molecules

MDPI

Article

iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets

Jianhua Jia [1,2,*], Zi Liu [1], Xuan Xiao [1,2,*], Bingxiang Liu [1] and Kuo-Chen Chou [2,3]

Received: 18 November 2015 ; Accepted: 7 January 2016 ; Published: 19 January 2016
Academic Editor: Derek J. McPhee

[1] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China; liuzi189836@163.com (Z.L.); lbx1966@163.com (B.L.)
[2] Gordon Life Science Institute, Boston, MA 02478, USA; kcchou@gordonlifescience.org
[3] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia
[*] Correspondence: jjia@gordonlifescience.org (J.J.); xxiao@gordonlifescience.org (X.X.); Tel.: +86-1397-9866-529 (J.J.); +86-1387-9809-729 (X.X.)

Abstract: Knowledge of protein-protein interactions and their binding sites is indispensable for in-depth understanding of the networks in living cells. With the avalanche of protein sequences generated in the postgenomic age, it is critical to develop computational methods for identifying in a timely fashion the protein-protein binding sites (PPBSs) based on the sequence information alone because the information obtained by this way can be used for both biomedical research and drug development. To address such a challenge, we have proposed a new predictor, called **iPPBS-Opt**, in which we have used: (1) the K-Nearest Neighbors Cleaning (KNNC) and Inserting Hypothetical Training Samples (IHTS) treatments to optimize the training dataset; (2) the ensemble voting approach to select the most relevant features; and (3) the stationary wavelet transform to formulate the statistical samples. Cross-validation tests by targeting the experiment-confirmed results have demonstrated that the new predictor is very promising, implying that the aforementioned practices are indeed very effective. Particularly, the approach of using the wavelets to express protein/peptide sequences might be the key in grasping the problem's essence, fully consistent with the findings that many important biological functions of proteins can be elucidated with their low-frequency internal motions. To maximize the convenience of most experimental scientists, we have provided a step-by-step guide on how to use the predictor's web server (http://www.jci-bioinfo.cn/iPPBS-Opt) to get the desired results without the need to go through the complicated mathematical equations involved.

Keywords: protein-protein binding sites; physicochemical property; stationary wavelet transform; PseAAC; Optimize training dataset; KNNC; IHTS; target cross-validation

1. Introduction

Individual proteins rarely function alone. Most proteins whose functions are essential to life are associated with protein-protein interactions [1]. Actually, these kinds of interactions affect the biological processes in a living cell. To really understand protein-protein interactions, however, it is indispensable to acquire the information of protein-protein binding site (PPBS). Despite many studies on the binding site of a protein or DNA with its ligand (small molecule) have been made [2–8], relatively much less studies have been conducted on PPBS, particularly based on the sequence information alone. It is both time-consuming and expensive to determine PPBS purely based on biochemical experiments. Facing the enormous number of protein sequences generated in the postgenomic era, it is highly desired to develop computational methods to identify PPBSs for uncharacterized proteins so that they can be timely used for both basic research and drug development, such as conducting mutagenesis studies [9] and prioritize drug targets.

Given a protein sequence, how can one identify which of its constituent amino acid residues are located in the binding sites? Actually, considerable efforts were made to address this problem [10,11]. Although the aforementioned works each have their own merits and did play a role in stimulating the development of this area, further work is needed due to the following shortcomings: (1) The datasets used by these authors to train their prediction methods were highly imbalanced or with a strong bias; *i.e.*, the number of non-PPBS samples was significantly larger than that of PPBS samples; (2) None of their prediction methods has a publicly accessible web server, and hence their practical application value is quite limited, particularly for the majority of experimental scientists.

The present study is initiated in an attempt to develop a new PPBS predictor by addressing the aforementioned shortcomings. According to the Chou's 5-step rule [12] and the demonstrations in a series of recent publications [13–20], to establish a really useful sequence-based statistical predictor for a biological system, we should make the following five aspects crystal 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; (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 Dataset

Two benchmark datasets were used for the current study. One is the "surface-residue" dataset and the other is "all-residue" dataset, as described 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 [21]. 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, R_2 the 2nd residue, and so forth. The residue R_i $(i = 1, 2, \cdots, L)$ is deemed as a surface residue if it satisfies the following condition:

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

where $\mathrm{ASA}(\mathrm{R}_i\,|\,\mathrm{P})$ is the accessible surface area 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 accessible surface area as given in Table 1 [22], and $\phi\,(\mathrm{R}_i)$ is the ratio of the two.

Table 1. Maximum accessible surface area (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	

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

Furthermore, the surface residue R_i is deemed as interfacial residue [23] if:

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

where ASA(R_i | **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 [24] to find out all its surface residues based on Equation (2), and use PSAIA program [25] to find all its interfacial residues based on Equation (3).

If only considering the surface residues as done in [26] for the 99 polypeptide chains extracted by Deng *et al.* [10] from the 54 heterocomplexes in the Protein Data Bank, we have obtained the results that can be formulated as follows:

$$\mathbb{S}_{\mathrm{surf}} = \mathbb{S}^{+}_{\mathrm{surf}} \bigcup \mathbb{S}^{-}_{\mathrm{surf}} \tag{4}$$

where $\mathbb{S}_{\mathrm{surf}}$ is called the "surface-residue dataset" that contains a total of 13,771 surfaces residues, of which 2828 are interfacial residues belonging to the positive subset $\mathbb{S}^{+}_{\mathrm{surf}}$ while 10,943 are non-interfacial residues belonging the negative subset $\mathbb{S}^{-}_{\mathrm{surf}}$, and $\bigcup$ is the symbol of union in the set theory.

If considering all the residues as done in [11], however, the corresponding benchmark dataset can be expressed by:

$$\mathbb{S}_{\mathrm{all}} = \mathbb{S}^{+}_{\mathrm{all}} \bigcup \mathbb{S}^{-}_{\mathrm{all}} \tag{5}$$

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

For readers' convenience, given in S1 Dataset (List of the 99 proteins and their residues' attributions associated with the protein-protein binding sites is in Supplementary Materials) is a combination of the two benchmark datasets, 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 [27] there is no need to separate a benchmark dataset into a training dataset and a testing dataset 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 obtained via this kind of approach is actually from a combination of many different independent dataset tests.

2.2. Flexible Sliding Window Approach

Given a protein chain as expressed in Equation (1), the sliding window approach [28] and flexible sliding window approach [29] are often used to investigate its various posttranslational modification (PTM) sites [16,30–34] and HIV (human immunodeficiency virus) protease cleavage sites [35]. Here, we also use it to study protein-protein binding sites. In the sliding window approach, a scaled window is denoted by [$-\xi$, $+\xi$] [28], and its width is $2\xi + 1$, where ξ is an integer. When sliding it along a protein chain **P**, one can see through the window a series of consecutive peptide segments as formulated by:

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

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 $\mathrm{L} - \xi$, the corresponding $\mathbf{P}_{\xi}\left(\mathbb{R}_0\right)$ is defined, rather than by **P** of Equation (1), but by the following dummy protein chain:

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

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**, we have a working protein segment as defined by Equation (6). In the current study, the $(2\xi+1)$-tuple peptide $\mathbf{P}_\xi\left(\mathbb{R}_0\right)$ can be further classified into the following categories:

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

where ∈ represents "a member of" in the set theory.

