SCIENTIFIC REPORTS

OPEN

Using a Classifier Fusion Strategy to Identify Anti-angiogenic Peptides

Lina Zhang, Runtao Yang & Chengjin Zhang

Received: 25 June 2018
Accepted: 7 September 2018
Published online: 14 September 2018

Anti-angiogenic peptides perform distinct physiological functions and potential therapies for angiogenesis-related diseases. Accurate identification of anti-angiogenic peptides may provide significant clues to understand the essential angiogenic homeostasis within tissues and develop antineoplastic therapies. In this study, an ensemble predictor is proposed for anti-angiogenic peptide prediction by fusing an individual classifier with the best sensitivity and another individual one with the best specificity. We investigate predictive capabilities of various feature spaces with respect to the corresponding optimal individual classifiers and ensemble classifiers. The accuracy and Matthew's Correlation Coefficient (MCC) of the ensemble classifier trained by Bi-profile Bayes (BpB) features are 0.822 and 0.649, respectively, which represents the highest prediction results among the investigated prediction models. Discriminative features are obtained from BpB using the Relief algorithm followed by the Incremental Feature Selection (IFS) method. The sensitivity, specificity, accuracy, and MCC of the ensemble classifier trained by the discriminative features reach up to 0.776, 0.888, 0.832, and 0.668, respectively. Experimental results indicate that the proposed method is far superior to the previous study for anti-angiogenic peptide prediction.

Angiogenesis is a process of new blood vessel formations[1], which involves multiple biological behaviors including endothelial cell proliferation, migration, apoptosis, cell-cell and cell-matrix adhesion[2]. It contributes to vascular remodeling and maturation[3]. Angiogenesis is tightly regulated by stimulators and inhibitors[4]. Appropriate balance between stimulators and inhibitors plays a pivotal function in maintaining and regulating angiogenesis, which often involves embryonic development, wound healing, menstrual cycle, and hair cycle[2]. Disruption of such an equilibrium is often associated with pathological processes[5,6], including heart diseases, stroke, diabetes, blindness[2], proliferative diabetic retinopathy, and atherosclerosis[7]. Especially, abundant evidence has indicated that imbalanced angiogenesis is involved in cancer progression[8,9], due to the fact that the newly formed tumor vasculature provides stable blood supply for the growing tumor mass and eventually disseminates tumor cells that have escaped from the primary tumor[10].

Angiogenesis inhibitors are needed to down-regulate the progression of angiogenesis, which would contribute to the development of therapeutic treatments for these angiogenesis-related diseases[11]. Previous studies have indicated that anti-angiogenic proteins or polypeptides can inhibit the angiogenesis process and have been applied in the therapies of cancers and other diseases[12]. However, most of anti-angiogenic proteins are large and complex, and they would cause some serious side effects[9,13]. In contrast to proteins and polypeptides, anti-angiogenic peptides have advantages for therapeutic application, in terms of their small size, lack of toxicity, lower immune reaction to the host system, higher solubility in water, higher stability, receptivity to chemical modification, and increased bio-availability[2]. In addition, they have a better ability to target and penetrate tissues[14]. Therefore, anti-angiogenic peptides have been shown as promising therapies for tumors and other angiogenesis-related diseases[15–17].

Several anti-angiogenic peptide candidates which are currently in clinical trials are showing promising results[9,18]. For example, YSNS, a cyclized anti-angiogenic peptide, has been demonstrated to inhibit the capillary network formation in vivo and limit tumor growth in the small cell lung cancer[19]. KV11, a 12-mer peptide, has an ability to suppress tumor growth and tumor microvasculature in breast cancer xenografts[20]. Anti-angiogenic SPARC peptides have been investigated to inhibit progression of neuroblastoma tumors[21]. In view of the physiological functions and potential therapeutic purposes in organisms, identification of anti-angiogenic peptides may not only contribute to better fundamental understanding of the essential angiogenic homeostasis within tissues[22]. but also have significant implications for development of antineoplastic therapies[6].

School of Mechanical, Electrical and Information Engineering, Shandong University at Weihai, Weihai, 264209, China. Correspondence and requests for materials should be addressed to R.Y. (email: yrt@sdu.edu.cn)

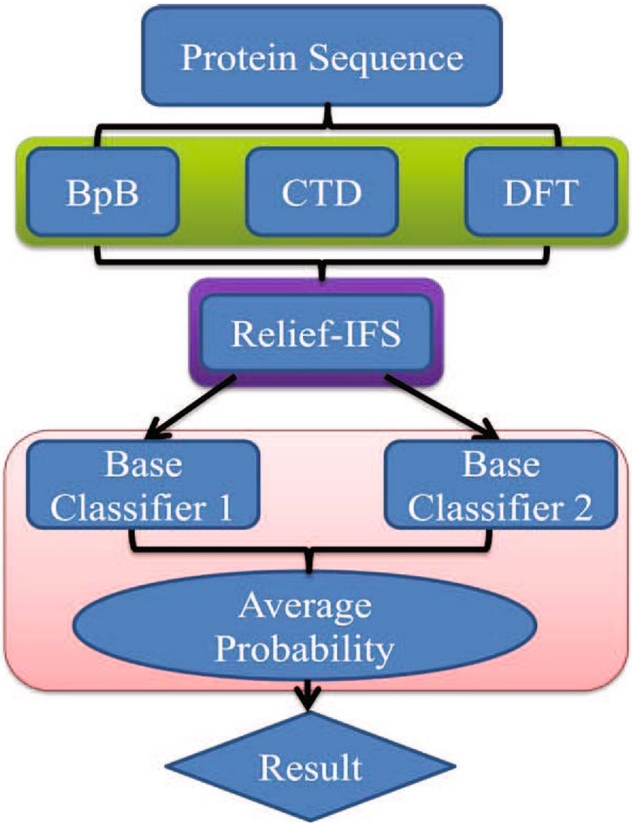

Figure 1. The construction process of the proposed anti-angiogenic peptide prediction model.

There are some computational and experimental methodologies to identify anti-angiogenic peptides. Based on the protein basic local alignment search tool (BLAST), searching the conserved domains of angiogenesis-associated proteins existing in the proteome is a common computational method to identify the putative anti-angiogenic peptides[23]. Homology modeling is another computational technique where the structure of an anti-angiogenic peptide is determined by comparison to a high-resolution structure or structures with sequence homology[9]. However, these two methods have a critical shortcoming that they can't work when there are no homology sequences existing in the proteome. Computational screening via docking is a viable method of peptide discovery[9]. However, its complexity leads to a prohibitively expensive cost. Molecular dynamics (MD) is a computational simulation technique to identify the anti-angiogenic peptides, but the high computational cost hinders the process of MD[9]. In addition, experimental identification of anti-angiogenic peptides relies on an empirical process[4], which is both labor intensive and time consuming[22].

