontier for investigating the functional relationship of proteins in a cell. However, it is by no means an easy job due to the extreme complexity of the problem concerned. As is well known, using graphical approaches to study complicated biological problems can provide an intuitive picture or useful insights for helping analyzing complicated relations in these systems, as demonstrated by many previous studies on a series of important biological topics, such as enzyme-catalyzed reactions (Zhou and Deng, 1984), inhibition of HIV-1 reverse transcriptase (Althaus et al., 1993a, 1993b) , protein folding kinetics (Chou, 1990), and using wenxiang diagram or graph (Chou et al., 2011) to study protein–protein interactions (Zhou,

2011a, 2011b; Zhou and Huang, 2013). In view of this, we are also using the vertex–arc graph (Fig. 1) to express a protein–protein interaction network, where the vertex denotes each of the proteins in the network system while the arc to indicate their relation. As we can see from the figure, the PPI systems are indeed very complicated. Therefore, it is absolutely necessary to combine the experimental and computational approaches together for really understanding this kind of systems.

During the last decade or so, various experimental techniques have been developed for determining PPIs, such as yeast two-hybrid systems (Fields and Song, 1989; Ito et al., 2001), mass spectrometry (Gavin et al., 2002), and protein chip (Zhu et al., 2001). But only very small portion of PPI's were identified (Han et al., 2005) because it was time-consuming, labor-intensive and expensive by using experimental technique alone.

Fortunately, the success of the human genome project has provided us with a significant amount of useful data to conduct statistical analyses in this regard, and hence made it feasible to predict the PPIs by computational approaches.

Our rationale is as follows. It is virtually axiomatic that the amino acid sequence of a protein will determine its 3D (three-dimensional) structure (Anfinsen, 1973); while the latter will determine its biological function. Accordingly, the sequence information alone of proteins can certainly determine their interaction relationship. Actually, many computational methods have been proposed in this regard (Chou and Cai, 2006; Espadaler et al., 2005; Gomez et al., 2003; Guo et al., 2008b; Marcotte et al., 1999; Shen et al., 2007b; Xia et al., 2010a, 2010b; Yang et al., 2010). Each of these methods has its own merit, and did play a role in stimulating the development of this area. However, all the aforementioned methods were based on a single learner without using the ensemble learning technique, and hence their power might be limited. Besides, in the aforementioned methods, none of physicochemical properties was taken into account, which might further limit the prediction quality.

Many evidences have indicated that using ensemble classifier can significantly enhance the success rates in recognizing protein fold pattern (Shen and Chou, 2006), identifying membrane protein types (Shen and Chou, 2007), and predicting protein subcellular localization (see, e.g., Chou and Shen (2006a), Shen et al. (2007a)). In other words, in comparison with a single classifier, the ensemble classifier formed by fusing multiple single classifiers can achieve much better prediction quality with more generalized ability (Chou and Shen, 2007a; Jia et al., 2011).

Stimulated by the successes of using ensemble classifiers for predicting protein attributes (Chou and Shen, 2006a; Shen and Chou, 2006), Nanni et al. (Nanni and Lumini, 2006) developed an ensemble classifier by fusing K-local hyperplanes for predicting PPIs, remarkably enhancing the success rate. Unfortunately, for the prediction method (Nanni and Lumini, 2006), no web-server whatsoever has been established. Therefore, its practical application value is considerably limited, particularly for most experimental scientists.

The present study was aimed at (1) developing a new and more powerful ensemble classifier by incorporating the physicochemical properties concerned, and (2) establishing a user-friendly web-server for the new PPI predictor.

As reflected by a number of recent articles (Chen et al., 2013, 2014a, 2014b; Guo et al., 2014; Lin et al., 2014; Liu et al., 2015b, 2014a, 2014b; Xu et al., 2014b) in response to the call (Chou, 2011), in presenting a sequence-based statistical predictor for a biological system, one should make the following five procedures very clear: (1) how to construct or select a valid benchmark dataset 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; and (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. Material and methods

2.1. Benchmark datasets

Two benchmark datasets were used for the current study. One is called the S.C. dataset used to study the PPIs in the cell of *Saccharomyces cerevisiae*, while the other called the H.P. dataset used to study the PPIs in the cell of *Helicobacter pylori*.

2.1.1. S.C. dataset

To obtain a high quality benchmark dataset, the source *S. cerevisiae* proteins for the S.C. dataset were collected according to the following criteria: (1) each of the included proteins must contain at least 50 residues in order to avoid fragments (Chou and Shen, 2007a); (2) none of the included proteins has $\geq 40\%$ pairwise sequence identity to any other in order to reduce the homology bias. From the 7374 source proteins thus obtained and using DIP (database of interacting proteins) (Xenarios et al., 2002), we can obtain 17,505 interactive protein pairs. As for the non-interacting pairs that are not readily available from DIP database, we constructed them as follows. The non-interactive pairs were generated based on such an assumption that proteins located at different subcellular localizations do not interact with each other (Guo et al., 2008a; Shen et al., 2007b). The subcellular location information of the proteins concerned was extracted from Swiss-Prot (http://www.expasy.org/sprot/) by considering the following seven locations: cytoplasm, nucleus, mitochondrion, endoplasmic reticulum, Golgi apparatus, peroxisome, and vacuole. Subsequently, the negative data were formed via pairing the proteins concerned from one location site with those from a different one. The following requirement must be satisfied when doing so (Guo et al., 2008a; Shen et al., 2007b): the non-interacting pairs thus formed should not also occur in the positive dataset of interactive pairs. A total of 5943 negative pairs were thus generated. As pointed out by the authors in (Ben-Hur and Noble, 2006), however, the restricting negative samples formed from different subcellular locations may lead to a biased estimate of the accuracy for a PPI predictor. Therefore, it is necessary to also generate the negative samples from the same subcellular location to reduce this kind of bias. In view of this, additional negative samples were generated according to the principle that the protein pairs at the same localization were considered as the negative pairs if none of them has occurred in the yeast positive pairs. Thus, additional 27,204 negative pairs were generated within each of the aforementioned seven subcellular locations: 8000 within cytoplasm, 8000 within nucleus, 8284 within mitochondrion, 1953 within endoplasmic reticulum, 300 within Golgi apparatus, 171 within peroxisome, and 496 within vacuole. Finally, the benchmark dataset thus established can be formulated below

$$\mathbb{S}_{\text{S.C.}} = \mathbb{S}_{\text{S.C.}}^{+} \cup \mathbb{S}_{\text{S.C.}}^{-} \tag{1}$$

where $\mathbb{S}_{\text{S.C.}}$ is the S.C. benchmark dataset for *S. cerevisiae* that contains 50,652 protein pairs, of which 17,505 are interactive pairs belonging to the positive subset $\mathbb{S}_{\text{S.C.}}^{+}$, $5943 + 27,204 = 33,147$ are non-interactive pairs belonging to the negative subset $\mathbb{S}_{\text{S.C.}}^{-}$, and $\cup$ represents the union in the set theory. For the details of these protein pairs and their DIP codes, see Supporting information S1.