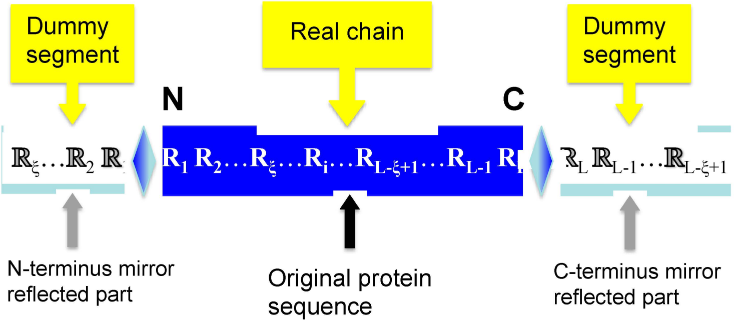

Dummy segment
Real chain
Dummy segment
N
C
ℝ_ξ…ℝ_2 ℝ
R_1 R_2…R_ξ…R_i…R_L-ξ+1…R_L-1 R
ℝ_L ℝ_L-1…ℝ_L-ξ+1
N-terminus mirror reflected part
Original protein sequence
C-terminus mirror reflected part

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, and DNA/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 [36]; (2) All the existing computational algorithms, such as optimization approach [37], correlation-angle approach [38], covariance discriminant (CD) [39], neural network [40], K-nearest neighbor (KNN) [41], OET-KNN [42], SLLE algorithm [43], random forest [44], Fuzzy K-nearest neighbor [45], and ML-KNN algorithm [46], can only handle vector but not sequence samples.

However, a vector defined in a discrete model may completely lose the sequence-order information as elaborated in [47,48]. To cope with such a dilemma, the approach of pseudo amino acid composition [36,49] or Chou's PseAAC [50,51] was proposed. Ever since it was introduced in 2001 [36], the concept of PseAAC has been penetrating into nearly all the areas of computational biology (see, e.g., [52–56] as well as a long list of references cited in [48,57] and a recent review [58]). It has also been selected as a special topic for a special issue on "drug development and biomedicine" [59]. Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA and nucleotides [60–64]. Because of its being widely and increasingly used, three types of open access soft-ware, called "PseAAC-Builder" [65], "propy" [50], and "PseAAC-General" [57], 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 [12], PseAAC can be generally formulated as:

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

where **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 datasets (*cf.* Equations (4) and (5)) to define the working protein segments via Equation (9). For the convenience of formulation below, we convert the $(2\xi + 1)$-tuple 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 [66] or $\Phi^{(1)}$; (2) hydrophilicity [67] or $\Phi^{(2)}$; (3) side-chain volume [68] or $\Phi^{(3)}$; (4) polarity [69] or $\Phi^{(4)}$; (5) polarizability [70] or $\Phi^{(5)}$; (6) solvent-accessible surface area (SASA) [71] or $\Phi^{(6)}$; and (7) side-chain net charge index (NCI) [72] or $\Phi^{(7)}$. Their numerical values are given in Table 2.

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

Amino Acid Code	**Physicochemical Property (*cf.* Equation (11))** [a]						
	$\mathbf{\Phi}^{(1)}$	$\mathbf{\Phi}^{(2)}$	$\mathbf{\Phi}^{(3)}$	$\mathbf{\Phi}^{(4)}$	$\mathbf{\Phi}^{(5)}$	$\mathbf{\Phi}^{(6)}$	$\mathbf{\Phi}^{(7)}$
	H1	**H2**	**V**	**P1**	**P2**	**SASA**	**NCI**
A	0.62	−0.5	27.5	8.1	0.046	1.181	0.007187
C	0.29	−1	44.6	5.5	0.128	1.461	−0.03661
D	−0.9	3	40	13	0.105	1.587	−0.02382
E	−0.74	3	62	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	9	0	0.881	0.179052
H	−0.4	−0.5	79	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	100	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	58.7	11.6	0.134	1.655	0.005392
P	0.12	0	41.9	8	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	105	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.

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 (10), $\Phi_2^{(2)}$ 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:

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

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 through 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., [73–75]. Many marvelous biological functions in proteins and DNA and their profound dynamic mechanisms, such as switch between active and inactive states [76,77], cooperative effects [78], allosteric transition [79–81], intercalation of drugs into DNA [82], and assembly of microtubules [83], can be revealed by studying their low-frequency internal motions as summarized in a comprehensive review [84]. Low frequency Fourier spectrum was also used by Liu *et al.* [85] 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 into the current study.

The stationary wavelet transform (SWT) [86] is a wavelet transform algorithm designed to overcome the lack of shift-invariance of the discrete wavelet transform (DWT) [87]. 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 j-th 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 coefficients for the ensemble learning framework that will be introduced later.

Preliminary tests indicated that, when $\xi = 7$, *i.e.*, the working segments are 15-tuple peptides, the outcomes thus obtained were most promising. Accordingly, we only consider the case of $\xi = 7$ hereafter.

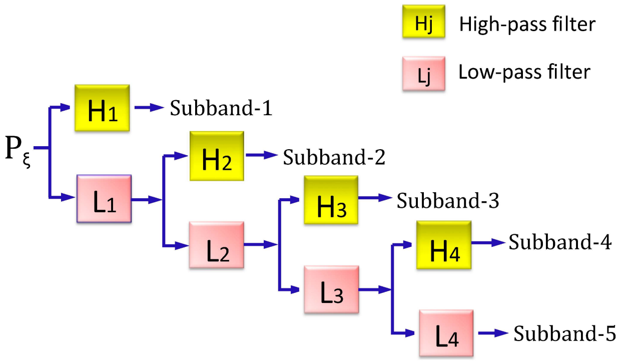


Figure 2. A schematic drawing to illustrate the procedure of multi-level SWT (stationary wavelets transform). See Equations (10)–(12) as well as the relevant text for further explanation.

Using the SWT approach, we have generated five sub-bands (Figure 2), each of which has four coefficients: (1) α_i, the maximum of the wavelet coefficients in the sub-band i $(1, 2, \cdots 5)$; (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:

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

where:

$$\Psi_\mu^{(k)} = \begin{cases} \alpha_\mu^{(k)} & \text{when } 1 \leqslant \mu \leqslant 5 \\ \beta_{\mu-5}^{(k)} & \text{when } 6 \leqslant \mu \leqslant 10 \\ \lambda_{\mu-10}^{(k)} & \text{when } 11 \leqslant \mu \leqslant 15 \\ \delta_{\mu-15}^{(k)} & \text{when } 11 \leqslant \mu \leqslant 20 \end{cases} \tag{14}$$

2.4. Optimizing Imbalanced Training Datasets

In the current benchmark dataset $\mathbb{S}_{\text{surf}}$ or $\mathbb{S}_{\text{all}}$, the negative subset $\mathbb{S}_{\text{all}}^-$ or $\mathbb{S}_{\text{surf}}^-$ is much larger than the corresponding positive subset $\mathbb{S}_{\text{all}}^+$ or $\mathbb{S}_{\text{surf}}^+$ as can be seen by the following equation:

$$\begin{cases} \mathbb{S}_{\text{surf}}\,(13771) = \mathbb{S}_{\text{surf}}^+\,(2828) \bigcup \mathbb{S}_{\text{surf}}^-\,(10943) & \text{for surface residues} \\ \mathbb{S}_{\text{all}}\,(27442) = \mathbb{S}_{\text{all}}^+\,(2828) \bigcup \mathbb{S}_{\text{all}}^-\,(24614) & \text{for all residues} \end{cases} \tag{15}$$

where the figures in the parentheses denote the sample numbers taken from Section 2.1. As we can see from Equation (15), the numbers of the negative samples are nearly nine and four times the sizes of the corresponding positive samples for the all-residue and surface-residue benchmark datasets, respectively.

