Zhang *et al. BMC Bioinformatics* (2016) 17:225
DOI 10.1186/s12859-016-1087-5

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Prediction of aptamer-protein interacting pairs using an ensemble classifier in combination with various protein sequence attributes

Lina Zhang[1], Chengjin Zhang[1,2*], Rui Gao[1], Runtao Yang[1] and Qing Song[3]

Abstract

Background: Aptamer-protein interacting pairs play a variety of physiological functions and therapeutic potentials in organisms. Rapidly and effectively predicting aptamer-protein interacting pairs is significant to design aptamers binding to certain interested proteins, which will give insight into understanding mechanisms of aptamer-protein interacting pairs and developing aptamer-based therapies.

Results: In this study, an ensemble method is presented to predict aptamer-protein interacting pairs with hybrid features. The features for aptamers are extracted from Pseudo K-tuple Nucleotide Composition (PseKNC) while the features for proteins incorporate Discrete Cosine Transformation (DCT), disorder information, and bi-gram Position Specific Scoring Matrix (PSSM). We investigate predictive capabilities of various feature spaces. The proposed ensemble method obtains the best performance with Youden's Index of 0.380, using the hybrid feature space of PseKNC, DCT, bi-gram PSSM, and disorder information by 10-fold cross validation. The Relief-Incremental Feature Selection (IFS) method is adopted to obtain the optimal feature set. Based on the optimal feature set, the proposed method achieves a balanced performance with a sensitivity of 0.753 and a specificity of 0.725 on the training dataset, which indicates that this method can solve the imbalanced data problem effectively. To evaluate the prediction performance objectively, an independent testing dataset is used to evaluate the proposed method. Encouragingly, our proposed method performs better than previous study with a sensitivity of 0.738 and a Youden's Index of 0.451.

Conclusions: These results suggest that the proposed method can be a potential candidate for aptamer-protein interacting pair prediction, which may contribute to finding novel aptamer-protein interacting pairs and understanding the relationship between aptamers and proteins.

Keywords: Aptamer-protein interacting pairs, Ensemble method, Hybrid features, Imbalanced data problem

Background

Aptamers, first reported by Ellington and Gold in 1990 [1, 2], are single stranded DNA/RNA molecules or peptide molecules [3]. They can fold into specific three-dimensional configurations that bind to targets with a high specificity and regulate their activities [4, 5]. The targets include proteins, nucleic acids, drugs, organic dyes, metal ions, and even whole cells or organisms [6, 7]. Figure 1 depicts the structures of two aptamers binding to specific targets. With a deeper understanding of aptamers in terms of their conformational and protein-binding properties, aptamer-protein interacting pairs have a potential to perform a variety of functions [8, 9]. Aptamers exhibit significant advantages over antibodies in flexibility of selection, chemical stability [4], and post-modifications [10].

Since they were discovered, aptamers have garnered tremendous attention and found wide applications in

*Correspondence: cjzhang@sdu.edu.cn
[1]School of Control Science and Engineering, Shandong University, Jingshi Road No.17923, 250061 Jinan, China
[2]School of Mechanical, Electrical and Information Engineering, Shandong University at Weihai, Wenhuaxi Road No.180, 264209 Weihai, China
Full list of author information is available at the end of the article