2.1.2. H.P. dataset

For facilitating comparison later, the benchmark dataset used to study the PPIs in the cell of *H. pylori* was taken from (Martin et al., 2005) since many investigators used it to test their own methods with the success rates well documented (see, e.g., (Nanni, 2005; Nanni and Lumini, 2006; Xia et al., 2010b)). Likewise, the H. P. dataset can be formulated as

$$\mathbb{S}_{\text{H.P.}} = \mathbb{S}_{\text{H.P.}}^{+} \cup \mathbb{S}_{\text{H.P.}}^{-} \tag{2}$$

where $\mathbb{S}_{\text{H.P.}}$ contains 2,916 *H. pylori* protein pairs, $\mathbb{S}_{\text{H.P.}}^{+}$ is the positive subset containing 1458 interactive protein pairs, and $\mathbb{S}_{\text{H.P.}}^{-}$ is the negative subset containing 1458 non-interactive protein pairs. For the details of these protein pairs and their corresponding protein sequences, see Supporting information S2 and S3, respectively.

2.2. Using pseudo amino acid composition to represent protein pairs

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 and Chou, 1992), correlation-angle approach (Chou, 1993), covariance discriminant (CD) (Chen et al., 2012a, , 2012b), LogitBoost (Feng et al., 2005), SLLE algorithm (Wang et al., 2005), support vector machine (SVM) (Lin et al., 2014; Xu et al., 2014a), random forests (Lin et al., 2011), conditional random field (Xu et al., 2013a), nearest neighbor (NN) (Cai and Chou, 2003), K-nearest neighbor (KNN) (Shen et al., 2006), OET-KNN (Shen and Chou, 2009a), Fuzzy K-nearest neighbor (Xiao et al., 2013a), and ML-KNN algorithm (Xiao et al., 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, 2001a, 2005b) or Chou's PseAAC (Cao et al., 2013; Du et al., 2012; Lin and Lapointe, 2013) was proposed. Ever since it was introduced in 2001 (Chou, 2001a), the concept of PseAAC has been widely used in almost all the areas of computational proteomics, such as in predicting antifreeze proteins (Mondal and Pai, 2014), predicting protein structural class (Kong et al., 2014; Zhang et al., 2014b), predicting anticancer peptides (Hajisharifi et al., 2014), identifying bacterial virulent proteins (Nanni et al., 2012b), predicting protein subcellular location in various organisms and levels (Kandaswamy et al., 2010; Li et al., 2014; Mei, 2012; Nanni and Lumini, 2008; Zhang et al., 2008; Zuo et al., 2014), predicting membrane protein types (Chen and Li, 2013; Han et al., 2014), discriminating outer membrane proteins (Hayat and Khan, 2012), analyzing genetic sequence (Georgiou et al., 2013), identifying cyclin proteins (Mohabatkar, 2010), predicting GABA(A) receptor proteins (Mohabatkar et al., 2011), identifying antibacterial peptides (Khosravian et al., 2013), identifying allergenic proteins (Mohabatkar et al., 2013), predicting metalloproteinase family (Mohammad Beigi et al., 2011), identifying GPCRs and their types (ZiaUr and Khan, 2012), identifying the types of conotoxins (Ding et al., 2014), identifying protein quaternary structural attributes (Sun et al., 2012), identifying risk type of human papillomaviruses (Esmaeili et al., 2010), identifying various PTM (post-translational modification) sites in proteins (Jia et al., 2014; Qiu et al., 2014b, 2014c; Xu et al., 2013a, 2013b; Zhang et al., 2014a), among many

others (see a long list of references cited in a recent article (Du et al., 2014)). It has also been used in some disciplines of drug development and biomedicine (Zhong and 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 and nucleotides (Chen et al., 2013, 2014c; Guo et al., 2014; Liu et al., 2015a; Qiu et al., 2014a), as well as other biological samples (see, e.g., Jiang et al. (2013)). Because it has been widely and increasingly used, recently three types of powerful open access soft-ware, 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 3rd one for those of Chou's general PseAAC.

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

$$\mathbf{P} = \begin{bmatrix} \psi_1 & \psi_2 & \cdots & \psi_u & \cdots & \psi_\Omega \end{bmatrix}^{\mathbf{T}} \tag{3}$$

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 Eq. (3) will depend on how to extract the desired information from a protein sequence. Below, we are to describe how to extract the useful information from the aforementioned benchmark datasets to define a pair of proteins via Eq. (3).

The wavelet transform (Mallat, 1989) is a very effective approach for using Eq. (3) to formulate a biological sequence, as demonstrated by a series of recent studies such as (1) using wavelets to formulate PseAAC (Chou, 2001a, 2005b) for predicting membrane protein types (Liu et al., 2005), predicting protein structural classes (Chen et al., 2012a; Li et al., 2009), predicting enzyme family classes (Qiu et al., 2010), predicting protein classification (Nanni et al., 2012a), predicting protein quaternary structural attributes (Sun et al., 2012), predicting types of homo-oligomers (Qiu et al., 2011a), as well as predicting G-protein-coupled receptor classes (Qiu et al., 2009); and (2) using wavelets to formulate PseKNC (pseudo-trinucleotide composition) for predicting promoters (Zhou et al., 2013). As is well known, in molecular and cellular biology many remarkable functions in proteins and DNA can be revealed through the low-frequency (or terahertz frequency) collective motion (Chou, 1988, 1989b; Chou and Mao, 1988; Gordon, 2008). In view of this, it would be particularly intriguing to define the components of Eq. (3) with the wavelets transform approach because it may help to capture the features important for studying PPI.

Below, we use the wavelets transform to define each of the components in Eq. (3) via the amino acid's physicochemical properties.

3. Physicochemical properties

Given a protein sample with L residues as expressed by

$$\mathbf{P} = R_1 R_2 R_3 R_4 R_5 R_6 R_7 \cdots R_L \tag{4}$$

where R_1 represents the 1st amino acid residue of the protein$\mathbf{P}$, R_2 the 2nd residue, and so forth. Different types of amino acid in Eq. (4) 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 and Woods, 1981) or $\Phi^{(2)}$; (3) side-chain volume (Krigbaum and Komoriya, 1979) or $\Phi^{(3)}$; (4) polarity (Grantham, 1974) or $\Phi^{(4)}$; (5) polarizability (Charton and Charton, 1982) or $\Phi^{(5)}$; (6) solvent-accessible surface area (SASA) (Rose et al., 1985) or $\Phi^{(6)}$; and (7) side-chain net charge index (NCI) (Zhou et al., 2006) or $\Phi^{(7)}$. Their numerical values are given in Table 1. Thus, the protein $\mathbf{P}$ of Eq. (4) can be encoded into seven different numerical

series, as formulated by

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

where $\Phi_1^{(1)}$ is the hydrophobicity value of R_1 in Eq. (4), $\Phi_2^{(2)}$ the hydrophilicity value of R_2, and so forth. Note that before substituting the physicochemical values of Table 1 into Eq. (5), they all are subjected to the following standard conversion:

$$\Phi_i^{(\xi)} \Leftarrow \frac{\Phi_i^{(\xi)} - \left\langle \Phi_i^{(\xi)} \right\rangle}{\mathrm{SD}\left(\Phi_i^{(\xi)} \right)} (\xi = 1, 2, \cdots, 7; \ i = 1, 2, \cdots, L) \tag{6}$$

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 Eq. (6) 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.