Although this might reflect the real world in which the non-binding sites are always the majority compared with the binding ones, a predictor trained by such a highly skewed benchmark dataset would inevitably have the bias consequence that many binding sites might be mispredicted as non-binding ones [88]. Actually, what is really the most intriguing information for us is the information about the binding sites. Therefore, it is important to find an effective approach to optimize the unbalanced training dataset and minimize this kind of bias consequence. To realize this, we took the following procedures.

First, we used the K-Nearest Neighbors Cleaning (KNNC) treatment to remove some redundant negative samples from the negative subset so as to reduce its statistical noise. The detailed process can be described below: (i) for each of the samples in the negative subset $\mathbb{S}^-$ find its K nearest neighbors, where K may be any integer (such as 3 or 8), and its final value will be discussed later; (ii) if one of its K nearest neighbors belongs to the positive subset $\mathbb{S}^+$, remove the negative sample from $\mathbb{S}^-$. A similar method, called the Neighborhood Cleaning Rule (NCR), was also been used by Laurikkala *et al.* [89], Xiao *et al.* [90], and Liu *et al.* [91] although their details are different with the current practice. Also, the current KNNC approach is more flexible because it contains a variable K and hence can be used to deal with various different training datasets.

Second, we used the Inserting Hypothetical Training Samples (IHTS) treatment to add some hypothetical positive samples into the positive subset so as to enhance the ability in identifying the interactive pairs. For the details of how to generate the hypothetical training samples, see the Monte Calo samples expanding approach in [92,93], or seed-propagation approach in [94], or the SMOTE (synthetic minority over-sampling technique) approach in [95].

After the above two treatments, we can change an original highly skewed training dataset to a balanced training dataset with its positive subset and negative subset having exactly the same size.

It is instructive to point out that the hypothetical samples generated via the IHTS treatment can only be expressed by their feature vectors as defined in Equation (13), but not the real peptide segment samples as given by Equations (6) or (10). Nevertheless, it would be perfectly reasonable to do so because the data directly used to train a predictor were actually the samples' feature vectors but not their sequence codes. This is the key to optimize an imbalanced benchmark dataset in the current study, and the rationale of such an interesting approach will be further elucidated later.

2.5. Fusing Multiple Physicochemical Properties

The random forest (RF) algorithm is a powerful algorithm, which has been used in many areas of computational biology (see, e.g., [44,96,97]). The detailed procedures and formulation of RF have been very clearly described in [98], and hence there is no need to repeat here.

As shown in Equations (11)–(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 random forest predictor after the KNNC and IHTS procedures. Accordingly, we have a total of seven individual predictors for identifying PPBS, as formulated by:

$$\text{PPBS individual predictor} = \mathbb{RF}(k) \quad (k = 1,\ 2,\ \cdots,\ 7) \tag{16}$$

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

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 [99,100] and protein quaternary structural attribute [101]. Encouraged by the previous investigators' studies, here we are also developing an ensemble classifier by fusing the seven individual predictors $\mathbb{RF}(k)\ (k = 1, 2, \cdots, 7)$ through a voting system, as formulated by:

$$\mathbb{RF}^{\mathrm{E}} = \mathbb{RF}(1) \forall \cdots \forall \mathbb{RF}(7) = \forall_{k=1}^{7} \mathbb{RF}(k) \tag{17}$$

where $\mathbb{RF}^{\mathrm{E}}$ stands for the ensemble classifier, and the symbol ∀ 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 [27], 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 3 to illustrate how the seven individual RF predictors are fused into the ensemble classifier.

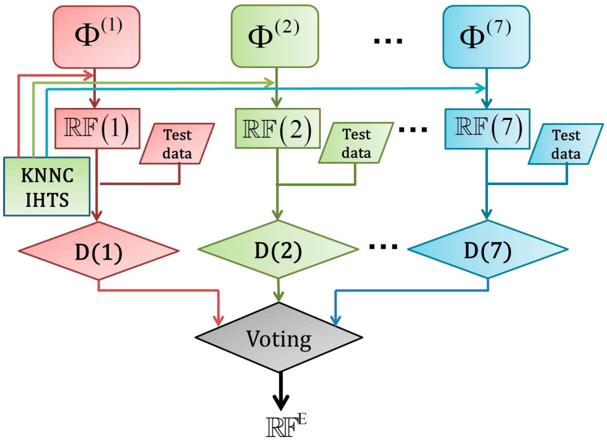


Figure 3. A flowchart to illustrate the ensemble classifier of Equation (17) that exploits all the different groups of features, where D(1) means the decision made by $\mathbb{RF}(1)$, D(2) means the decision made by $\mathbb{RF}(2)$, and so forth. See the text as well as Equations (11) and (16) for further explanation.

The final predictor thus obtained is called "**iPPBS-Opt**", where "i" stands for "identify", "PPBS" for "protein-protein binding site", and "Opt" for "optimizing" training datasets. Note that the **iPPBS-Opt** predictor contains a parameter K, reflecting how many nearest neighbors should be considered in removing the redundant negative samples from the training dataset during the KNNC treatment (*cf.* Section 2.4). Its final value is determined by maximizing the overall success rate via cross-validation, as will be described later.

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 [12]. Towards 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. Metrics for Measuring Success Rates

For measuring the success rates in identifying PPBS, a set of four metrics are usually 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., [102]). 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 [103] for studying signal peptides, the aforementioned four metrics can be formulated by a set of equations given below [14,30,60,61,104]:

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \leqslant \mathrm{Sn} \leqslant 1 \\ \mathrm{Sp} = 1 - \dfrac{N_+^-}{N^-} & 0 \leqslant \mathrm{Sp} \leqslant 1 \\ \mathrm{Acc} = \Lambda = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \leqslant \mathrm{Acc} \leqslant 1 \\ \mathrm{Mcc} = \dfrac{1 - \left(\dfrac{N_-^+ + N_+^-}{N^+ + N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \leqslant \mathrm{Mcc} \leqslant 1 \end{cases} \tag{18}$$

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 (18), 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 Sn = 1. When $N_-^+ = N^+$ meaning that all the PPBSs are incorrectly predicted to be of non-PPBS, we have the sensitivity Sn = 0. Likewise, when $N_+^- = 0$ meaning none of the non-PPBSs are incorrectly predicted to be of PPBS, we have the specificity Sp = 1; whereas $N_+^- = N^-$ meaning that all the non-PPBSs are incorrectly predicted to be of PPBS, we have the specificity Sp = 0. When $N_-^+ = N_+^- = 0$ meaning that none of PPBSs in the positive dataset and none of the non-PPBSs in the negative dataset are incorrectly predicted, we have the overall accuracy Acc = 1 and MCC = 1; when $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the PPBSs in the positive dataset and all the non-PPBSs in the negative dataset are 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 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 (18), particularly for the meaning of MCC.

It should be pointed out, however, the set of metrics as defined in Equation (18) is valid only for the single-label systems. For the multi-label systems whose emergence has become more frequent in system biology [46,105,106] and system medicine [107], a completely different set of metrics as defined in [108] is needed.

3.2. Cross-Validation and Target Cross-Validation

Once established the evaluation metrics, the next issue is the selection of the most appropriate validation method should be used to derive the values of these metrics. Three cross-validation methods are often used to derive metrics values in statistical prediction: the independent dataset test, subsampling (or K-fold cross-validation) test, and jackknife test [109]. Of the three the jackknife test is deemed the least arbitrary as it can always yield a unique outcome for a given benchmark dataset, as elucidated in [12] 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., [46,53,54,110–115]. However, to reduce the computational time, in this study we adopted the 10-fold cross-validation, as done by most investigators with SVM and random forests algorithms as the prediction engine.