Recently, machine learning methods have been potential tools and have achieved promising results for identifying protein attributes. Ettayapuram Ramaprasad AS *et al.*[24] developed a support vector machine (SVM)-based predictor to identify anti-angiogenic peptides, using various features extracted from peptide sequences including Binary Profile Patterns (BPP), Amino Acid Composition (AAC), and Dipeptide Compositions (DPC). The accuracy and Matthew's Correlation Coefficient (MCC) of the method are 0.748 and 0.500, respectively. The prediction performance is acceptable, but there still exist the following shortcomings. (1) No feature selection technique was employed by the predictor proposed in the existing method[24], which would lead to dimension disaster and poor performance[25]. Feature selection has the ability to get rid of redundancy information or noise and decrease model complexity[26]. (2) The method[24] was based on an individual classifier which could have its own inherent defects[27]. It is generally accepted that the ensemble predictor integrating multiple basic classifiers of diverse learning policies (or diversely trained) is superior in carrying out statistics, calculation, and characterization analysis compared to its base classifiers[27]. Therefore, ensemble methods have been suggested as the promising measures for protein classification problems[28].

In view of the above shortcomings, a classifier fusion method as illustrated in Fig. 1 is proposed in this paper to promote the ability to predict anti-angiogenic peptides. We investigate predictive capabilities of various feature spaces including CTD (Composition, Transition and Distribution), BpB (Bi-profile Bayes), and DFT (Discrete Fourier Transform). These features are all related with the properties of the target peptides. To decrease the complexity of computation, the relevance of features and categories is assessed by Relief algorithm, and then IFS (Incremental Feature Selection) method is applied to capture a set of important features. Several individual classifiers are separately adopted to construct anti-angiogenic peptide prediction models. To achieve a better prediction accuracy, the classifier with the best sensitivity and the classifier with the best specificity are selected as the base classifiers. The final output of the prediction model is equal to the average probability for a given sample to

Feature Space	Optimal Classifier	Sn	Sp	Acc	MCC	AUC
BpB	NB	0.682	0.925	0.804	0.626	0.902
CTD	RBFNetwork	0.551	0.766	0.659	0.325	0.698
DFT	NNA	0.692	0.579	0.636	0.273	0.636
BpB + CTD	RBFNetwork	0.701	0.804	0.752	0.507	0.806
BpB + DFT	RF	0.710	0.850	0.780	0.566	0.843
CTD + DFT	RF	0.664	0.682	0.673	0.346	0.699
BpB + CTD + DFT	RF	0.673	0.794	0.734	0.471	0.802

Table 1. Prediction performance of various feature spaces with respect to the corresponding optimal individual classifiers.

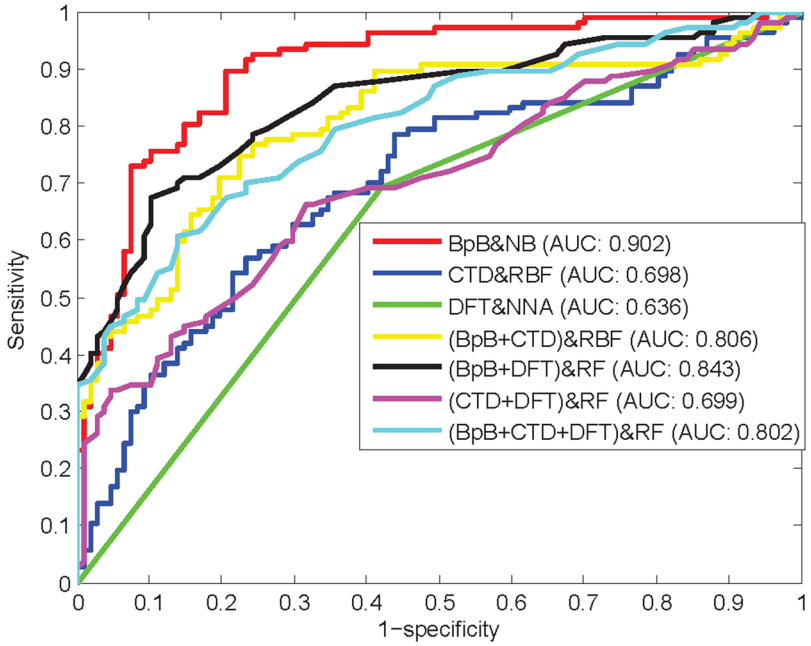

Figure 2. ROC curves of various feature spaces with respect to the corresponding optimal individual classifiers.

be an anti-angiogenic peptide predicted by the base classifiers. 10-fold cross validation is carried out to verify the effectiveness of the prediction model. Simulation results show that the sensitivity, specificity, accuracy, and MCC of the proposed method reach up to 0.776, 0.888, 0.832, and 0.668, respectively, higher than those of the existing method[24]. The comparison results indicate that the proposed method is a promising tool for identifying anti-angiogenic peptides.

Results and Discussion

Performance of Various Feature Spaces on Different Individual Classifiers. To investigate the optimal individual classifiers for different feature types, we evaluate the impact of various features on the performance of multiple individual classifiers. The prediction results of various feature spaces with respect to the corresponding optimal classifiers are given in Table 1. Figure 2 illustrates the receiver operating characteristic (ROC) curves of various feature spaces with respect to the corresponding optimal individual classifiers. As listed in Table 1, the prediction accuracy of various feature spaces with respect to the corresponding optimal classifiers is in the range of 0.636 to 0.804, indicating an ideal prediction effect for anti-angiogenic peptides. As shown in Fig. 2, the accuracy, MCC, and area under the ROC curve (AUC) of BpB is 0.804, 0.626, and 0.902, respectively, which represents the highest prediction results among the various feature spaces. These results demonstrate that statistical differences about the position-specific amino acid composition at the N-terminal region and C-terminal region are relatively discriminative in anti-angiogenic peptide identification, which is in accordance with research results in the previous study[24].

In addition, the optimal classifiers for different individual feature types are totally different (i.e., Naïve Bayes (NB) for BpB, Radial Basis Function Network (RBFNetwork) for CTD, and Nearest Neighbor Algorithm (NNA) for DFT). Except that the optimal classifier for BpB + CTD is identical to that for CTD, the optimal classifiers for hybrid feature spaces are totally different from those for their component feature types. These results show that an individual classifier is good at dealing with data classification with specific feature distribution. Except for BpB + CTD, other hybrid feature spaces have the identical optimal classifier, i.e., Random Forest (RF), demonstrating that RF is remarkable on managing data classification with complicated structure. In addition, except

Feature Space	Optimal Classifier	Sn	Sp	Acc	MCC	AUC
BpB	NB + LR	0.766	0.879	0.822	0.649	0.870
CTD	RBFNetwork + NNA	0.617	0.57	0.593	0.187	0.676
DFT	NB + NNA	0.701	0.579	0.64	0.282	0.645
BpB + CTD	NB + LR	0.794	0.72	0.757	0.515	0.842
BpB + DFT	NB + LR	0.748	0.701	0.724	0.449	0.831
CTD + DFT	NB + RF	0.542	0.72	0.631	0.266	0.700
BpB + CTD + DFT	NB + LR	0.748	0.738	0.743	0.486	0.838

Table 2. Prediction performance of various feature spaces with respect to the corresponding optimal ensemble classifiers.