4. Discrete wavelet transform (DWT)

Wavelet transform (WT) is a multi-resolution analysis tool (Mallat, 1989). It is quite popular for analyzing, de-noising and compressing signals and images. The WT approach can overcome the shortcoming of Fourier analysis, which is based on the functions that are localized in frequency domain but not in time domain. A digital signal can be decomposed into many groups of coefficients in different scales with WT, and these coefficient vectors can exhibit characteristics in time domain and frequency domain.

The DWT approach can transform a discrete time signal to a discrete wavelet representation. When using the DWT on any of the seven numerical series for protein $\mathbf{P}$ (cf. Eq. (5)), we can view it as a discrete time series, with the 1st residue as $t = 1$, 2nd residue as $t = 2$,

Table 1
The original values of the seven physicochemical properties $\Phi^{(\xi)}$ ($\xi = 1, 2, \cdots, 7$) for the 20 native amino acids (cf. Eq. (5)).

Amino acid code	$\Phi^{(1)}$	$\Phi^{(2)}$	$\Phi^{(3)}$	$\Phi^{(4)}$	$\Phi^{(5)}$	$\Phi^{(6)}$	$\Phi^{(7)}$
A	0.620	−0.500	27.500	8.100	0.046	1.181	7.187×10^{-3}
C	0.290	−1.000	44.600	5.500	0.128	1.461	-3.661×10^{-2}
D	−0.900	3.000	40.000	13.000	0.105	1.587	-2.382×10^{-2}
E	−0.740	3.000	62.000	12.300	0.151	1.862	6.802×10^{-3}
F	1.190	−2.500	115.500	5.200	0.290	2.228	3.755×10^{-2}
G	0.480	0.000	0.000	9.000	0.000	0.881	1.791×10^{-1}
H	−0.400	−0.500	79.000	10.400	0.230	2.025	-1.069×10^{-2}
I	1.380	−1.800	93.500	5.200	0.186	1.810	2.163×10^{-2}
K	−1.500	3.000	100.000	11.300	0.219	2.258	1.771×10^{-2}
L	1.060	−1.800	93.500	4.900	0.186	1.931	5.167×10^{-2}
M	0.640	−1.300	94.100	5.700	0.221	2.034	2.683×10^{-3}
N	−0.780	2.000	58.700	11.600	0.134	1.655	5.392×10^{-3}
P	0.120	0.000	41.900	8.000	0.131	1.468	2.395×10^{-1}
Q	−0.850	0.200	80.700	10.500	0.180	1.932	4.921×10^{-2}
R	−2.530	3.000	105.000	10.500	0.291	2.560	4.359×10^{-2}
S	−0.180	0.300	29.300	9.200	0.062	1.298	4.627×10^{-3}
T	−0.050	−0.400	51.300	8.600	0.108	1.525	3.352×10^{-3}
V	1.080	−1.500	71.500	5.900	0.140	1.645	5.700×10^{-2}
W	0.810	−3.400	145.500	5.400	0.409	2.663	3.798×10^{-2}
Y	0.260	−2.300	117.300	6.200	0.298	2.368	2.360×10^{-2}

and so forth. The discrete time series thus obtained is input into one high-pass filter and one low-pass filter. The coefficients thus obtained can be approximately used for the signal's high scale and low frequency components. In practice, such transform will be applied recursively on the low-pass series with the Mallat algorithm (Mallat, 1999) until the desired number of iterations is reached. The block diagram of Fig. 2 illustrates the digital implementation of DWT. In this study, the decomposition level $\lambda = 4$ was selected to represent a protein, which is similar to the treatment of Qiu et al. (2011b). Accordingly, we can obtain $(4+1) = 5$ sub-bands when the discrete series **P** was decomposed by DWT with level $\lambda = 4$ (see Fig. 2). Each of the five sub-bands has four coefficients: (1) α_j, the maximum of the wavelet coefficients in the $j-$th sub-band; (2) β_j the mean of the wavelet coefficients in the $j-$th sub-band; (3) γ_j the minimum of the wavelet coefficients in the $j-$th sub-band; (4) δ_j the standard deviation of the wavelet coefficients in the $j-$th sub-band ($j = 1, 2, \cdots, 5$). Thus, in a way quite similar to the treatment in (Qiu et al., 2014b; Xu et al., 2013a, 2013b), each of the components in Eq. (3) can be formulated as

$$\psi_u = \begin{cases} \alpha_u & \text{if } 1 \le u \le 5 \\ \beta_u & \text{if } 6 \le u \le 10 \\ \gamma_u & \text{if } 11 \le u \le 15 \\ \delta_u & \text{if } 16 \le u \le 20 = \Omega \end{cases} \tag{7}$$

For a protein pair formed by $\mathbf{P}^{k1}$ and $\mathbf{P}^{k2}$, the corresponding PseAAC can be formulated by their orthogonal sum (Chou and Cai, 2006); i.e.,

$$\mathbf{P}^{k1} \oplus \mathbf{P}^{k2} = \begin{bmatrix} \psi_1^{k1} & \psi_2^{k1} & \cdots & \psi_{20}^{k1} & \psi_1^{k2} & \psi_2^{k2} & \cdots & \psi_{20}^{k2} \end{bmatrix}^{\mathbf{T}} \tag{8}$$

where $\mathbf{P}^{k1}$ and $\mathbf{P}^{k2}$ as well as their components have exactly the same meaning as those in Eq. (3) except for that they are now referred to a specified protein $\mathbf{P}^{k1}$ or $\mathbf{P}^{k2}$ instead of a general protein **P**, and the symbol $\oplus$ represents the sign of orthogonal sum (Chou and Cai, 2006).

Note that when in turn using each of the seven different physicochemical features (cf. Eq. (5)), we can generate seven different PseAAC vectors to represent a same protein pair, as formulated by

$$\text{Protein} - \text{pair} = \mathbf{P}^{k1} \oplus \mathbf{P}^{k2}(\xi) = \begin{cases} \text{hydrophobicity} & \xi = 1 \\ \text{hydrophicility} & \xi = 2 \\ \text{side} - \text{chain volume} & \xi = 3 \\ \text{polarity} & \xi = 4 \\ \text{polarizability} & \xi = 5 \\ \text{solvent} - \text{accessible surface} & \xi = 6 \\ \text{side} - \text{chain net charge} & \xi = 7 \end{cases} \tag{9}$$
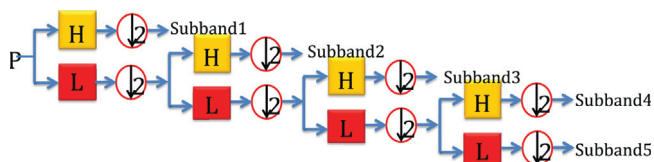
H is high pass filter   L is low pass filter

② is the operator of downsampling with 2

Fig. 2. A schematic drawing to illustrate the procedure of multi-level DWT (discrete wavelet transform). See the text for further explanation.