When conducting the 10-fold cross-validation for the current predictor **iPPBS-Opt,** however, some special consideration is needed. This is because a dataset, after optimized by the KNNC and ITHTS treatments, may miss many experimental negative samples and contain some hypothetical positive samples. It would be fine to use such a dataset to train a predictor, but not for validation. Since the validation should be conducted based on all the experimental data in the benchmark dataset but not on the added hypothetical samples nor only on the data in the reduced negative subset, a special cross-validation, the so-called target cross-validation, has been introduced here. During the target cross-validation process for the positive samples, only the experiment-confirmed samples are singled out as the targets (or test samples) for validation; but during the target cross-validation process for the negative samples, even all the excluded experimental data are taken into account. The detailed procedures of the target 10-fold cross-validation are as follows:

Step 1. Before optimizing the original benchmark dataset, both its positive and negative subsets were randomly divided into 10 parts with about the same size. For example, for the all-residue benchmark dataset $\mathbb{S}_{\text{all}}$, after such evenly division we have:

$$\mathbb{S}_{\text{all}} = \mathbb{S}_{\text{all}(1)} \bigcup \mathbb{S}_{\text{all}(2)} \bigcup \cdots \bigcup \mathbb{S}_{\text{all}(10)} = \bigcup_{i=1}^{10} \mathbb{S}_{\text{all}(i)} \tag{19}$$

and:

$$\mathbb{S}_{\text{all}(1)} \triangleq \mathbb{S}_{\text{all}(2)} \triangleq \cdots \triangleq = \mathbb{S}_{\text{all}(10)} \tag{20}$$

where the symbol $\triangleq$ means that the divided 10 datasets are about the same in size, and so are their subsets.

Step 2. One of the 10 sets, say $\mathbb{S}_{\text{all}(1)}$, was singled out as the testing dataset and the remaining nine sets as the training dataset.

Step 3. The training set was optimized using the KNNC and IHTS treatments as described in Section 2.4. After such a process, the original imbalanced training dataset would become a balanced one; *i.e.*, its positive subset and negative subset would contain a same number of samples. Note that although the starting value for K in the KNNC treatment could be arbitrary, the following empirical approach might be of help to reduce the time for finally finding its optimal value. Suppose the starting value for K is $K\,(0)$, then we have according to our experience

$$K\,(0) = \text{Int}\left[\frac{N^-}{N^+}\right] \tag{21}$$

where N^{+} and N^{-} are the numbers of the total positive and negative samples in the benchmark dataset, respectively, and Int is the "integer truncation operator" meaning to take the integer part for the number in the brackets right after it [116]. Substituting the data of Equation (15) into Equation (21), we obtained $K\,(0) = 3$ or 8 for the surface-residue case or of all-residue case, respectively.

Step 4. Use the aforementioned balanced dataset to train the operation engine, followed by applying the **iPPBS-Opt** predictor to calculate the prediction scores for the testing dataset, which had been singled out in Step 2 before the optimized treatment and hence contained the experiment-confirmed samples only.

Step 5. Repeat Steps 2–4 until all the 10 divided sets had been singled out one-by-one for testing validation.

Step 6. Substituting the scores obtained from the above 10-round tests into Equation (18) to calculate Sn, Sp, Acc, and MCC. The metrics values thus obtained should be a function of *K*; for instance, the overall accuracy Acc can be expressed as Acc(*K*).

Step 7. Repeat Steps 2–6 by increasing *K* with a gap of 1, we consecutively obtained Acc(3), Acc(4), ..., Acc(12) for the surface-residue case or Acc(8), Acc(9), ..., Acc(17) for the all-residue case, respectively (Figure 4). The value of *K* that maximized Acc would be taken for **iPPBS-Opt** in the current study, as given in the footnote c of Table 3.

It is instructive to emphasize again that it is absolutely fair to use the above 10-fold cross-validation steps to compare the current predictor with the existing ones. This is because all the predictors concerned were tested using exactly the same experiment-confirmed samples and that all the added hypothetical samples had been completely excluded from the testing datasets.

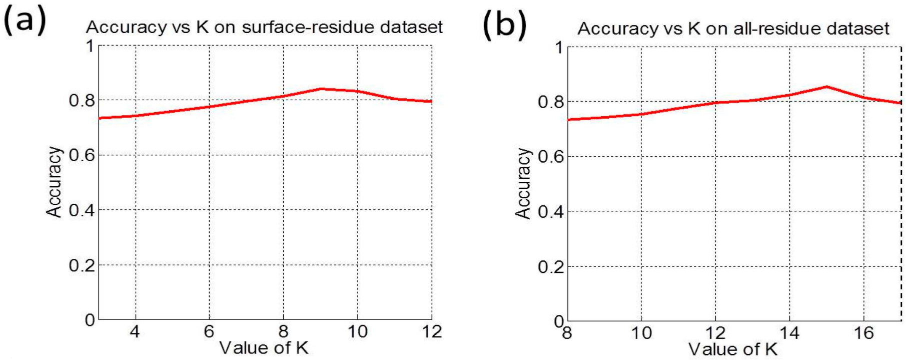


Figure 4. A plot of Acc *vs.* *K* for (**a**) the surface-residue benchmark dataset (*cf.* Equation (4)); and (**b**) the all-residue benchmark dataset (*cf.* Equation (5)). It can be seen from panel (**a**) that the overall accuracy reaches its peak at $K = 9$, and from panel (**b**) that the overall accuracy reaches its peak at $K = 15$.

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

Benchmark Dataset	Method	Acc (%)	MCC	Sn (%)	Sp (%)	AUC
Surface-residue	Deng [a]	N/A	0.3456	76.77	63.16	0.7976
	Chen [b]	75.09	0.4248	43.81	92.12	0.8004
	iPPBS-PseAAC [c]	**84.04**	**0.5821**	58.26	94.14	**0.8934**
All-residue	Deng [a]	N/A	0.3763	76.33	78.61	0.8465
	Chen [b]	73.77	0.3286	24.95	96.52	0.8001
	iPPBS-PseAAC [c]	**85.45**	**0.4662**	39.14	96.66	**0.8820**

[a] Results reported by Deng *et al.* [10]; [b] Results reported by Chen *et al.* [11]; [c] Results obtained on the same testing dataset by the current predictor **iPPBS-Opt** with its parameter $K = 9$ for the surface-residue benchmark dataset $\mathbb{S}_{\text{surf}}$ (*cf.* Equation (4)) and $K = 15$ for the all-residue benchmark dataset $\mathbb{S}_{\text{all}}$ (*cf.* Equation (5)). Also see Figure 4 for the details.

3.3. Comparison with the Existing Methods

Listed in Table 3 are the values of the four metrics (*cf.* Equation (18)) obtained by the current **iPPBS-Opt** predictor using the target 10-fold cross-validation on the surface-residue benchmark dataset $\mathbb{S}_{\text{surf}}$ (Equation (4)) and the all-residue benchmark dataset $\mathbb{S}_{\text{all}}$ (Equation (5)), respectively. See S1 Dataset for the details of the two benchmark datasets. For facilitating comparison, the corresponding results obtained by the existing methods [10,11] are also given there.

As we can see from the table, the new predictor **iPPBS-Opt** 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 [10] is higher than that of the current predictor when tested by the surface-residue benchmark dataset, its corresponding Sp value is more than 30% lower than that of the latter, indicating the method [10] is very unstable with extremely high noise.