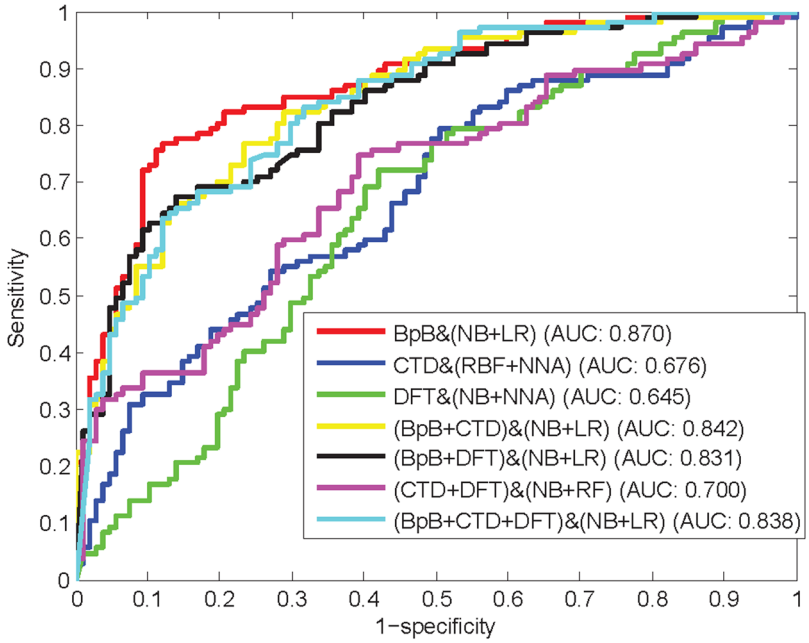

Figure 3. ROC curves of various feature spaces with respect to the corresponding optimal ensemble classifiers.

CTD + DFT, the accuracy values of hybrid feature spaces are not better than those of individual feature types. These results indicate that much redundant information may exist in hybrid feature spaces, which would deteriorate prediction performance in anti-angiogenic peptide prediction.

Performance of Various Feature Spaces on Ensemble Classifiers. To investigate the best ensemble classifiers with respect to different feature types, we first examine the prediction performance of various features on multiple individual classifiers. Then, the ensemble classifier is determined by combining an individual classifier with a better sensitivity and another one with a better specificity. Table 2 shows the prediction results of various feature spaces with respect to the corresponding optimal ensemble classifiers. The ROC curves of various feature spaces with respect to the corresponding optimal ensemble classifiers are depicted in Fig. 3. From Table 2, for various feature spaces, the corresponding ensemble classifiers are not identical. However, except CTD, the ensemble classifiers for other feature spaces all have an NB classifier, indicating that an NB classifier can predict negative samples better than other individual classifiers. For half of different feature spaces, NB + LR (Logistic Regression) is the optimal ensemble classifier to identify anti-angiogenic peptides. Therefore, to verify the effectiveness of the ensemble method, the individual performance of NB classifier and LR classifier on the feature spaces with which the ensemble classifier NB + LR achieves best performance is separately given in Tables 3 and 4.

As shown in Table 3, there is a big difference between Sn and Sp achieved by NB classifier on different feature spaces. As shown in Table 4, although LR classifier achieves a much balanced Sn and Sp on different feature spaces, the Accs are not satisfactory. Compared with the NB classifier and LR classifier, the ensemble classifier NB + LR as given in Table 2 achieves a much better prediction performance on the corresponding feature spaces.

From Tables 1 and 2, hybrid features on the ensemble classifiers do not outperform the corresponding component individual feature types due to the redundant information in the hybrid features. The accuracy of BpB on the ensemble classifier is improved from 0.804 to 0.822. DFT, BpB + CTD, and BpB + CTD + DFT are all the same case with BpB on the corresponding ensemble classifiers. These comparison results reveal that an ensemble classifier can effectively improve prediction performance. However, there are exceptions for other feature spaces whose performance on the ensemble classier is worse than that on the optimal individual classier. In general, an

Feature Space	Classifier	Sn	Sp	Acc	MCC	AUC
BpB	NB	0.682	0.925	0.804	0.626	0.902
BpB + CTD	NB	0.626	0.832	0.734	0.478	0.729
BpB + DFT	NB	0.570	0.804	0.687	0.384	0.704
BpB + CTD + DFT	NB	0.589	0.841	0.715	0.444	0.715

Table 3. The individual performance of NB classifier on different feature spaces.

Feature Space	Classifier	Sn	Sp	Acc	MCC	AUC
BpB	LR	0.785	0.748	0.766	0.533	0.766
BpB + CTD	LR	0.757	0.720	0.738	0.477	0.782
BpB + DFT	LR	0.738	0.682	0.710	0.421	0.710
BpB + CTD + DFT	LR	0.748	0.710	0.729	0.458	0.729

Table 4. The individual performance of LR classifier on different feature spaces.

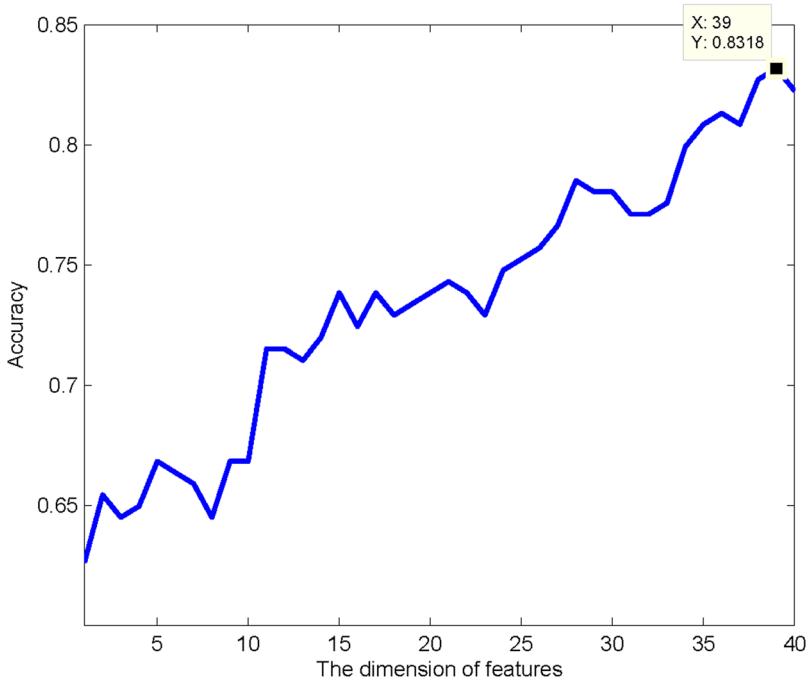

Figure 4. The IFS curve: the accuracy of the prediction model trained by different feature subsets.

ensemble classifier that integrates multiple basic classifiers of diverse learning policies (or diversely trained) can achieve better prediction performance than its component classifiers for protein attribute predictions[28]. These exceptions may be due to lack of diversity in learning policies of the component individual classifiers. The accuracy and MCC of the ensemble classifier trained by Bi-profile Bayes (BpB) features are 0.822 and 0.649, respectively, which represents the highest prediction results among the investigated prediction models using various feature spaces with different classifiers. In addition, from Fig. 3, BpB with the optimal ensemble classifier of NB and LR yields the best AUC of 0.870. Therefore, this study employs BpB with NB + LR to construct the final prediction model.