4.1. Random forest and ensemble classifier

The random forests (RF) algorithm is a powerful algorithm and has been used in many areas of computational biology (see, e.g. (Kandaswamy et al., 2011; Lin et al., 2011; Pugalenthi et al., 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.

As shown in Eq. (9), a protein pair can be formulated with seven different PseAAC forms, each of which can be used to train the RF predictor. Accordingly, we have a total of seven individual predictors for identifying PPIs, as formulated by

$$\text{PPI individual predictor} = \mathbb{RF}(\xi) \quad (\xi = 1, 2, \cdots, 7) \tag{10}$$

where $\mathbb{RF}(\xi)$ represents the RF predictor based on the $\xi-$th physicochemical property (cf. Eqs. (5), (6), (9)).

Now, the problem is how to combine the results from the seven individual predictors to maximize the prediction quality. As indicated by a series previous studies, using the ensemble classifier formed by fusing many individual classifiers can remarkably enhance the success rates in predicting protein subcellular localization (Chou and Shen, 2006b, 2007b) and protein quaternary structural attribute (Shen and Chou, 2009b). Encouraged by the previous investigators' studies, here we are also to develop an ensemble classifier by fusing the seven individual predictors $\mathbb{RF}(\xi)$ ($\xi = 1, 2, \cdots, 7$) through a voting system, as formulated by

$$\mathbb{RF}^{\text{E}} = \mathbb{RF}(1) \forall \mathbb{RF}(2) \forall \cdots \forall \mathbb{RF}(7) = \forall_{\xi=1}^{7} \mathbb{RF}(\xi) \tag{11}$$

where $\mathbb{RF}^{\text{E}}$ represents the ensemble classifier, and the symbol $\forall$ denotes 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 Eqs. 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 Fig. 3 to illustrate how the seven individual RF predictors are fused into the ensemble classifier.

The final predictor thus obtained is called "iPPI-Esml", where "i" stands for "identify", "PPI" for "protein–protein interaction", and "Esml" for "ensemble learning".

4.2. Evaluation metrics and validation method

For identifying whether the two counterparts in a pair of proteins are interacting with each other, four metrics are often used in 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 et al. (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 (2001b) 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 et al., 2014a):

$$\begin{cases} \text{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \le \text{Sn} \le 1 \\[2mm] \text{Sp} = 1 - \dfrac{N_+^-}{N^-} & 0 \le \text{Sp} \le 1 \\[2mm] \text{Acc} = \Lambda = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \text{Acc} \le 1 \\[2mm] \text{MCC} = \dfrac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-} \right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+} \right) \left(1 + \frac{N_-^+ - N_+^-}{N^-} \right)}} & -1 \le \text{MCC} \le 1 \end{cases} \tag{12}$$

where N^+ represents the total number of interactive protein pairs investigated whereas N_-^+ the number of true interactive pairs incorrectly predicted as the non-interactive pairs; N^- the total
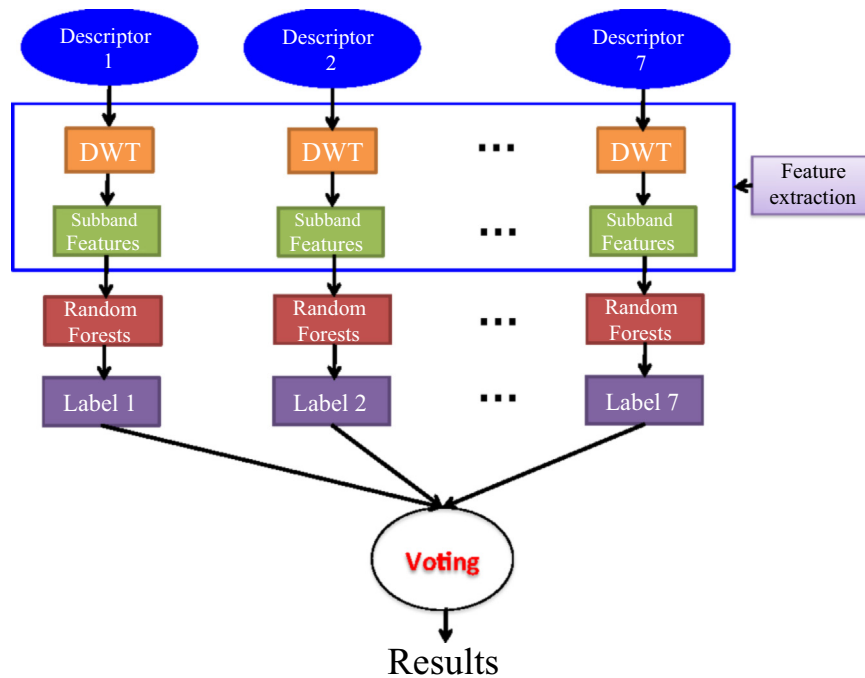
Fig. 3. A flowchart to show how an ensemble classifier is formed via a voting system.

number of the non-interactive protein pairs investigated whereas N_+^- the number of non-interactive protein pairs incorrectly predicted as the interactive pairs.

Now with Eq. (12) at hands, it is crystal clear to see the following. When $N_-^+ = 0$ meaning none of the interactive protein pairs is incorrectly predicted to be a non-interactive pairs, we have the sensitivity $Sn = 1$. When $N_-^+ = N^+$ meaning that all the interactive protein pairs are incorrectly predicted to be non-interactive protein pairs, we have the sensitivity $Sn = 0$. Likewise, when $N_+^- = 0$ meaning none of the non-interactive protein pairs was incorrectly predicted to be the interactive protein pairs, we have the specificity $Sp = 1$; whereas $N_+^- = N^-$ meaning that all the non-interactive protein pairs were incorrectly predicted as interactive pairs, we have the specificity $Sp = 0$. When $N_-^+ = N_+^- = 0$ meaning that none of interactive protein pairs in the positive dataset and none of the non-interactive protein pairs in the negative dataset was incorrectly predicted, we have the overall accuracy $Acc = 1$ and $MCC = 1$; when $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the interactive protein pairs in the positive dataset and all the non-interactive protein pairs in the negative dataset were incorrectly predicted, we have the overall accuracy $Acc = 0$ and $MCC = -1$; whereas when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$ we have $Acc = 0.5$ and $MCC = 0$ meaning no better than random prediction. As we can see from the above discussion based on Eq. (12), the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient have become much more intuitive and easier-to-understand.