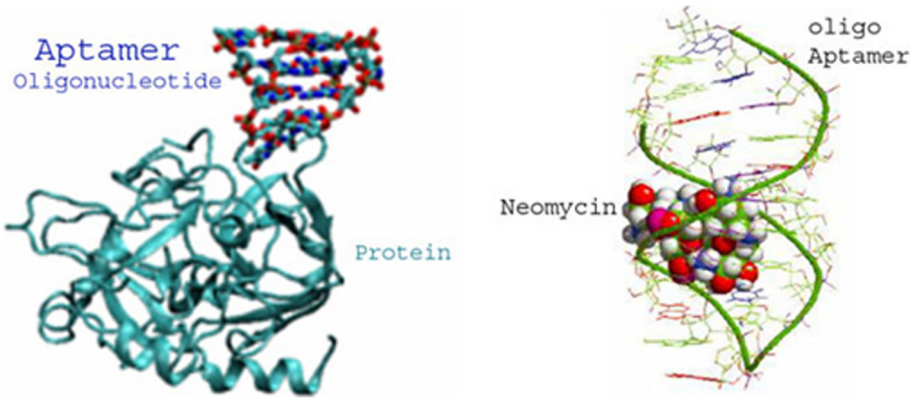

Fig. 1 The structures of aptamers binding to specific targets

biosensing, target imaging, diagnostics, and therapeutics [11, 12]. In the field of therapy, aptamers are thought to have an excellent potential in treating Age-related Macular Degeneration (AMD) [13], thrombus [14], glomerulonephritis, pulmonary hypertension, and chronic diseases [15, 16]. In the food industry, some aptamers can act as pesticides [17]. What's more, aptamers have a potential in cancer diagnosis and developing target-based therapeutic drugs delivery to cancer cells, which can reduce side effects of most chemotherapeutic drugs [12].

Due to the physiological functions and practical applications of aptamers, designing aptamers binding to certain interested proteins is crucial to gain insight into mechanisms of aptamer-protein interacting pairs and develop aptamer-based therapies for various diseases. Generally, aptamers can be artificially generated in vitro by a process commonly referred to as SELEX (Systematic Evolution of Ligands by Exponential Enrichment) [18], which consists of several repeated rounds of binding, partition, and amplification [2]. In SELEX experiments, aptamers are identified for their abilities to bind a protein of interest from libraries containing up to 10^{16} different RNA or DNA sequences [19]. Obviously, it is time-consuming and costly to design aptamers for specific proteins using experimental methods. Therefore, it would be of help to develop a computational method for rapidly and effectively predicting the aptamer-protein interacting pairs based on sequence information.

To the best of our knowledge, only one machine learning method has been reported to predict aptamer-protein interacting pairs. Li et al. [20], utilized random forest to establish the prediction model, which integrated information from nucleotide composition, amino acid composition, pseudo-amino acid composition. The maximum Relevance Minimum Redundancy (mRMR) combined with Incremental Feature Selection (IFS) strategy was applied to select high discriminative features. This method has its own merits and does facilitate the development of this field, but achieves severely imbalanced performance with a high specificity and a low sensitivity, which may be attributed to the following shortcomings. (1) Previous study merely extracted composition-based features based on the alphabetic sequences, and failed to capture some sequence-order information. Some useful features based on structural and evolutionary information are also missing. It has been a major focus in bioinformatics to integrate heterogeneous features (in some cases coupled with feature selection to remove redundant and irrelevant features from the original feature sets). For example, Li et al. [21] integrated various sources of features including protein functional domains, protein subcellular locations, and protein-protein interaction information to improve the prediction accuracy of kinase-specific phosphorylation sites. In another work, Wang et al. [22] combined heterogeneous features with a two-step feature selection procedure to improve the prediction performance of caspase substrate cleavage sites. In another recent work, Li et al. [23] attained a promising result for glycosylation site prediction by using heterogeneous feature selection. Generally, multiple features can not only preserve enough discriminative information for protein attribute prediction, but also complement each other to enhance the performance and robustness of a predictor [24]. Therefore, the combination of various features from different sources (heterogeneous features) is a good strategy for constructing classifiers [25]. (2) The existing method was based on an individual classifier whose own inherent defects would lead to unsatisfactory prediction performance [26]. In general, an ensemble predictor that integrates diverse learning policies of multiple basic classifiers can outperform its component classifiers [27]. Therefore, the ensemble predictor has been considered as a promising strategy to improve the prediction performance. (3) The previous method did not deal with the serious class imbalance problem, which would lead to a high prediction accuracy

for the majority class but a poor prediction accuracy for the minority class [28]. When there is a big difference between the number of positive samples and the number of negative samples, machine learning algorithms will not have sufficient information to learn a function to distinguish the classes due to the inherent learning biases of the imbalanced dataset [29]. Therefore, balanced dataset is needed for avoiding biases in the machine learning [30].