Feature Selection Results and Corresponding Analysis. The features extracted from the BpB method are sorted according to the weights from highest to lowest given by the Relief algorithm. As provided in Table S1, the feature with a higher ranking suggests that its ability to identify anti-angiogenic peptides is more powerful. Based on the feature ranking, the IFS method is implemented using the ensemble classifier NB + LR. Table S2 shows the detailed prediction results of the prediction model at each iteration based on 10-fold cross validation. As given in Fig. 4, the IFS curve that displays the accuracy of the prediction model at each iteration reaches a peak value when the prediction model is built by the first 39 features in Table S1. Thus, the first 39 features in Table S1 are selected to constitute the optimal feature subset for anti-angiogenic peptide prediction.

Method	Sn	Sp	Acc	MCC	AUC
Without feature selection	0.766	0.879	0.822	0.649	0.870
With feature selection	0.776	0.888	0.832	0.668	0.872

Table 5. Prediction results with the proposed feature selection method or not.

Method	Sn	Sp	Acc	MCC	AUC
Ref.[24]	0.757	0.738	0.748	0.50	—
This study	0.776	0.888	0.832	0.668	0.872

Table 6. Performance comparisons with the existing method on benchmark dataset.

To analyze the effectiveness of the proposed feature selection method, using the ensemble classifier NB + LR, the prediction models with and without the proposed feature selection method are separately constructed. As shown in Table 5 and Fig. 5, with the optimal feature subset generated by the proposed feature selection method, the sensitivity, specificity, accuracy, MCC, and AUC of the prediction model are 0.776, 0.888, 0.832, 0.668, and 0.872, respectively, better than those of the prediction model using all features. Therefore, the Relief combine with IFS is effective to eliminate irrelevant and redundant features existing in the BpB feature space. The final anti-angiogenic peptide prediction model will be constructed by the ensemble classifier NB + LR combined with the proposed feature selection method.

Performance Comparisons with the Existing Method on Benchmark Dataset. To objectively access the prediction ability for anti-angiogenic peptide prediction, performance measures obtained by our method and the existing method[24] on the same benchmark dataset are compared. The detailed prediction results based on 10-fold cross validation are listed in Table 6. As given in Table 6, the proposed method achieves ideal results, obviously outperforming the previous study[24]. More specifically, the specificity, accuracy, and MCC of the proposed method are significantly (i.e., approximately 0.150, 0.084, 0.168) higher than those of the existing method[24]. Therefore, the proposed ensemble method is effective in predicting anti-angiogenic peptides, which may provide a deeper understanding for the essential angiogenic homeostasis, thereby beneficial to develop anti-neoplastic therapies.

The outstanding performance of our predictor is mainly attributed to 3 aspects. (1) The BpB features contain discriminative information for distinguish anti-angiogenic pepetides from non-anti-angiogenic pepetides. (2) The Relief combined with IFS can make a distinct contribution to selecting the optimal features for identifying anti-angiogenic pepetides. (3) The ensemble learning method proposed here takes advantage of superiorities of individual classifiers with respect to specific data structure and distribution.

For classification problems, numerous studies have demonstrated that an effective way to improve prediction performance is to design an advanced learning algorithm. Based on laplacian regularized sparse subspace learning, extreme gradient boosting machine, and ensemble learning, respectively, the computational models developed by Chen X *et al*. achieved superior prediction accuracy for miRNA-disease association[29–31]. Based on ensemble rotation forest learning, Wang L *et al*. proposed an effective computation method for large-scale identification of protein-protein interactions[32]. Based on ensemble learning, a new sequence-based method proposed by Li JQ *et al*. shows a good performance for self-interacting protein prediction[33]. These existing learning algorithms will inspires us to propose novel machine learning models or other ensemble models to identify anti-angiogenic peptides in the future work.

Materials and Methods

Benchmark Dataset. In order to objectively make comparisons with the previous study for anti-angiogenic peptide prediction, the benchmark dataset constructed by Ettayapuram Ramaprasad AS *et al*.[24] containing 107 positive and 107 negative samples is employed to construct the proposed prediction model. None of the peptides has 70% sequence identity to any other in the positive samples. For detailed information of the benchmark dataset, please refer to Table S3.

Feature Extraction. The selection of appropriate protein feature representation methods that can truly reflect their intrinsic correlation with the attribute to be predicted is critical to establish a powerful protein attribute predictor[34]. Appropriate feature representations make it easier for the classifier to recognize underlying regularities, which is vital to the success of classifier learning[35]. Generally, one single feature extraction method cannot capture enough discriminative information for protein attribute predictions. Multiple features from different sources can complement each other in enhancing the discrimination power of a hypothesis. It is an extremely difficult task to discover the best combination of features that are distinctively responsible for accurate classification as no standard technique is available for it[36]. In this study, after investigating the sequence properties of anti-angiogenic peptides carefully, hybrid features extracted from CTD, BpB, and DFT, which are all correlated with the intrinsic properties of these peptides, are adopted for anti-angiogenic peptide identification.

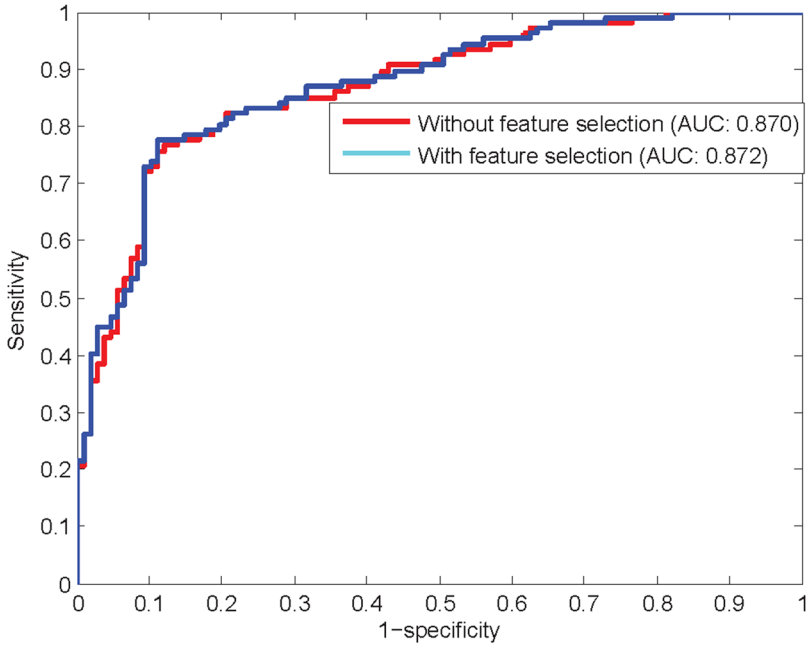

Figure 5. ROC curves with the proposed feature selection method or not.