Because graphic approaches can provide useful intuitive insights (see, e.g., [117–122]), here we also provide a graphic comparison of the current predictor with their counterparts via the Receiver Operating Characteristic (ROC) plot [123], as shown in (Figure 5). According to ROC [123], 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.

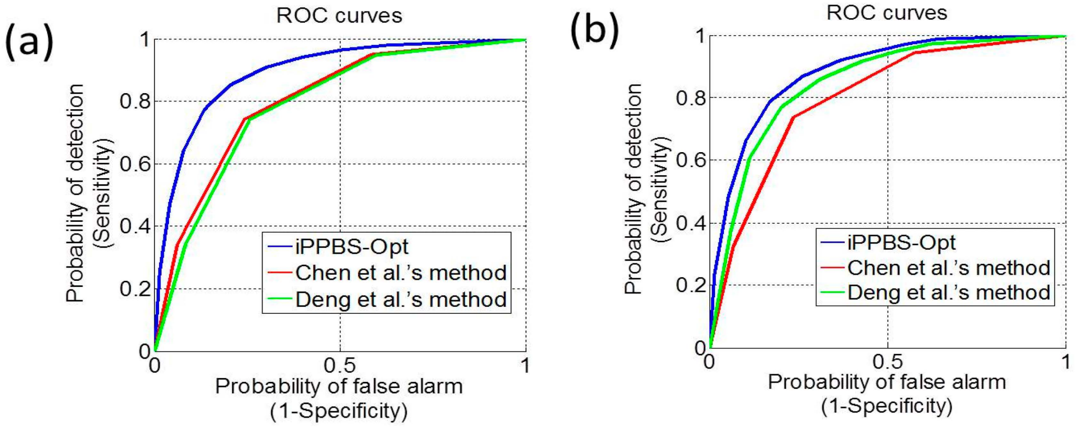


Figure 5. The ROC (Receiver Operating Characteristic) curves to show the 10-fold cross validation by **iPPBS-Opt**, Deng *et al.*'s method [10], and Chen *et al.*'s method [11] on (**a**) surface-residue benchmark dataset; and (**b**) the all-residue benchmark dataset. As shown on the figure, the area under the ROC curve for **iPPBS-Opt** is obviously larger than those of their counterparts, indicating a clear improvement of the new predictor in comparison with the existing ones.

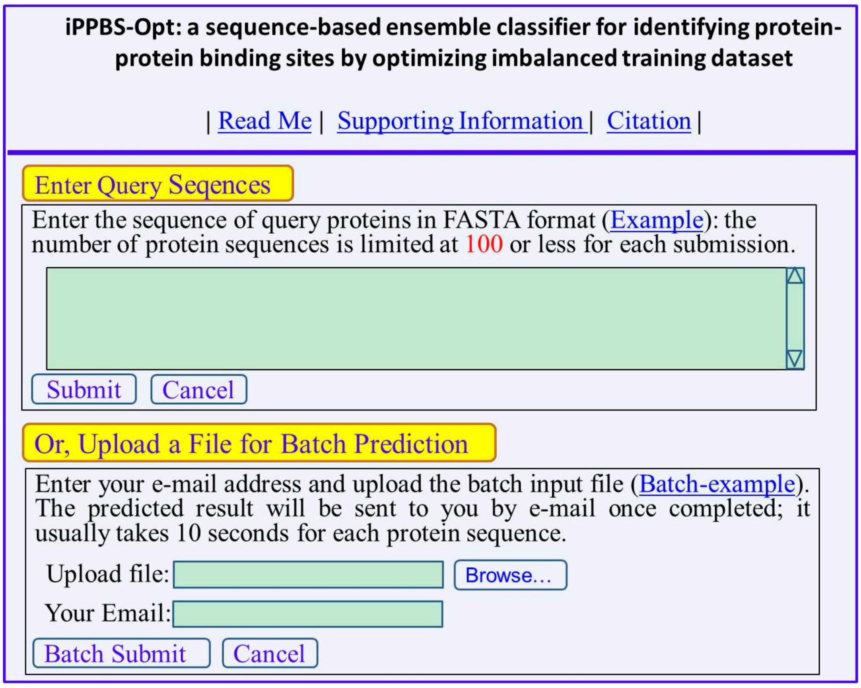
All the above facts have shown that **iPPBS-Opt** is really a very promising predictor for identifying protein-protein binding sites. Or at the very least, it can play a complementary role to the existing prediction methods in this area. Particularly, none of the existing predictors has provided a web server. In contrast to this, a user-friendly and publically accessible web server has been established for **iPPBS-Opt** at http://www.jci-bioinfo.cn/iPPBS-Opt, 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? The reasons are as follows: First, the KNNC and IHTS treatments have been introduced to optimize the training datasets, so as to avoid many misprediction events caused by the highly imbalanced training datasets used in previous studies. Second, the ensemble technique has been utilized in this study to select the most relevant one from seven classes of different physicochemical properties. Third, the wavelets transform technique has been applied to extract some important key features, which are deeply hidden in complicated protein sequences. This is just like the studies in dealing with the extremely complicated internal motions of proteins, it is the key to grasp the low-frequency collective motion [74,75] for in-depth understanding or revealing the dynamic mechanisms of their various important biological functions [84], such as cooperative effects [78], allosteric transition [80,81], assembly of microtubules [83], and switch between active and inactive states [76]. Fourth, the PseAAC approach has been introduced to formulate the statistical samples, which has been proved very useful not only in dealing with protein/peptide sequences, but also in dealing with DNA/RNA sequences, as elaborated in a recent review paper [124].

3.4. Web Server and User Guide

To enhance the value of its practical applications, a web-server for **iPPBS-Opt** has been established at http://www.jci-bioinfo.cn/iPPBS-Opt. 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/iPPBS-Opt, you will see the top page of **iPPBS-Opt** on your computer screen, as shown in Figure 6. Click on the Read Me button to see a brief introduction about the i**PPBS-Opt** predictor.



iPPBS-Opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training dataset

Enter the sequence of query proteins in FASTA format (Example): the number of protein sequences is limited at 100 or less for each submission.

Enter your e-mail address and upload the batch input file (Batch-example). The predicted result will be sent to you by e-mail once completed; it usually takes 10 seconds for each protein sequence.

| Read Me | Supporting Information | Citation |
Enter Query Seqences
Submit
Cancel
Or, Upload a File for Batch Prediction
Upload file:
Browse…
Your Email:
Batch Submit
Cancel

Figure 6. A semi-screenshot of the top page for the web server **iPPBS-Opt** at http://www.jci-bioinfo.cn/iPPBS-Opt.

Step 2. Either type or copy/paste the query protein sequences into the input box at the center of Figure 6. 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 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 109 amino acid residues, of which 11 are highlighted with red, meaning belonging to binding site; (2) Sequence-2 contains 275 residues, of which 25 are highlighted with red, belonging binding site. All these predicted results are fully consistent with experimental observations except for residues 53 in sequence-1 and residues 62 and 249 in sequence-2 that are overpredicted.

Step 4. As shown on the lower panel of Figure 6, batch prediction can also be selected by entering an e-mail address and the desired batch input file (in FASTA format naturally) 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-Opt**.

Step 6. Click the Supporting Information button to download the benchmark dataset used in this study.