To address the above limitations and further improve the prediction performance, an ensemble method is developed in this paper to predict the aptamer-protein interacting pairs with Pseudo K-tuple Nucleotide Composition (PseKNC), Discrete Cosine Transformation (DCT), disorder information, and bi-gram Position Specific Scoring Matrix (PSSM). In order to reduce the computational complexity and enhance the prediction accuracy, the Relief-IFS method is employed to select high discriminative features. The ensemble random forest classifier is introduced to deal with the imbalanced dataset problem that exists in predicting the aptamer-protein interacting pairs. 10-fold cross validation is carried out to evaluate the performance of the proposed method. Our method achieves promising prediction performance with a balanced sensitivity and a specificity. Further analysis of the optimal features provides insights into the mechanisms of aptamer-protein interacting pairs.

Methods

Data collection

In order to evaluate the proposed method and facilitate its comparison with previous studies in predicting aptamer-protein interacting pairs, we use the benchmark dataset constructed recently by Li et al. [20]. The dataset is obtained from Apatmer Base [31]. It is divided into a training dataset and an independent testing dataset. The training dataset consists of 580 positive and 1740 negative samples while the independent testing dataset consists of 145 positive and 435 negative samples. The samples in the independent testing dataset are not in the training dataset. The training dataset and independent testing dataset are given in Additional file 1.

Feature extraction

An important issue in designing a predictor is how to convert an input sample sequence into a set of numerical features that are fed into a classifier. Appropriate input representations make it easier for the classifier to recognize underlying regularities, which is vital to the success of classifier learning [32]. In general, an individual feature extraction strategy can only represent partial sample's characteristics, which may limit the prediction performance. Multiple feature extraction strategies can complement each other to enhance the prediction accuracy.

Since each sample in the current dataset consists of an aptamer (DNA or RNA) and a target protein, PseKNC is adopted to formulate the aptamer sequences while hybrid features extracted from DCT, disorder information, and bi-gram PSSM are utilized for encoding target protein sequences.

Represent aptamers with pseudo K-tuple nucleotide composition

Suppose a DNA/RNA sequence D with L nucleic acid residues, i.e.

$$\begin{aligned} &D = R_1, R_2, \cdots, R_i, \cdots, R_L, \\ &R_i \in \{denine\,(A)\,, cytosine\,(C)\,, guanine\,(G)\,, \\ &thymine\,(T)\, oruracil\,(U)\}, \end{aligned} \tag{1}$$

where R_i denotes the ith nucleic acid residue along the given sequence.

Nucleic Acid Composition (NAC) is the most simple feature to encode a DNA/RNA sequence. The sequence D can be formulated by NAC as the following feature vector:

$$F_1 = [f\,(A)\,, f\,(C)\,, f\,(G)\,, f\,(T)\, orf\,(U)]\,, \tag{2}$$

where $f\,(A)\,, f\,(C)\,, f\,(G)\,, f\,(T)\, orf\,(U)$ are the normalized occurrence frequencies of the corresponding nucleotides.

In this type of representation, the sequence order information is completely lost which in turn affects the prediction performance. In order to capture local order information and global sequence-order information, Pseudo K-tuple Nucleotide Composition (PseKNC) [33, 34] is introduced here. Recent studies indicate that PseKNC have been successfully applied in identifying recombination spots [35], promoters [36], and nucleosomes [37]. In this paper, K is set as 2 for dinucleotide and 3 for trinucleotide, respectively.

As known, DNA physicochemical properties have been proved to play a significant impact on gene expression regulation [38]. Therefore, physicochemical properties of nucleotides are used to formulate PseKNC for DNA/RNA sequences. Results in [34] have shown that DNA/RNA dinucleotide physical structures, including twist, tilt, roll, shift, slide and rise, contribute to dealing with DNA/RNA sequences. Therefore, these six dinucleotide physical structures are employed to encode the pseudo 2-tuple nucleotide composition. Meanwhile, 12 physicochemical properties of trinucleotides are all included to encode the pseudo 3-tuple nucleotide composition. The values for both the 6 physicochemical properties of dinucleotides and the 12 physicochemical properties of trinucleotides can be referred to [33].