Bi-profile Bayes. Statistical differences between positive sample set and negative sample set exist in the frequencies of 20 native amino acids occurred along peptide sequences, i.e. Cys, Pro, Ser, Arg, Trp, Thr and Gly are predominant in anti-angiogenic peptides while Ala, Asp, Ile, Leu, Val and Phe are not preferred in these peptides[24]. Important single peptides of a protein are usually hidden at its N- or C-terminal region, which is considered as a key factor for protein function determination[37]. As demonstrated in preliminary analysis[24], there are statistical differences about the position-specific amino acid composition between positive and negative samples at the N-terminal region and C-terminal region. Certain residues (Ser, Pro, Trp, Thr, Arg, and Cys) are preferred at various positions at the N-terminal region of anti-angiogenic peptides while Ala, Val, Glu, Met, Phe, and Asn are prominent at various positions at the N-terminal region of non-anti-angiogenic peptides. For anti-angiogenic peptides, Cys, Gly, Asp, Ser, and Arg are prominent at different positions of the C-terminal region while Ala, Leu, Trp, Ile, and Val are preferred at distinct positions at the C-terminal region of non-anti-angiogenic peptides. In this study, BpB[38] is utilized to calculate statistically significant differences in the distribution of amino acid residues at the N-terminal region and C-terminal region between positive and negative datasets.

Given a peptide segment $P = \{n_1, \ldots, n_i, \ldots, n_m, c_1, \ldots, c_i, \ldots, c_m\}$ with m amino acids at the N-terminus and m amino acids at the C-terminus, where n_i is the ith residue at the N-terminus and c_i represents the ith residue at the C-terminus. After calculating the posterior probabilities of 20 natural amino acids at each position of the C-terminus and N-terminus from the benchmark dataset, a peptide sample can be formulated as

$$(P_N^1, P_N^2, \ldots, P_N^m; P_C^1, P_C^2, \ldots, P_C^m; P_N^{m+1}, P_N^{m+2}, \ldots, P_N^{2m}; P_C^{m+1}, P_C^{m+2}, \ldots, P_C^{2m}), \quad (1)$$

where $(P_N^1, P_N^2, \ldots, P_N^m)$ and $(P_C^1, P_C^2, \ldots, P_C^m)$ denote the posterior probabilities of the corresponding amino acids at each position of the N-terminus and C-terminus compared to the positive dataset, respectively.

Similarly, $(P_N^{m+1}, P_N^{m+2}, \ldots, P_N^{2m})$ and $(P_C^{m+1}, P_C^{m+2}, \ldots, P_C^{2m})$ represent the posterior probabilities of each amino acid at each position of the N-terminus and C-terminus compared to the negative dataset, respectively. The length of N-terminus or C-terminus is set as 10, then each sample is converted into a 40-dimensional feature vector.

Composition, Transition, and Distribution. Primary analysis based on the amino acid composition and residue propensities in the existing method[24] reveals that certain residues (Cys, Trp, Ser, Arg, and Pro) are preferred in anti-angiogenic peptides[24]. In addition, research results in[39] have demonstrated that anti-angiogenic peptides, for the most part, are compositionally similar and they have a relatively high incidence of hydrophobic and cationic residues. In view of the essential physicochemical properties of anti-angiogenic peptides mentioned above, 20 natural amino acids are divided into four groups on the basis of their hydrophobicity and charged character, that is the hydrophobic group $C_1 = \{A, F, G, I, L, M, P, V, W\}$, the polar group $C_2 = \{C, N, Q, S, T, Y\}$, the positively charged group $C_3 = \{H, K, R\}$, and the negatively charged group $C_4 = \{D, E\}$[40]. Based on the four groups, the concept of CTD proposed d by Dubchak I *et al.*[41] is introduced to extract information on global composition, physicochemical property, and sequence order from peptide sequences.

With a particular property, composition (C) calculates the frequencies of each group in a given peptide, which is defined as

$$\left(\frac{N_1}{L}, \frac{N_2}{L}, \frac{N_3}{L}, \frac{N_4}{L}\right), \tag{2}$$

where N_i, $i \in \{1, 2, 3, 4\}$ is the number of each group and L is the length of the peptide.

In a given peptide, transition (T) describes the frequencies of an amino acid with a particular property followed by an amino acid with another property, which is formulated by the following equation.

$$\frac{N_{i,j} + N_{j,i}}{L - 1}, \tag{3}$$

where $i, j \in \{1, 2, 3, 4\}$ represents the corresponding group. $N_{i,j}$ is the number of the dipeptide containing two residues from two different groups.

Distribution (D) expresses the distribution pattern of each group which is measured by the position of the first, 25%, 50%, 75%, and 100% of each of the four groups along the sequence, which can be calculated by

$$\left(\frac{N_{1,1}}{L}, \ldots, \frac{N_{1,5}}{L}, \frac{N_{2,1}}{L}, \ldots, \frac{N_{2,5}}{L}, \ldots, \frac{N_{4,1}}{L}, \ldots, \frac{N_{4,5}}{L}\right), \tag{4}$$

where $N_{i,1}$ is the chain length within which the first amino acid of the ith group is located. $N_{i,2}$, $N_{i,3}$, $N_{i,4}$, $N_{i,5}$ measure the chain lengths within which the 25%, 50%, 75%, and 100% of the amino acids of the ith group are located, respectively.

Discrete Fourier Transform. Physicochemical properties of amino acids are the most important features for protein biochemical reactions, which have a deep influence on protein structure and function forming[42]. Dings RP *et al.*[39] have reported that hydrophobic residues are prone to occur in anti-angiogenic peptides. In addition, a protein sequence occasionally shows periodicity of hydrophobicity and hydrophilicity, which can contribute to protein attribute predictions[43].

In this study, based on the hydrophobicity and hydrophilicity, a peptide sequence is transformed into a numerical feature vector. Then the frequency information reflecting the periodicity is merged into a set of discrete components by transferring the coded sequence to its corresponding frequency domain, which reflects the distribution of power contained in a peptide sequence over the frequencies[44]. Via the transformation, some important features hidden in the sequence could be revealed without information loss. This goal can be achieved with the help of DFT. DFT[45], a transformation approach converting numerical values into frequency domain, reveals periodicities of input data as well as the relative strengths of various periodic components.

The DFT of a given peptide sequence with the length of L is defined as

$$F(k) = \sum_{n=1}^{L} H(p_n) e^{-2\pi nkj/L}, \quad k = 0, 1, \ldots, L - 1, \tag{5}$$

where $F(k)$ represents the periodicity characteristics of the sequence and the compositional patterns by sinusoidal waves with various frequencies. $H(p_n)$, $n = 0, 1, \ldots, L-1$ denotes physicochemical property values of each amino acid of the given peptide sequence.

The DFT power spectrum at frequency k is defined as

$$PS(k) = |F(k)|^2, \; k = 0, 1, \ldots, L - 1. \tag{6}$$

The fourier coefficients partially reflect the sequence order information. Generally, the low-frequency components of DFT contain more biological significance than high frequency noisy ones[46]. Hence the DFT power spectrums at low frequencies are chosen as effective features. The minimum length of peptide sequences in the benchmark dataset is 10. For the hydrophobicity or hydrophilicity of amino acids, we use 10 low frequency DFT power spectrums to represent the sequences.

Feature Selection. Not all the extracted features can contribute to the prediction accuracy. Commonly, hybrid features from various sources would bring some redundant or irrelevant features, which may deteriorate the generalization ability and the performance of learning algorithms[25]. To eliminate the redundant features and improve prediction performance, it is necessary to develop feature selection techniques to pick out the optimal features, which can also contribute to simplifying the classifier and gaining deeper insights into the intrinsic properties of protein sequences. To obtain the optimal feature subset, the Relief algorithm combined with IFS method is employed in this study.