4. Conclusions

It is a very effective approach to optimize the training dataset via the KNNC treatment and IHTS treatment to enhance the prediction quality in identifying the protein-protein binding sites. This is because the training datasets constructed in this area without undergoing such an optimization procedure are usually extremely skewed and unbalanced, with the negative subset being overwhelmingly larger than the positive one. It is anticipated that the **iPPBS-Opt** web server presented in this paper will become a very useful high throughput tool for identifying protein-protein binding sites, or at the very least, a complementary tool to the existing prediction methods in this area.

Supplementary Materials: Supplementary materials can be accessed at: http://www.mdpi.com/1420-3049/21/1/95/s1.

Acknowledgments: The authors wish to thank the two anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of this paper. This work was partially supported by the National Nature Science Foundation of China (No. 61261027, 61262038, 31260273, 61202313), the Natural Science Foundation of Jiangxi Province, China (No. 20122BAB211033, 20122BAB201044, 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.

Author Contributions: Jianhua Jia: conducted the computation and wrote the preliminary version for the paper; Zi Liu: helped to establish the web-server; Xuan Xiao: provided the facilities and participated in the analysis and discussion; Bingxiang Liu: provided the facilities and participated in the analysis and discussion; Kuo-Chen Chou: guided the entire research, analyzed the computed results, and finalizing the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chou, K.C.; Cai, Y.D. Predicting protein-protein interactions from sequences in a hybridization space. *J. Proteome Res.* **2006**, *5*, 316–322. [CrossRef] [PubMed]
2. Ma, D.L.; Chan, D.S.; Leung, C.H. Group 9 organometallic compounds for therapeutic and bioanalytical applications. *Acc. Chem. Res.* **2014**, *47*, 3614–3631. [CrossRef] [PubMed]
3. Ma, D.L.; He, H.Z.; Leung, K.H.; Chan, D.S.; Leung, C.H. Bioactive luminescent transition-metal complexes for biomedical applications. *Angew. Chem.* **2013**, *52*, 7666–7682. [CrossRef] [PubMed]
4. Tomasselli, A.G.; Heinrikson, R.L. Prediction of the Tertiary Structure of a Caspase-9/Inhibitor Complex. *FEBS Lett.* **2000**, *470*, 249–256.
5. Leung, C.H.; Chan, D.S.; He, H.Z.; Cheng, Z.; Yang, H.; Ma, D.L. Luminescent detection of DNA-binding proteins. *Nucleic Acids Res.* **2012**, *40*, 941–955. [CrossRef] [PubMed]
6. Leung, C.H.; Chan, D.S.; Ma, V.P.; Ma, D.L. DNA-binding small molecules as inhibitors of transcription factors. *Med. Res. Rev.* **2013**, *33*, 823–846. [CrossRef] [PubMed]
7. Jones, D.; Heinrikson, R.L. Prediction of the tertiary structure and substrate binding site of caspase-8. *FEBS Lett.* **1997**, *419*, 49–54.
8. Wei, D.Q.; Zhong, W.Z. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. (Erratum: ibid., 2003, Vol. 310, 675). *Biochem. Biophys. Res. Commun.* **2003**, *308*, 148–151.
9. Chou, K.C. Review: Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.* **2004**, *11*, 2105–2134. [CrossRef] [PubMed]
10. Deng, L.; Guan, J.; Dong, Q.; Zhou, S. Prediction of protein-protein interaction sites using an ensemble method. *BMC Bioinform.* **2009**, *10*. [CrossRef] [PubMed]

11. Chen, X.W.; Jeong, J.C. Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics* **2009**, *25*, 585–591. [CrossRef] [PubMed]
12. Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef] [PubMed]
13. Ding, H.; Deng, E.Z.; Yuan, L.F.; Liu, L. iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed Res. Int.* **2014**, *2014*. [CrossRef] [PubMed]
14. Lin, H.; Deng, E.Z.; Ding, H.; Chen, W. iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.* **2014**, *42*, 12961–12972. [CrossRef] [PubMed]
15. Liu, B.; Xu, J.; Lan, X.; Xu, R.; Zhou, J. 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* **2014**, *9*, e106691. [CrossRef] [PubMed]
16. Xu, Y.; Wen, X.; Wen, L.S.; Wu, L.Y. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS ONE* **2014**, *9*, e105018. [CrossRef] [PubMed]
17. Xu, R.; Zhou, J.; Liu, B.; He, Y.A.; Zou, Q. Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. *J. Biomol. Struct. Dyn.* **2015**, *33*, 1720–1730. [CrossRef] [PubMed]
18. Liu, B.; Fang, L.; Wang, S.; Wang, X.; Li, H. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. *J. Theor. Biol.* **2015**, *385*, 153–159. [CrossRef] [PubMed]
19. Chen, W.; Feng, P.; Ding, H. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.* **2015**, *490*, 26–33. [CrossRef] [PubMed]
20. Liu, B.; Fang, L.; Long, R.; Lan, X. iEnhancer-2L: A two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics* **2015**. [CrossRef] [PubMed]
21. Yan, C.; Dobbs, D.; Honavar, V. A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics* **2004**, *20*, i371–i378. [CrossRef] [PubMed]
22. Ofran, Y.; Rost, B. Predicted protein-protein interaction sites from local sequence information. *FEBS Lett.* **2003**, *544*, 236–239. [CrossRef]
23. Jones, S.; Thornton, J.M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 13–20. [CrossRef] [PubMed]
24. Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637. [CrossRef] [PubMed]
25. Mihel, J.; Šikić, M.; Tomić, S.; Jeren, B.; Vlahovicek, K. PSAIA-rotein structure and interaction analyzer. *BMC Struct. Biol.* **2008**, *8*. [CrossRef] [PubMed]
26. Wang, B.; Huang, D.S.; Jiang, C. A new strategy for protein interface identification using manifold learning method. *IEEE Trans. Nanobiosci.* **2014**, *13*, 118–123. [CrossRef] [PubMed]
27. Chou, K.C.; Shen, H.B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.* **2007**, *370*. [CrossRef]
28. Chou, K.C. Prediction of signal peptides using scaled window. *Peptides* **2001**, *22*, 1973–1979. [CrossRef]
29. Shen, H.B. Signal-CF: A subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun.* **2007**, *357*, 633–640.
30. Xu, Y.; Ding, J.; Wu, L.Y. iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS ONE* **2013**, *8*, e55844. [CrossRef] [PubMed]
31. Xu, Y.; Shao, X.J.; Wu, L.Y.; Deng, N.Y. iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine *S*-nitrosylation sites in proteins. *PeerJ* **2013**, *1*, e171. [CrossRef] [PubMed]
32. Qiu, W.R.; Xiao, X.; Lin, W.Z. iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach. *Biomed. Res. Int.* **2014**, *2014*. [CrossRef] [PubMed]
33. Qiu, W.R.; Xiao, X. iUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model. *J. Biomol. Struct. Dyn.* **2015**, *33*, 1731–1742. [CrossRef] [PubMed]
34. Xu, Y.; Wen, X.; Shao, X.J.; Deng, N.Y. iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *Int. J. Mol. Sci.* **2014**, *15*, 7594–7610. [CrossRef] [PubMed]