It should be pointed out, however, the set of metrics as defined in Eq. (12) is valid only for the single-label systems. For the multi-label systems whose emergence has become more frequent in system biology (Chou et al., 2012; Lin et al., 2013; Xiao et al., 2011) and system medicine (Chen et al., 2012b; Xiao et al., 2013b), 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 dataset test, subsampling (or K-fold cross-validation)

test, and jackknife test (Chou and 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 dataset as elucidated in Chou (2011) and demonstrated by Eqs. 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 and Elrod (2002; Chou and Cai, 2003; Hajisharifi et al., 2014; Mohabatkar et al., 2013; Mondal and Pai, 2014; Nanni et al., 2014; Shen et al., 2007a; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003)). However, to reduce the computational time, in this study we adopted the 5-fold cross-validation and 10-fold cross validations, as done by most investigators with random forests algorithm as the prediction engine.

4.3. Web-server and user guide

To enhance the value of its practical applications, a web-server for iPPI-Esml has been established at http://www.jci-bioinfo.cn/iPPI-Esml. Furthermore, to maximize the convenience for most experimental scientists, a step-to-step guide or protocol is provided below:

Step 1: Opening the web-server at http://www.jci-bioinfo.cn/iPPI-Esml, you will see the top page of iPPI-Emsl on your computer screen, as shown in Fig. 4. Click on the Read Me button to see a brief introduction about the PPI predictor.
Step 2: Either type or copy/paste the query protein sequences into the input box at the center of Fig. 4. The input sequence should be in the FASTA format. 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 query protein sequences in the Example window as the input, you will see the following shown on the screen of your computer: (1) Proteins example-1 and 2 belong to non-interacting pair because their voting score for interaction is $3/7 \approx 0.43$, smaller than $4/7 \approx 0.57$. (2) Proteins example-1 and 3 belong to non-interacting pair

because their voting score for interaction is $3/7 \approx 0.43$, smaller than $4/7 \approx 0.57$. (3) Proteins example-2 and 3 belong to interacting pair because their voting score for interaction is $5/7 \approx 0.71 \geq 4/7 \approx 0.57$. All these results are fully consistent with the experimental observations.

Step 4: As shown on the lower panel of Fig. 4, you may also choose the batch prediction by entering your e-mail address and your desired batch input file (in FASTA format) 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 iPPI-Esml.

Step 6: Click the Supporting Information button to download the benchmark dataset used to train and test the current PPI predictor.

5. Results and discussion

The proposed predictor was first tested by the benchmark dataset $\mathbb{S}_{S.C.}$ in Eq. (1) from *S. cerevisiae*, which contains 17,505 interactive protein pairs and 33,147 non-interactive protein pairs (cf. Supporting Information S1). The benchmark dataset was randomly separated into a training dataset $\mathbb{S}_{S.C.}$ (train) and a testing dataset $\mathbb{S}_{S.C.}$ (test); i.e.,

$$\mathbb{S}_{S.C.} = \mathbb{S}_{S.C.} \text{ (train)} \cup \mathbb{S}_{S.C.} \text{ (test)} \tag{13}$$

where $\mathbb{S}_{S.C.}$ (train) contains 5943 interactive pairs and 5943 non-interactive pairs, while $\mathbb{S}_{S.C.}$ (test) contains 11,562 interactive pairs and 27,204 non-interactive pairs.

Listed in Table 2 are the values of four metrics (cf. Eq. (12)) obtained by iPPI-Emsl via the 5-fold cross-validation on $\mathbb{S}_{S.C.}$ (train). For facilitating comparison, listed in that table are also the corresponding rates obtained by the method proposed by Guo et al. (2008b).

Listed in Table 3 are the corresponding results on the $(11,562 + 27,204) = 38,766$ samples in the independent testing dataset $\mathbb{S}_{S.C.}$ (test) but trained with $(5943 + 5943) = 11,886$ samples in the training dataset $\mathbb{S}_{S.C.}$ (train).

It can be clearly seen from Tables 2 and 3, the new predictor iPPI-Emsl remarkably outperformed the Guo et al.'s method (Guo et al., 2008b) via both the 5-fold cross-validation and independent dataset tests, indicating the proposed predictor is indeed a quite powerful one.

Fig. 4. A semi-screenshot to show the top-page of the iPPIs-Emsl web-server at http://www.jci-bioinfo.cn/iPPI-Esml.

Table 2
The results obtained by the 5-fold cross-validation on the dataset. $\mathbb{S}_{S.C.}$ (train) (cf. Eq. (13)). See Eq. (12) for the definitions of Acc, MCC, Sn, and Sp.

Method	Acc (%)	MCC	Sn (%)	Sp (%)
This paper[a]	84.39	0.6897	87.03	82.13
Guo et al.[b]	77.96	0.5099	76.84	78.22

[a] The proposed predictor iPPI-Emsl.
[b] See Guo et al. (2008b).

Table 3
The results on the $(11,562 + 27,204) = 38,766$ samples in $\mathbb{S}_{S.C.}$ (test) but trained with $(5943 + 5943) = 11,886$ samples in $\mathbb{S}_{S.C.}$ (train).

Method	Acc (%)	MCC	Sn (%)	Sp (%)
This paper[a]	86.45	0.6832	75.59	91.53
Guo et al.[b]	78.65	0.5171	64.85	85.00

[a] See footnote a of Table 2.
[b] See footnote a of Table 2.

Table 4
Compared with the other six state-of-art methods via the 10-cross-validation on the *H. pylori* dataset (Martin et al., 2005).

Method	Acc (%)	MCC	Sn (%)	Sp (%)	Web-server
Bock and Gough[a]	75.80	N/A	69.80	80.20	No
Gao et al.[b]	80.96	0.5577	78.65	83.20	No
Martin[c]	83.40	N/A	79.90	85.70	No
Nanni[d]	83.00	N/A	80.60	85.10	No
Nanni and Lumini[e]	86.60	N/A	86.70	85.00	No
Xia et al.[f]	88.40	N/A	88.20	89.20	No
iPPI-Esml[g]	**90.75**	**0.8151**	90.41	87.50	Yes

[a] Results reported by Bock and Gough (2003).
[b] Results reported by Guo et al. (2008a).
[c] Results reported by Martin et al. (2005).
[d] Results reported by Nanni (2005).
[e] Results reported by Nanni and Lumini (2006).
[f] Results reported by Xia et al. (2010b).
[g] Results obtained by the current predictor using the same cross-validation method on the same benchmark dataset as the aforementioned six state-of-art-methods.

As mentioned in Section 2.1, many state-of-the art prediction methods in this area have used the benchmark dataset $\mathbb{S}_{H.P.}$ (cf. Eq. (2)) constructed by Martin et al. (2005) from the cell of *H. pylori* to examine their success rates. Below, we are also to use the same benchmark dataset to examine the proposed iPPI-Emsl predictor.