In order to describe the correlation between class labels and features, Kira K and Rendell LA developed a feature selection algorithm called Relief in 1992[47]. It runs in low-order polynomial time, and is noise-tolerant to feature interactions. With the ability to differentiate the class labels of close samples, Relief is an effective iterative algorithm to evaluate the prediction ability of each feature by setting feature weights within the interval $[0, 1]$[48], which is represented as

$$W_p^{i+1} = W_p^i - \frac{diff(Y, s_i, NH(s_i))}{n} + \frac{diff(S, s_i, NM(s_i))}{n}, \tag{7}$$

$$diff(*, x, y) = \begin{cases} \|x - y\|, & x \neq y \\ 0, & x = y \end{cases}, \tag{8}$$

where W_p^i and W_p^{i+1} stands for the assigned weights for a given feature p at the ith iteration and the $(i+1)$th iteration, respectively. s_i denotes one of the samples in the dataset numbered i. $NH(s_i)$, called nearest hit, denotes the neighbor samples of s_i in the sample set Y where the samples have the same class label as s_i. $NM(s_i)$, called nearest miss, denotes the neighbor samples of s_i in the sample set S where the samples have the different class labels as s_i. n denotes the number of samples generated randomly. In order to search the nearest neighbor sample, this paper uses the formula (8) to calculate the distance of different samples.

In general, the feature weight calculated by the Relief algorithm is positively correlated with the prediction ability of the corresponding feature. According to the weights from highest to lowest, the considered features can be sorted.

In order to obtain the optimal feature subset, the IFS (Incremental Feature Selection) method, a proverbial searching strategies in feature selection, is adopted in this study. Based on the feature ranking, the IFS method is implemented in the following steps: First, generate an empty feature subset, and then add the features to the feature subset one by one with the weight from highest to lowest. At each iteration, with a new feature added, a new feature subset is generated to construct a new prediction model. The feature subset that achieves the highest prediction accuracy will be selected as the optimal feature subset.

Machine Learning Method. *Random Forest.* The random forest (RF) algorithm, proposed by Breiman[49], is a supervised learning algorithm that has been successfully employed in classification problems[50,51] and achieves satisfactory performance. It is an ensemble classifier generating a multitude of decision trees, where each decision tree is constructed based on the benchmark dataset and produces a classification label. To predict a test sample, its feature vector is put into each of the decision trees in the forest, and each tree gives a vote suggesting one class. The predicted result of the RF is decided based on the most votes given by all the individual trees[52]. RF can reduce the output variance of individual trees, and thus improves the stability and accuracy of classification. In addition, it is relatively robust to noise and outliers[49].

Radial Basis Function Network. The radial basis function network (RBFNetwork) is suitable for solving function approximation and pattern classification problems due to its faster training procedure and better approximation capabilities[53]. In the classical RBFNetwork, there is an input layer, a hidden layer with a non-linear RBF activation function, and a linear output layer[54]. It uses the k-means clustering to provide the basis functions. The basis functions are usually chosen as Gaussian and the number of hidden units are fixed a priori using some properties of input data. RBFNetworks have advantages of strong tolerance to noise and good generalization[55].

Naïve Bayes. Naïve bayes (NB) is generally known as a simple probabilistic classifier, which has been successfully used in the realm of bioinformatics[56,57]. The naïve bayes assumes the attribute variables to be independent from each other, which can greatly reduce the complexity of the development of the classifier.

Logistic Regression. The crucial limitation of linear regression is that it cannot deal with dependent variables that are dichotomous and categorical. Logistic regression (LR) is an effective method to find the best fitting model to describe the relationship between the categorical dependent variable and a set of independent numeric variables[58].

Nearest Neighbor Algorithm. Nearest neighbor algorithm (NNA) is a machine learning technique based on cluster theory. Despite its simplicity, NNA often performs nearly as well as more sophisticated methods. Based on the NNA classification principle, a new sample is assigned to the same class as the one in the benchmark dataset that is nearest to the query sample[59].

Classifier Fusion. Every single learning strategy has its own shortcomings and could not always perform well on all datasets[60]. The classifier fusion emerges as a promising measure to overcome this problem[28,61]. A fusion of classifiers is a collection of multiple basic individual classifiers with diverse learning policies and then aggregates the outputs of all independent classifiers to tackle the same classification task[62]. In general, the outputs of different single classifiers tend to be different for a given classification problem. But at the same time they have the ability to correct each other's mistakes. Therefore, the prediction ability of classifier fusion is usually superior to that of its component single classifier[63]. Hansen LK and Salamon P[64] has theoretically demonstrated that the classifier fusion gives much better performance compared to its base classifiers.

In this study, we evaluate prediction performance of different classifiers including radial basis function network, naïve bayes, logistic regression, nearest neighbor algorithm and random forest, respectively. Then the ultimate result is determined by the average probability of the outputs obtained from one classifier which is good at predicting negative class (with a higher specificity) and another one which is good at predicting positive class (with a higher sensitivity). WEKA machine learning platform[65] is used for implementing all the algorithms and the classifier fusion method.

Performance Evaluation. Independent dataset test, jackknife test, and sub-sampling test are the 3 common methods to measure the performance of a predictor[66]. For a given prediction problem, the output result generated by the jackknife test is unique while the other 2 methods are not[67,68]. Therefore, the jackknife test can obtain a more strict and objective prediction result, which make it extensively applied to verify the performance

of prediction models[27,69]. For the purpose of reducing the complexity of computing, 10-fold cross validation test[24], one of sub-sampling test, is used to measure the performance of the anti-angiogenic peptide predictors.

Based on the prediction result generated by the 10-fold cross validation test, the following evaluation indexes are calculated to compare the proposed method with the existing method.

Sensitivity (*Sn*) represents the prediction accuracy of anti-angiogenic peptides, which is expressed as:

$$S_n = \frac{TP}{TP + FN}, \tag{9}$$

Specificity (*Sp*) represents the prediction accuracy of non-anti-angiogenic peptides, which is given by:

$$S_p = \frac{TN}{TN + FP}, \tag{10}$$

Accuracy (*Acc*) represents the overall prediction accuracy of all samples in the dataset, which is defined as:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}, \tag{11}$$

Matthew's correlation coefficient (MCC)[70] is another effective measure for performance evaluation and calculated as:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)\,(TP + FP)\,(TN + FP)\,(TN + FN)}}, \tag{12}$$

where *TP*, *TN*, *FP*, and *FN* denote number of correctly predicted anti-angiogenic peptides, number of correctly predicted non-anti-angiogenic peptides, number of non-anti-angiogenic peptides incorrectly predicted as anti-angiogenic peptides, and numer of anti-angiogenic peptides incorrectly predicted as non-anti-angiogenic peptides, respectively.

To provide more insight into the prediction performance for anti-angiogenic peptides, the receiver operating characteristic (ROC) curve[71] is plotted, and the area under the ROC curve (AUC) is calculated. The prediction model with a higher AUC value indicates that it achieves a better prediction performance[49].