35. Chou, K.C. Review: Prediction of human immunodeficiency virus protease cleavage sites in proteins. *Anal. Biochem.* **1996**, *233*. [CrossRef]
36. Chou, K.C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins* **2001**, *43*, 246–255. [CrossRef] [PubMed]
37. Zhang, C.T. An optimization approach to predicting protein structural class from amino acid composition. *Protein Sci.* **1992**, *1*, 401–408. [CrossRef] [PubMed]
38. Chou, J.J. Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach. *J. Protein Chem.* **1993**, *12*, 291–302. [CrossRef] [PubMed]
39. Elrod, D.W. Bioinformatical analysis of G-protein-coupled receptors. *J. Proteome Res.* **2002**, *1*, 429–433.
40. Feng, K.Y.; Cai, Y.D. Boosting classifier for predicting protein domain structural class. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 213–217. [CrossRef] [PubMed]
41. Shen, H.B.; Yang, J. Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. *J. Theor. Biol.* **2006**, *240*, 9–13. [CrossRef] [PubMed]
42. Shen, H.B. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Anal. Biochem.* **2009**, *394*, 269–274. [CrossRef] [PubMed]
43. Wang, M.; Yang, J.; Xu, Z.J. SLLE for predicting membrane protein types. *J. Theor. Biol.* **2005**, *232*, 7–15. [CrossRef] [PubMed]
44. Lin, W.Z.; Fang, J.A. iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. *PLoS ONE* **2011**, *6*, e24756. [CrossRef] [PubMed]
45. Xiao, X.; Min, J.L.; Wang, P. iGPCR-Drug: A web server for predicting interaction between GPCRs and drugs in cellular networking. *PLoS ONE* **2013**, *8*, e72234. [CrossRef] [PubMed]
46. Xiao, X.; Wu, Z.C. iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J. Theor. Biol.* **2011**, *284*, 42–51. [CrossRef] [PubMed]
47. Chou, K.C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteom.* **2009**, *6*, 262–274. [CrossRef]
48. Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L. Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* **2015**, *43*, W65–W71. [CrossRef] [PubMed]
49. Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**, *21*, 10–19. [CrossRef] [PubMed]
50. Cao, D.S.; Xu, Q.S.; Liang, Y.Z. propy: A tool to generate various modes of Chou’s PseAAC. *Bioinformatics* **2013**, *29*, 960–962. [CrossRef] [PubMed]
51. Lin, S.X.; Lapointe, J. Theoretical and experimental biology in one—A symposium in honour of Professor Kuo-Chen Chou’s 50th anniversary and Professor Richard Giegé’s 40th anniversary of their scientific careers. *J. Biomed. Sci. Eng.* **2013**, *6*, 435–442. [CrossRef]
52. Mohabatkar, H.; Mohammad Beigi, M.; Esmaeili, A. Prediction of GABA(A) receptor proteins using the concept of Chou’s pseudo-amino acid composition and support vector machine. *J. Theor. Biol.* **2011**, *281*, 18–23. [CrossRef] [PubMed]
53. Mondal, S.; Pai, P.P. Chou’s pseudo amino acid composition improves sequence-based antifreeze protein prediction. *J. Theor. Biol.* **2014**, *356*, 30–35. [CrossRef] [PubMed]
54. Hajisharifi, Z.; Piryaiee, M.; Mohammad Beigi, M.; Behbahani, M.; Mohabatkar, H. Predicting anticancer peptides with Chou’s pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* **2014**, *341*, 34–40. [CrossRef] [PubMed]
55. Nanni, L.; Lumini, A.; Gupta, D.; Garg, A. 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.* **2012**, *9*, 467–475. [CrossRef] [PubMed]
56. Hayat, M.; Khan, A. Discriminating Outer Membrane Proteins with Fuzzy K-Nearest Neighbor Algorithms Based on the General Form of Chou’s PseAAC. *Protein Pept. Lett.* **2012**, *19*, 411–421. [CrossRef] [PubMed]
57. Du, P.; Gu, S.; Jiao, Y. 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.* **2014**, *15*, 3495–3506. [CrossRef] [PubMed]
58. Chou, K.C. Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* **2015**, *11*, 218–234. [CrossRef] [PubMed]
59. Zhong, W.Z.; Zhou, S.F. Molecular science for drug development and biomedicine. *Int. J. Mol. Sci.* **2014**, *15*, 20072–20078. [CrossRef] [PubMed]

60. Qiu, W.R.; Xiao, X. iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.* **2014**, *15*, 1746–1766. [CrossRef] [PubMed]
61. Chen, W.; Feng, P.M.; Lin, H. iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **2013**, *41*, e68. [CrossRef] [PubMed]
62. Guo, S.H.; Deng, E.Z.; Xu, L.Q.; Ding, H. iNuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* **2014**, *30*, 1522–1529. [CrossRef] [PubMed]
63. Chen, W.; Lei, T.Y.; Jin, D.C. PseKNC: A flexible web-server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.* **2014**, *456*, 53–60. [CrossRef] [PubMed]
64. Liu, B.; Liu, F.; Fang, L.; Wang, X. 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* **2015**, *31*, 1307–1309. [CrossRef] [PubMed]
65. Du, P.; Wang, X.; Xu, C.; Gao, Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Anal. Biochem.* **2012**, *425*, 117–119. [CrossRef] [PubMed]
66. Tanford, C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J. Am. Chem. Soc.* **1962**, *84*, 4240–4274. [CrossRef]
67. Hopp, T.P.; Woods, K.R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA* **1981**, *78*, 3824–3828. [CrossRef] [PubMed]
68. Krigbaum, W.R.; Knutton, S.P. Prediction of the amount of secondary structure in a globular protein from its amino acid composition. *Proc. Natl. Acad. Sci. USA* **1973**, *70*, 2809–2813. [CrossRef] [PubMed]
69. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **1974**, *185*, 862–864. [CrossRef] [PubMed]
70. Charton, M.; Charton, B.I. The structural dependence of amino acid hydrophobicity parameters. *J. Theor. Biol.* **1982**, *99*, 629–644. [CrossRef]
71. Rose, G.D.; Geselowitz, A.R.; Lesser, G.J.; Lee, R.H.; Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins. *Science* **1985**, *229*, 834–838. [CrossRef] [PubMed]
72. Zhou, P.; Tian, F.; Li, B.; Wu, S.; Li, Z. Genetic algorithm-based virtual screening of combinative mode for peptide/protein. *Acta Chim. Sin. Chin. Ed.* **2006**, *64*, 691–697.
73. Martel, P. Biophysical aspects of neutron scattering from vibrational modes of proteins. *Prog. Biophys. Mol. Biol.* **1992**, *57*, 129–179. [CrossRef]
74. Gordon, G. Extrinsic electromagnetic fields, low frequency (phonon) vibrations, and control of cell function: A non-linear resonance system. *J. Biomed. Sci. Eng.* **2008**, *1*, 152–156. [CrossRef]
75. Madkan, A.; Blank, M.; Elson, E.; Goodman, R. Steps to the clinic with ELF EMF. *Nat. Sci.* **2009**, *1*, 157–165. [CrossRef]
76. Wang, J.F. Insight into the molecular switch mechanism of human Rab5a from molecular dynamics simulations. *Biochem. Biophys. Res. Commun.* **2009**, *390*, 608–612. [CrossRef] [PubMed]
77. Wang, J.F.; Gong, K.; Wei, D.Q. Molecular dynamics studies on the interactions of PTP1B with inhibitors: From the first phosphate-binding site to the second one. *Protein Eng. Des. Sel.* **2009**, *22*, 349–355. [CrossRef] [PubMed]
78. Chou, K.C. Low-frequency resonance and cooperativity of hemoglobin. *Trends Biochem. Sci.* **1989**, *14*, 212–213. [CrossRef]
79. Wang, J.F. Insights from studying the mutation-induced allostery in the M2 proton channel by molecular dynamics. *Protein Eng. Des. Sel.* **2010**, *23*, 663–666. [CrossRef] [PubMed]
80. Chou, K.C. The biological functions of low-frequency phonons: 6. A possible dynamic mechanism of allosteric transition in antibody molecules. *Biopolymers* **1987**, *26*, 285–295. [CrossRef] [PubMed]
81. Schnell, J.R.; Chou, J.J. Structure and mechanism of the M2 proton channel of influenza A virus. *Nature* **2008**, *451*, 591–595. [CrossRef] [PubMed]
82. Mao, B. Collective motion in DNA and its role in drug intercalation. *Biopolymers* **1988**, *27*, 1795–1815.
83. Zhang, C.T.; Maggiora, G.M. Solitary wave dynamics as a mechanism for explaining the internal motion during microtubule growth. *Biopolymers* **1994**, *34*, 143–153.
84. Chou, K.C. Review: Low-frequency collective motion in biomacromolecules and its biological functions. *Biophys. Chem.* **1988**, *30*, 3–48. [CrossRef]
85. Liu, H.; Wang, M. Low-frequency Fourier spectrum for predicting membrane protein types. *Biochem. Biophys. Res. Commun.* **2005**, *336*, 737–739. [CrossRef] [PubMed]