The results obtained by iPPI-Emsl on the benchmark dataset $\mathbb{S}_{H.P.}$ (cf. Supporting information S2 and S3) via the 10-fold cross-validation test are given in Table 4, where, for facilitating comparison, the rates obtained by the other methods using exactly the same benchmark dataset and exactly the same cross-validation approach are also given. As we can see from the table, the new method proposed in this paper remarkably outperformed all the other existing methods, once again demonstrating that iPPI-Emsl is really a very promising predictor for identifying protein–protein interactions. Particularly, as clearly sown in Table 4, in contrast to all the other six existing prediction methods without any web-server provided, the current proposed predictor does provide a use-friendly web-server that is no doubt very useful for the majority of experimental scientists in this or related areas.

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 (Chou, 1983, 1984; Chou and Chen, 1977; Chou et al., 1981; Gordon, 2008; Madkan et al., 2009; Sobell et al., 1983; Zhou, 1989) for in-depth understanding or revealing the dynamic mechanisms of their various important biological functions (Chou, 1988), such as cooperative effects (Chou, 1989b), allosteric transition (Chou, 1987; Schnell and Chou, 2008), assembly of microtubules (Chou et al., 1994), and switch between active and inactive states (Wang and Chou, 2009).

6. Conclusion

In the new PPI predictor, a protein pair is formulated by a general form of PseAAC whose components are defined via the following procedures: (1) a protein sequence is converted into a numerical series via the physicochemical properties of amino acids; (2) the numerical series is subsequently converted into a 20-D (dimensional) feature vector by means of the DWT technique; and (3) the protein pair sample is an orthogonal sum of the two 20-D vectors generated from its two counterparts.

The operation engine to run the PPI prediction is an ensemble classifier formed via a voting system to fuse seven different random forests classifiers based on seven different physicochemical properties, respectively.

Rigorous cross-validations have indicted that the new predictor established with the above procedures is very powerful and promising. It is anticipated that iPPI-Emsl will become a very useful high throughput tool for predicting protein–protein interactions in cells, stimulating a series of interesting follow-up researches in this and related areas.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgments

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2015.04.011.

References

Althaus, I.W., Gonzales, A.J., Chou, J.J., Diebel, M.R., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993a. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J. Biol. Chem. 268, 14875–14880.
Althaus, I.W., Chou, J.J., Gonzales, A.J., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993b. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32, 6548–6554.

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. Science 181, 223–230.
Ben-Hur, A., Noble, W.S., 2006. Choosing negative examples for the prediction of protein–protein interactions. BMC Bioinformatics 7, S2.
Bock, J.R., Gough, D.A., 2003. Whole-proteome interaction mining. Bioinformatics 19, 125–134.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
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. Biochem. Biophys. Res. Commun. 305, 407–411.
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.
Charton, M., Charton, B.I., 1982. The structural dependence of amino acid hydro-phobicity parameters. J. Theor. Biol. 99, 629–644.
Chen, C., Shen, Z.B., Zou, X.Y., 2012a. Dual-layer wavelet SVM for predicting protein structural class via the general form of Chou's pseudo amino acid composition. Protein Peptide Lett. 19, 422–429.
Chen, J., Liu, H., Yang, J., 2007. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33, 423–428.
Chen, L., Zeng, W.M., Cai, Y.D., Feng, K.Y., 2012b. Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical–chemical inter-actions and similarities. PLoS One, 7.
Chen, W., Feng, P.M., Lin, H., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 41, e68.
Chen, W., Feng, P.M., Lin, H., 2014a. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed. Res. Int. 2014, 623149. http://dx.doi.org/10.1155/2014/623149.
Chen, W., Feng, P.M., Deng, E.Z., Lin, H., 2014b. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal. Biochem. 462, 76–83.
Chen, W., Lei, T.Y., Jin, D.C., Lin, H., 2014c. PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. Anal. Biochem. 456, 53–60. http://dx.doi.org/10.1016/j.ab.2014.04.001.
Chen, Y.K., Li, K.B., 2013. Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. J. Theor. Biol. 318, 1–12.
Chou, J.J., 1993. Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach. J. Protein Chem. 12, 291–302.
Chou, K.C., 1983. Identification of low-frequency modes in protein molecules. Biochem. J. 215, 465–469.
Chou, K.C., 1984. Low-frequency vibration of DNA molecules. Biochem. J. 221, 27–31.
Chou, K.C., 1987. The biological functions of low-frequency phonons: 6. A possible dynamic mechanism of allosteric transition in antibody molecules. Biopolymers 26, 285–295.
Chou, K.C., 1988. Review: low-frequency collective motion in biomacromolecules and its biological functions. Biophys. Chem. 30, 3–48.
Chou, K.C., 1989a. Graphic rules in steady and non-steady enzyme kinetics. J. Biol. Chem 264, 12074–12079.
Chou, K.C., 1989b. Low-frequency resonance and cooperativity of hemoglobin. Trends Biochem. Sci. 14, 212–213.
Chou, K.C., 1990. Review: applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophys. Chem. 35, 1–24.
Chou, K.C., 2001a. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct. Funct. Genet. 43, 246–255 (Erratum: ibid., 2001, Vol. 44, 60).
Chou, K.C., 2001b. Using subsite coupling to predict signal peptides. Protein Eng. 14, 75–79.
Chou, K.C., 2005a. Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. J. Proteome Res. 4, 1681–1686.
Chou, K.C., 2005b. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.
Chou, K.C., 2010. Graphic rule for drug metabolism systems. Curr. Drug Metab. 11, 369–378.
Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review). J. Theor. Biol. 273, 236–247.
Chou, K.C., 2013. Some remarks on predicting multi-label attributes in molecular biosystems. Mol. Biosyst. 9, 1092–1100.
Chou, K.C., 2015. Impacts of bioinformatics to medicinal chemistry. Med. Chem.11, 218–234. http://dx.doi.org/10.2174/1573406411666141229162834.
Chou, K.C., Chen, N.Y., 1977. The biological functions of low-frequency phonons. Sci. Sin. 20, 447–457.
Chou, K.C., Mao, B., 1988. Collective motion in DNA and its role in drug intercala-tion. Biopolymers 27, 1795–1815.
Chou, K.C., Zhang, C.T., 1995. Review: prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.
Chou, K.C., Elrod, D.W., 2002. Bioinformatical analysis of G-protein-coupled receptors. J. Proteome Res. 1, 429–433.
Chou, K.C., Cai, Y.D., 2003. Predicting protein quaternary structure by pseudo amino acid composition. Proteins: Struct. Funct. Genet. 53, 282–289.
Chou, K.C., Shen, H.B., 2006a. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem. Biophys. Res. Commun. 347, 150–157.

Chou, K.C., Shen, H.B., 2006b. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J. Proteome Res. 5, 1888–1897.

Chou, K.C., Cai, Y.D., 2006. Predicting protein–protein interactions from sequences in a hybridization space. J. Proteome Res. 5, 316–322.

Chou, K.C., Shen, H.B., 2007a. Review: recent progresses in protein subcellular location prediction. Anal. Biochem. 370, 1–16.

Chou, K.C., Shen, H.B., 2007b. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J. Proteome Res. 6, 1728–1734. http://dx.doi.org/10.1021/pr060635i.