Conclusions

Anti-angiogenic peptides are thought to have physiological functions and excellent therapeutic potential for angiogenesis-related diseases. Identification of anti-angiogenic peptides accurately may not only contribute to better understanding essential angiogenic homeostasis within tissues, but also provide significant clues to develop antineoplastic therapies. To identify anti-angiogenic peptides, an ensemble learning method has been presented in this study by fusing an individual classifier with the best sensitivity and another classifier with the best specificity. To decrease the complexity of computation, the Relief algorithm followed by the IFS method is employed to eliminate the redundant features. Based on the benchmark dataset, the accuracy of various feature spaces (i.e., BpB, CTD, DFT) with respect to the corresponding optimal individual classifiers lies in the range of 0.636 to 0.804, indicating discriminative power of features. The accuracy, MCC, and AUC of BpB with an NB classifier are 0.804, 0.626, and 0.902, respectively, which represents the highest prediction results among the various feature spaces, demonstrating that position-specific statistical differences at the N and C-terminal region are suitable to identify anti-angiogenic peptides. The accuracy of BpB on the ensemble classifier (i.e., NB + LR) is 0.822, revealing that an appropriate ensemble classifier can effectively improve prediction performance. In addition, by means of the Relief-IFS, the sensitivity, specificity, accuracy, MCC, and AUC of the prediction model are 0.776, 0.888, 0.832, 0.668, and 0.872, respectively, better than those of the prediction model using all features. Performance comparisons with the existing method on the same dataset indicate that the proposed ensemble method is effective in predicting anti-angiogenic peptides.

References

1. Sacewicz, I., Wiktorska, M., Wysocki, T. & Niewiarowska, J. Mechanisms of cancer angiogenesis. *Postepy Hig. Med. Dosw.* **63**, 159–168 (2009).
2. Sulochana, K. N. & Ge, R. Developing antiangiogenic peptide drugs for angiogenesis-related diseases. *Curr. Pharm. Des.* **13**, 2074–2086 (2007).
3. Carmeliet, P. Mechanisms of angiogenesis and arteriogenesis. *Nat. Med.* **6**, 389–395 (2000).
4. Folkman, J. Angiogenesis: an organizing principle for drug discovery? *Nature Rev. Drug Discov.* **6**, 273–286 (2007).
5. Chuang, I. C. *et al.* The anti-angiogenic action of 2-deoxyglucose involves attenuation of VEGFR2 signaling and MMP-2 expression in HUVECs. *Life Sci.* **139**, 52–61 (2015).
6. Chiavacci, E. *et al.* The zebrafish/tumor xenograft angiogenesis assay as a tool for screening anti-angiogenic miRNAs. *Cytotechnology.* **67**, 969–975 (2015).
7. Robinet, A. *et al.* Elastin-derived peptides enhance angiogenesis by promoting endothelial cell migration and tubulogenesis through upregulation of MT1-MMP. *J. Cell. Sci.* **118**, 343–356 (2005).
8. Schneider, B. P. & Miller, K. D. Angiogenesis of breast cancer. *J. Clin. Oncol.* **23**, 1782–1790 (2005).
9. Rosca, E. V. *et al.* Anti-angiogenic peptides for cancer therapeutics. *Curr. Pharm. Biotechnol.* **12**, 1101–1116 (2011).
10. Tozer, G. M., Kanthou, C. & Baguley, B. C. Disrupting tumour blood vessels. *Nat. Rev. Cancer.* **5**, 423–435 (2005).
11. Albini, A., Tosetti, F., Li, V. W., Noonan, D. M. & Li, W. W. Cancer prevention by targeting angiogenesis. *Nat. Rev. Clin. Oncol.* **9**, 498–509 (2012).
12. Nakamura, T. & Matsumoto, K. Angiogenesis inhibitors: from laboratory to clinical application. *Biochem. Biophys. Res. Commun.* **333**, 289–291 (2005).
13. Wijngaarden, P. V., Coster, D. J. & Williams, K. A. Inhibitors of ocular neovascularization: promises and potential problems. *JAMA.* **293**, 1509–1513 (2005).