86. Shensa, M. The discrete wavelet transform: Wedding the a trous and Mallat algorithms. *IEEE Trans. Signal Process.* **1992**, *40*, 2464–2482. [CrossRef]
87. Mallat, S.G. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674–693. [CrossRef]
88. Sun, Y.; Wong, A.K.; Kamel, M.S. Classification of imbalanced data: A review. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 687–719. [CrossRef]
89. Laurikkala, J. *Improving Identification of Difficult Small Classes by Balancing Class Distribution, 63–66*; Springer: Berlin, Heidelberg, Germany, 2001.
90. Xiao, X.; Min, J.L.; Lin, W.Z.; Liu, Z. iDrug-Target: Predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. *J. Biomol. Struct. Dyn.* **2015**, *33*, 2221–2233. [CrossRef] [PubMed]
91. Liu, Z.; Xiao, X.; Qiu, W.R. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. *Anal. Biochem.* **2015**, *474*, 69–77. [CrossRef] [PubMed]
92. Zhang, C.T. Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. *Biophys. J.* **1992**, *63*, 1523–1529. [CrossRef]
93. Chou, K.C. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J. Biol. Chem.* **1993**, *268*, 16938–16948. [PubMed]
94. Zhang, C.T. An analysis of protein folding type prediction by seed-propagated sampling and jackknife test. *J. Protein Chem.* **1995**, *14*, 583–593. [CrossRef] [PubMed]
95. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2011**, *16*, 321–357.
96. Kandaswamy, K.K.; Moller, S.; Suganthan, P.N.; Sridharan, S.; Pugalenthi, G. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *J. Theor. Biol.* **2011**, *270*, 56–62. [CrossRef] [PubMed]
97. Pugalenthi, G.; Kandaswamy, K.K.; Kolatkar, P. RSARF: Prediction of Residue Solvent Accessibility from Protein Sequence Using Random Forest Method. *Protein Pept. Lett.* **2012**, *19*, 50–56. [CrossRef] [PubMed]
98. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
99. Shen, H.B. Euk-mPLoc: A fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.* **2007**, *6*, 1728–1734.
100. Shen, H.B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *J. Proteome Res.* **2006**, *5*, 1888–1897.
101. Shen, H.B. QuatIdent: A web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. *J. Proteome Res.* **2009**, *8*, 1577–1584. [CrossRef] [PubMed]
102. Chen, J.; Liu, H.; Yang, J. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids* **2007**, *33*, 423–428. [CrossRef] [PubMed]
103. Chou, K.C. Using subsite coupling to predict signal peptides. *Protein Eng.* **2001**, *14*, 75–79. [CrossRef] [PubMed]
104. Chen, W.; Lin, H.; Feng, P.M.; Ding, C. iNuc-PhysChem: A Sequence-Based Predictor for Identifying Nucleosomes via Physicochemical Properties. *PLoS ONE* **2012**, *7*, e47843. [CrossRef] [PubMed]
105. Wu, Z.C.; Xiao, X. iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. BioSyst.* **2012**, *8*, 629–641.
106. Lin, W.Z.; Fang, J.A.; Xiao, X. iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. BioSyst.* **2013**, *9*, 634–644. [CrossRef] [PubMed]
107. Xiao, X.; Wang, P.; Lin, W.Z. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.* **2013**, *436*, 168–177. [CrossRef] [PubMed]
108. Chou, K.C. Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. *Mol. BioSyst.* **2013**, *9*, 1092–1100. [CrossRef] [PubMed]
109. Zhang, C.T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275–349. [CrossRef]
110. Zhou, G.P.; Assa-Munt, N. Some insights into protein structural class prediction. *Proteins: Struct. Funct. Genet.* **2001**, *44*, 57–59. [CrossRef] [PubMed]
111. Chou, K.C.; Cai, Y.D. Prediction and classification of protein subcellular location:-sequence-order effect and pseudo amino acid composition. *J. Cell. Biochem.* **2003**, *90*, 1250–1260. [CrossRef] [PubMed]

112. Dehzangi, A.; Heffernan, R.; Sharma, A.; Lyons, J.; Paliwal, K.; Sattar, A. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. *J. Theor. Biol.* **2015**, *364*, 284–294. [CrossRef] [PubMed]
113. Khan, Z.U.; Hayat, M.; Khan, M.A. Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. *J. Theor. Biol.* **2015**, *365*, 197–203. [CrossRef] [PubMed]
114. Kumar, R.; Srivastava, A.; Kumari, B.; Kumar, M. Prediction of beta-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.* **2015**, *365*, 96–103. [CrossRef] [PubMed]
115. Shen, H.B.; Yang, J. Euk-PLoc: An ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids* **2007**, *33*, 57–67. [CrossRef] [PubMed]
116. Chou, K.C.; Shen, H.B. Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun.* **2006**, *347*, 150–157. [CrossRef] [PubMed]
117. Forsen, S. Graphical rules for enzyme-catalyzed rate laws. *Biochem. J.* **1980**, *187*, 829–835.
118. Chou, K.C. Graphic rules in steady and non-steady enzyme kinetics. *J. Biol. Chem.* **1989**, *264*, 12074–12079. [PubMed]
119. Wu, Z.C.; Xiao, X. 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. *J. Theor. Biol.* **2010**, *267*, 29–34. [CrossRef] [PubMed]
120. Althaus, I.W.; Chou, J.J.; Gonzales, A.J.; Kezdy, F.J.; Romero, D.L.; Aristoff, P.A.; Tarpley, W.G.; Reusser, F. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. *Biochemistry* **1993**, *32*, 6548–6554. [CrossRef] [PubMed]
121. Chou, K.C. Graphic rule for drug metabolism systems. *Curr. Drug Metab.* **2010**, *11*, 369–378. [CrossRef] [PubMed]
122. Zhou, G.P. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. *J. Theor. Biol.* **2011**, *284*, 142–148. [CrossRef] [PubMed]
123. Fawcett, J.A. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2005**, *27*, 861–874. [CrossRef]
124. Chen, W.; Lin, H. Pseudo nucleotide composition or PseKNC: An effective formulation for analyzing genomic sequences. *Mol. BioSyst.* **2015**, *11*, 2620–2634. [CrossRef] [PubMed]

Sample Availability: All the samples used in this study for training and testing the predictor are available by downloading them from the web-server at http://www.jci-bioinfo.cn/iPPBS-Opt.