Chou, K.C., Chen, N.Y., Forsen, S., 1981. The biological functions of low-frequency phonons: 2. Cooperative effects. Chem. Scr. 18, 126–132.

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.

Chou, K.C., Lin, W.Z., Xiao, X., 2011. Wenxiang: a web-server for drawing wenxiang diagrams. Nat. Sci. 3, 862–865.

Chou, K.C., Wu, Z.C., Xiao, X., 2012. iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 8, 629–641. http://dx.doi.org/10.1039/c1mb05420a.

Ding, H., Deng, E.Z., Yuan, L.F., Liu, L., Lin, H., 2014. iCTX-Type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed. Res. Int. 2014, 286419. http://dx.doi.org/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. Int. J. Mol. Sci. 15, 3495–3506.

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 composi-tions. Anal. Biochem. 425, 117–119.

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

Espadaler, J., Romero-Isart, O., Jackson, R.M., Oliva, B., 2005. Prediction of protein–protein interactions using distant conservation of sequence patterns and structure relationships. Bioinformatics 21, 3360–3368.

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

Fields, S., Song, O., 1989. A novel genetic system to detect protein–protein interactions. Nature 340, 245–246.

Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., Cruciat, C.-M., 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147.

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. Open Bioinform. J. 7, 41–48.

Gomez, S.M., Noble, W.S., Rzhetsky, A., 2003. Learning to predict protein–protein interactions from protein sequences. Bioinformatics 19, 1875–1881.

Gordon, G., 2008. Extrinsic electromagnetic fields, low frequency (phonon) vibra-tions, and control of cell function: a non-linear resonance system. J. Biomed. Sci. Eng. 1, 152–156.

Grantham, R., 1974. Amino acid difference formula to help explain protein evolution. Science 185, 862–864.

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

Guo, Y., Yu, L., Wen, Z., Li, M., 2008a. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res. 36, 3025–3030. http://dx.doi.org/10.1093/nar/gkn159.

Guo, Y., Yu, L., Wen, Z., Li, M., 2008b. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res. 36, 3025–3030.

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

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. J. Theor. Biol. 344, 31–39.

Han, J.-D.J., Dupuy, D., Bertin, N., Cusick, M.E., Vidal, M., 2005. Effect of sampling on topology predictions of protein–protein interaction networks. Nat. Biotechnol. 23, 839–844.

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 Lett. 19, 411–421.

Hopp, T.P., Woods, K.R., 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. 78, 3824–3828.

Huang, R.B., Du, Q.S., Wang, C.H., 2008. An in-depth analysis of the biological functional studies based on the NMR M2 channel structure of influenza A virus. Biochem. Biophys. Res. Commun. 377, 1243–1247.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y., 2001. A comprehen-sive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. 98, 4569–4574.

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.

Jia, J., Xiao, X., Liu, B., Jiao, L., 2011. Bagging-based spectral clustering ensemble selection. Pattern Recognit. Lett. 32, 1456–1467.

Jiang, Y., Huang, T., Chen, L., Gao, Y.F., Cai, Y., 2013. Signal propagation in protein interaction network during colorectal cancer progression. Biomed. Res. Int., 287019.

Kandaswamy, K.K., Pugalenthi, G., Moller, S., Hartmann, E., Kalies, K.U., Suganthan, P.N., 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 Peptide Lett. 17, 1473–1479.

Kandaswamy, K.K., Chou, K.C., Martinetz, T., Moller, S., Suganthan, P.N., Sridharan, S., Pugalenthi, G., 2011. AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J. Theor. Biol. 270, 56–62.

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 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.

Krigbaum, W.R., Komoriya, A., 1979. Local interactions as a structure determinant for protein molecules: II. Biochim. Biophys. Acta – Protein Struct. 576, 204–228.

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.

Li, Z.C., Zhou, X.B., Dai, Z., Zou, X.Y., 2009. Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis. Amino Acids 37, 415–425.

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 Res. 42, 12961–12972. http://dx.doi.org/10.1093/nar/gku1019.

Lin, S.X., Lapointe, J., 2013. Theoretical and experimental biology in one. J. Biomed. Sci. Eng. 6, 435–442.

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.

Lin, W.Z., Fang, J.A., Xiao, X., 2013. iLoc-animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Mol. BioSyst. 9, 634–644.

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. http://dx.doi.org/10.1093/bioinformatics/btu820.

Liu, B., Fang, L., Liu, F., Wang, X., 2015b. Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS One 10, e0121501.

Liu, B., Xu, J., Lan, X., Xu, R., Zhou, J., 2014a. 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.

Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., Chen, Q., 2014b. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics 30, 472–479.

Liu, H., Wang, M., Chou, K.C., 2005. Low-frequency Fourier spectrum for predicting membrane protein types. Biochem. Biophys. Res. Commun. 336, 737–739.

Madkan, A., Blank, M., Elson, E., Goodman, R., 2009. Steps to the clinic with ELF EMF. Nat. Sci. 1, 157–165.

Mallat, S.G., 1989. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Int. 11, 674–693.

Mallat, S.G., 1999. A Wavelet Tour of Signal Processing. Academic Press, Cambridge, UK.

Marcotte, E.M., Pellegrini, M., Ng, H.-L., Rice, D.W., Yeates, T.O., Eisenberg, D., 1999. Detecting protein function and protein–protein interactions from genome sequences. Science 285, 751–753.

Martin, S., Roe, D., Faulon, J.-L., 2005. Predicting protein–protein interactions using signature products. Bioinformatics 21, 218–226.

Mei, S., 2012. 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 Peptide 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 composi-tion and support vector machine. J. Theor. Biol. 281, 18–23.

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. Med. Chem. 9, 133–137.

Mohammad Beigi, M., Behjati, M., Mohabatkar, H., 2011. Prediction of metallopro-teinase 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, P.P., 2014. Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. J. Theor. Biol. 356, 30–35.

Nanni, L., 2005. Hyperplanes for predicting protein–protein interactions. Neuro-computing 69, 257–263.

Nanni, L., Lumini, A., 2006. An ensemble of K-local hyperplanes for predicting protein–protein interactions. Bioinformatics 22, 1207–1210.

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., Brahnam, S., Lumini, A., 2012a. Wavelet images and Chou's pseudo amino acid composition for protein classification. Amino Acids 43, 657–665.

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.

Nanni, L., Lumini, A., Gupta, D., Garg, A., 2012b. 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.

OuYang, B., Xie, S., Berardi, M.J., Zhao, X.M., Dev, J., Yu, W., Sun, B., Chou, J.J., 2013. Unusual architecture of the p7 channel from hepatitis C virus. Nature 498, 521–525.

Perutz, M.F., 1942. X-ray analysis of haemoglobin. Nature 149, 491–496.

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 Lett. 19, 50–56.

Qiu, J.D., Huang, J.H., Liang, R.P., Lu, X.Q., 2009. Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. Anal. Biochem. 390, 68–73.