14. Ruoslahti, E., Duza, T. & Zhang, L. Vascular homing peptides with cell-penetrating properties. *Curr. Pharm. Des.* **11**, 3655–3660 (2005).
15. Sitohy, B., Nagy, J. A. & Dvorak, H. F. Anti-VEGF/VEGFR therapy for cancer: reassessing the target. *Cancer Res.* **72**, 1909–1914 (2012).
16. Yi, J. M., Bang, O. S. & Kim, N. S. An evaluation of the anti-angiogenic effect of the Korean medicinal formula "Sa-mi-yeon-geon-tang" *in vitro* and in ovo. *BMC Complement Altern Med*. **15** (2015).
17. Yuan, D. *et al*. Anti-angiogenic efficacy of 5′-triphosphate siRNA combining VEGF silencing and RIG-I activation in NSCLCs. *Oncotarget*. **6**, 29664–29674 (2015).
18. Manegold, C. *et al*. Randomized phase II study of three doses of the integrin inhibitor cilengitide versus docetaxel as second-line treatment for patients with advanced non-smallcell lung cancer. *Invest. New Drugs.* **31**, 175–182 (2013).
19. Koskimaki, J. E. *et al*. Pentastatin-1, a collagen IV derived 20-mer peptide, suppresses tumor growth in a small cell lung cancer xenograft model. *BMC Cancer*. **10** (2010).
20. Yi, Z. F. *et al*. A novel peptide from human apolipoprotein(a) inhibits angiogenesis and tumor growth by targeting c-Src phosphorylation in VEGF-induced human umbilical endothelial cells. *Int. J. Cancer.* **124**, 843–852 (2009).
21. Chlenski, A. *et al*. Anti-angiogenic SPARC peptides inhibit progression of neuroblastoma tumors. *Mol. Cancer*. **9** (2010).
22. Karagiannis, E. D. & Popel, A. S. A systematic methodology for proteome-wide identification of peptides inhibiting the proliferation and migration of endothelial cells. *Proc. Natl. Acad. Sci. USA* **105**, 13775–13780 (2008).
23. Johnson, M. *et al*. NCBI BLAST: a better web interface. *Nucleic. Acids Res*. **36** (2008).
24. Ettayapuram Ramaprasad, A. S., Singh, S., Gajendra, P. S. R. & Venkatesan, S. AntiAngioPred: a server for prediction of anti-angiogenic peptides. *PLoS One.* **10**, e0136990 (2015).
25. Qian, J., Miao, D. Q., Zhang, Z. H. & Li, W. Hybrid approaches to attribute reduction based on indiscernibility and discernibility relation. *International Journal of Approximate Reasoning.* **52**, 212–230 (2011).
26. Wang, P. & Xiao, X. NRPred-FS: a feature selection based two level predictor for nuclear receptors. *J. Proteomics Bioinform*. S9 (2014).
27. Dehzangi, A., Phon-Amnuaisuk, S. & Dehzangi, O. Enhancing protein fold prediction accuracy by using ensemble of different classifiers. *Australian Journal of Intelligent Information Processing Systems.* **26**, 32–40 (2010).
28. Si, J., Zhang, Z., Lin, B., Schroeder, M. & Huang, B. MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. *BMC Syst. Biol.* **5**, S7 (2011).
29. Chen, X. & Huang, L. LRSSLMDA: laplacian regularized sparse subspace learning for miRNA-disease association prediction. *Plos Computational Biology.* **13**, e1005912 (2017).
30. Chen, X., Huang, L., Xie, D. & Zhao, Q. EGBMMDA: extreme gradient boosting machine for miRNA-disease association prediction. *Cell Death & Disease.* **9**, 3 (2018).
31. Chen, X., Zhou, Z. & Zhao, Y. ELLPMDA: ensemble learning and link prediction for miRNA-disease association prediction. *RNA Biology.* **25**, 1–12 (2018).
32. Wang, L. *et al*. An ensemble approach for large-scale identification of protein- protein interactions using the alignments of multiple sequences. *Oncotarget.* **8**, 5149–5159 (2017).
33. Li, J. Q., You, Z. H., Li, X., Ming, Z. & Chen, X. PSPEL: in silico prediction of self-interacting proteins from amino acids sequences using ensemble learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* **14**, 1165–1172 (2017).
34. Chou, K. C. Some remarks on protein attribute prediction and pseudo amino acidcomposition. *J. Theor. Biol.* **273**, 236–247 (2011).
35. Ali, S., Majid, A. & Khan, A. IDM-PhyChm-Ens: intelligent decision-making ensemble methodology for classification of human breast cancer using physicochemical properties of amino acids. *Amino Acids.* **46**, 977–993 (2014).
36. Nath, A. & Subbiah, K. Maximizing lipocalin prediction through balanced and diversified training set and decision fusion. *Comput. Biol. Chem*. **59**, 101–110 (2015).
37. Kaundal, R. & Raghava, G. P. RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information. *Proteomics.* **9**, 2324–2342 (2009).
38. Shao, J., Xu, D., Tsai, S. N., Wang, Y. & Ngai, S. M. Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PLoS One.* **4**, e4920 (2009).
39. Dings, R. P., Nesmelova, I., Griffioen, A. W. & Mayo, K. H. Discovery and development of anti-angiogenic peptides: a structural link. *Angiogenesis.* **6**, 83–91 (2003).
40. Shao, J. *et al*. PLMLA: prediction of lysine methylation and lysine acetylation by combining multiple features. *Mol. Biosyst.* **8**, 1520–1527 (2012).
41. Dubchak, I., Muchnik, I., Holbrook, S. R. & Kim, S. H. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. USA* **92**, 8700–8704 (1995).
42. Hou, T. *et al*. LAceP: Lysine acetylation site prediction using logistic regression classifiers. *PLoS One.* **9**, e89575 (2014).
43. Panda, B., Mishra, A. P., Majhi, B. & Rout, M. Prediction of protein structural class by functional link artificial neural network using hybrid feature extraction method. *SEMCCO (2), Springer, In Bijaya Ketan Panigrahi; Ponnuthurai Nagaratnam Suganthan; Swagatam Das & Subhransu Sekhar Dash*. **8298**, 298–307 (2013).
44. Sahu, S. S. & Panda, G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Comput. Biol. Chem.* **34**, 320–327 (2010).
45. Hoang, T. *et al*. A new method to cluster DNA sequences using Fourier power spectrum. *J. Theor. Biol.* **372**, 135–145 (2015).
46. Zhan, T. L. & Ding, Y. S. Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. *Amino Acids.* **33**, 623–629 (2007).
47. Kira, K. & Rendell, L. A. The feature selection problem: traditional methods and a new algorithm. *In Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, USA*, *12–16 July*, 12–134 (1992).
48. Sun, Y. Iterative RELIEF for feature weighting: Algorithms, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **26**, 1035–1051 (2007).
49. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
50. Li, C., Wang, X. F., Chen, Z., Zhang, Z. & Song, J. Computational characterization of parallel dimeric and trimeric coiled-coils using eff ective amino acid indices. *Mol. BioSyst.* **11**, 354–360 (2015).
51. Li, Y. *et al*. Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features. *Sci. Rep.* **4**, 57–65 (2014).
52. Lou, W. C. *et al*. Sequence based prediction of dna-binding proteins based on hybrid feature selection using random forest and gaussian Naïve Bayes. *PLoS One.* **9**, e86703 (2014).
53. Samantray, S. R., Dash, P. K. & Panda, G. Fault classification and location using HS-transform and radial basis function neural network. *Electric Power Syst. Res.* **76**, 897–905 (2006).
54. Yuan, L. F. *et al*. Prediction of the types of ion channel-targeted conotoxins based on radial basis function network. *Toxicology in Vitro.* **27**, 852–856 (2013).
55. Yu, H., Xie, T., Paszczyñski, S. & Wilamowski, B. M. Advantages of radial basis function networks for dynamic system design. *IEEE Transactions on Industrial Electronics.* **58**, 5438–5450 (2011).
56. Murakami, Y. & Mizuguchi, K. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics.* **26**, 1841–1848 (2010).

57. Sambo, F., Trifoglio, E., Di Camillo, B., Toffolo, G. M. & Cobelli, C. Bag of Naïve Bayes: biomarker selection and classification from genome-wide SNP data. *BMC Bioinformatics*. **13** (2012).
58. Peng, C. Y. J., Lee, K. L. & Ingersoll, G. M. An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*. **96**, 3–14 (2002).
59. Hall, P., Park, B. U. & Samworth, R. J. Choice of neighbor order in nearest-neighbor classification. *Annals of Statistics*. **36**, 2135–2152 (2008).
60. Zou, C., Gong, J. & Li, H. An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis. *BMC Bioinformatics*. **14** (2013).
61. Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **33**, 1–39 (2010).
62. Xu, R. F. *et al*. enDNA-Prot: Identification of dna-binding proteins by applying ensemble learning. *BioMed Res. Int* (2014).
63. Lo, S. L., Chiong, R. & Cornforth, D. Using support vector machine ensembles for target audience classification on Twitter. *PLoS One*. **10**, e0122855 (2015).
64. Hansen, L. K. & Salamon, P. Neural network ensembles. *IEEE Transactions on Pattern Analysis & Machine Intelligence*. **12**, 993–1001 (1990).
65. Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I. H. Data mining in bioinformatics using Weka. *Bioinformatics*. **20**, 2479–2481 (2004).
66. Chou, K. C. & Zhang, C. T. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **30**, 275–349 (1995).
67. Chou, K. C. & Shen, H. B. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nature Protocols*. **3**, 153–162 (2008).
68. Chou, K. C. & Shen, H. B. Recent progress in protein subcellular location prediction. *Crit. Rev. Biochem. Mol. Biol.* **370**, 1–16 (2007).
69. Ding, H. *et al*. iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *Biomed Res. Int*. 2014 (2014).
70. Ding, H., Feng, P. M., Chen, W. & Lin, H. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Mol. Biosyst.* **10**, 2229–2235 (2014).
71. Gribskov, M. & Robinson, N. L. Use of receiver operating characteristic(ROC) analysis to evaluate sequence matching. *J. Comput. Chem*. **20**, 25–33 (1996).

Acknowledgements

This research is supported by China Postdoctoral Science Foundation (Grant Nos 2017M612270 and 2018M630778) and National Natural Science Foundation of China (Grant Nos 61473335 and 61533011).

Author Contributions

L.N.Z. conceived and designed the experiments, R.T.Y. and C.J.Z. conducted the experiments, L.N.Z. and R.T.Y. analysed the results. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-32443-w.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018