Qiu, J.D., Huang, J.H., Shi, S.P., Liang, R.P., 2010. Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform. Protein Peptide Lett. 17, 715–722.

Qiu, J.D., Suo, S.B., Sun, X.Y., Shi, S.P., Liang, R.P., 2011a. OligoPred: a web-server for predicting homo-oligomeric proteins by incorporating discrete wavelet transform into Chou's pseudo amino acid composition. J. Mol. Graph. Model. 30, 129–134.

Qiu, J.D., Sun, X.Y., Suo, S.B., Shi, S.P., Huang, S.Y., Liang, R.P., Zhang, L., 2011b. Predicting homo-oligomers and hetero-oligomers by pseudo-amino acid composition: an approach from discrete wavelet transformation. Biochimie 93, 1132–1138.

Qiu, W.R., Xiao, X., Chou, K.C., 2014a. iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int. J. Mol. Sci. 15, 1746–1766.

Qiu, W.R., Xiao, X., Lin, W.Z., 2014b. iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model. J. Biomol. Struct. Dyn. http://dx.doi.org/10.1080/07391102.2014.968875.

Qiu, W.R., Xiao, X., Lin, W.Z., 2014c. iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. Biomed. Res. Int. 2014, 947416.

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.

Schnell, J.R., Chou, J.J., 2008. Structure and mechanism of the M2 proton channel of influenza A virus. Nature 451, 591–595.

Shen, H.B., Chou, K.C., 2006. Ensemble classifier for protein fold pattern recognition. Bioinformatics 22, 1717–1722.

Shen, H.B., Chou, K.C., 2007. Using ensemble classifier to identify membrane protein types. Amino Acids 32, 483–488.

Shen, H.B., Chou, K.C., 2009a. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. Anal. Biochem. 394, 269–274.

Shen, H.B., Chou, K.C., 2009b. QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. J. Proteome Res. 8, 1577–1584.

Shen, H.B., Yang, J., Chou, K.C., 2006. Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. J. Theor. Biol. 240, 9–13.

Shen, H.B., Yang, J., Chou, K.C., 2007a. Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33, 57–67.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., Jiang, H., 2007b. Predicting protein–protein interactions based only on sequences information. Proc. Natl. Acad. Sci. 104, 4337–4341.

Sobell, H.M., Baberjee, A., Lozansky, E.D., Zhou, G.P., 1983. The role of low frequency (acoustic) phonons in etermining the premelting and melting behaviors of DNA. In: Clementi, E., Sarma, R.H. (Eds.), Structure and Dynamics: Nucleic Acids and Proteins. Adenine Press, New York, pp. 181–195.

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. Mol. BioSyst. 8, 3178–3184. http://dx.doi.org/10.1039/c2mb25280e.

Tanford, C., 1962. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. J. Am. Chem. Soc. 84, 4240–4247.

Wang, J.F., Chou, K.C., 2009. Insight into the molecular switch mechanism of human Rab5a from molecular dynamics simulations. Biochem. Biophys. Res. Commun. 390, 608–612.

Wang, M., Yang, J., Xu, Z.J., 2005. SLLE for predicting membrane protein types. J. Theor. Biol. 232, 7–15.

Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., Eisenberg, D., 2002. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 30, 303–305.

Xia, J.-F., Zhao, X.-M., Huang, D.-S., 2010a. Predicting protein–protein interactions from protein sequences using meta predictor. Amino Acids 39, 1595–1599.

Xia, J.-F., Han, K., Huang, D.-S., 2010b. Sequence-based prediction of protein–protein interactions by means of rotation forest and autocorrelation descriptor. Protein Peptide Lett. 17, 137–145.

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. J. Theor. Biol. 284, 42–51.

Xiao, X., Min, J.L., Wang, P., 2013a. iGPCR-drug: a web server for predicting interaction between GPCRs and drugs in cellular networking. PLoS One 8, e72234.

Xiao, X., Wang, P., Lin, W.Z., Jia, J.H., 2013b. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal. Biochem. 436, 168–177.

Xu, R., Zhou, J., Liu, B., He, Y.A., Zou, Q., 2014a. Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. J. Biomol. Struct. Dyn. http://dx.doi.org/10.1080/07391102.2014.968624.

Xu, Y., Ding, J., Wu, L.Y., 2013a. 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.

Xu, Y., Shao, X.J., Wu, L.Y., Deng, N.Y., 2013b. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ 1, e171.

Xu, Y., Wen, X., Wen, L.S., Wu, L.Y., 2014b. iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS One 9, e105018.

Yang, L., Xia, J.-F., Gui, J., 2010. Prediction of protein–protein interactions from protein sequence using local descriptors. Protein Peptide Lett. 17, 1085–1090.

Zhang, C.T., Chou, K.C., 1992. An optimization approach to predicting protein structural class from amino acid composition. Protein Sci. 1, 401–408.

Zhang, J., Zhao, X., Sun, P., Ma, Z., 2014a. 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., 2014b. 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.

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.

Zhong, W.Z., Zhou, S.F., 2014. Molecular science for drug development and biomedicine. Int. J. Mol. Sci. 15, 20072–20078.

Zhou, G.P., 1989. Biological functions of soliton and extra electron motion in DNA structure. Phys. Scr. 40, 698–701. http://dx.doi.org/10.1088/0031-8949/40/5/021.

Zhou, G.P., 1998. An intriguing controversy over protein structural class prediction. J. Protein Chem. 17, 729–738.

Zhou, G.P., 2011a. 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, G.P., 2011b. The structural determinations of the leucine zipper coiled-coil domains of the cGMP-dependent protein kinase I alpha and its interaction with the myosin binding subunit of the myosin light chains phosphase. Proteins Peptide Lett. 18, 966–978.

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

Zhou, G.P., Assa-Munt, N., 2001. Some insights into protein structural class prediction. Proteins: Struct. Funct. Genet. 44, 57–59. http://dx.doi.org/10.1002/prot.1071.

Zhou, G.P., Doctor, K., 2003. Subcellular location prediction of apoptosis proteins. Proteins: Struct. Funct. Genet. 50, 44–48. http://dx.doi.org/10.1002/prot.10251.

Zhou, G.P., Huang, R.B., 2013. The pH-triggered conversion of the PrP(c) to PrP(sc.). Curr. Top. Med. Chem. 13, 1152–1163 CTMC-EPUB-20130503-3.

Zhou, P., Tian, F., Li, B., Wu, S., Li, Z., 2006. Genetic algorithm-based virtual screening of combinative mode for peptide/protein. Acta Chim. Sin. – Chin. Ed. 64, 691.

Zhou, X., Li, Z., Dai, Z., Zou, X., 2013. Predicting promoters by pseudo-trinucleotide compositions based on discrete wavelets transform. J. Theor. Biol. 319, 1–7.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., 2001. Global analysis of protein activities using proteome chips. Science 293, 2101–2105.

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 Peptide Lett. 19, 890–903.

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's pseudo amino acid patterns. Anal. Biochem. 458, 14–19.