For PseKNC, the sequence-order information of a DNA/RNA sequence can be reflected by a series of correlation factors, defined as

$$\begin{cases} \theta_1 = \frac{1}{L-K}\sum_{i=1}^{L-K}\Theta_{i,i+1} \\ \theta_2 = \frac{1}{L-K-1}\sum_{i=1}^{L-K-1}\Theta_{i,i+2} \\ \theta_3 = \frac{1}{L-K-2}\sum_{i=1}^{L-K-2}\Theta_{i,i+3} \\ \cdots\cdots \\ \theta_\lambda = \frac{1}{L-K-\lambda+1}\sum_{i=1}^{L-K-\lambda+1}\Theta_{i,i+\lambda} \end{cases} \quad \lambda = L_{\min} - K, \tag{3}$$

where

$$\begin{cases} \Theta_{i,i+j} = \frac{1}{N}\sum_{n=1}^{N}\left[H_n\left(R_iR_{i+1}\cdots R_{i+K-1}\right) - H_n\left(R_{i+j}R_{i+j+1}\cdots R_{i+j+K-1}\right)\right]^2, \\ i = 1, 2, \cdots, L-K+1; j = 1, 2, \cdots, \lambda \end{cases} \tag{4}$$

where θ_λ is the λth tier correlation factor that reflects the sequence order correlation between all the λth most contiguous K-tuple nucleotides along a DNA sequence. λ is the highest rank of correlation factor along the DNA/RNA sequence, and $L_{\min}$ is the length of the DNA/RNA sequence with minimum length in the training dataset. Here, we set $\lambda = L_{\min} - K$. $\Theta_{i,i+j}$ is the correlation function; $H_n(R_iR_{i+1}\cdots R_{i+K-1})$ denotes the normalized value of the nth physicochemical property for K-tuple nucleotide $R_iR_{i+1}\cdots R_{i+K-1}$ at position i and $H_n(R_{i+j}R_{i+j+1}\cdots R_{i+j+K-1})$ the corresponding value for K-tuple nucleotide $R_{i+j}R_{i+j+1}\cdots R_{i+j+K-1}$ at position $i+j$. N is the total number of physicochemical properties for K-tuple nucleotides. Here, N equals to 6 for pseudo 2-tuple nucleotide composition and N equals to 12 for pseudo 3-tuple nucleotide composition.

Finally, a DNA/RNA sequence can be represented by a $(4^K + \lambda)$-dimensional feature vector using the PseKNC,

$$D_{PseKNC} = \left[d_1 \ \cdots \ d_{4^K} \ d_{4^K+1} \ \cdots \ d_{4^K+\lambda}\right], (\lambda = L_{\min} - K), \tag{5}$$

where

$$d_u = \begin{cases} \frac{f_u^{K-tuple}}{\sum_{i=1}^{4^K} f_i^{K-tuple} + w\sum_{j=1}^{\lambda}\theta_j} & \left(1 \le u \le 4^K\right) \\ \frac{w\theta_{u-4^K}}{\sum_{i=1}^{4^K} f_i^{K-tuple} + w\sum_{j=1}^{\lambda}\theta_j} & \left(4^K + 1 \le u \le 4^K + \lambda\right) \end{cases}, \tag{6}$$

where $f_u^{K-tuple}$ is the normalized occurrence frequency of the uth K-tuple nucleotide. ω is the weight factor.

Represent target proteins with hybrid features

Discrete cosine transform A protein sequence occasionally shows periodicity of hydrophobicity and hydrophilicity, which plays a significant role in protein attribute prediction [39]. To achieve this goal, hydrophobicity and hydrophilicity of amino acids along the protein sequence are employed and transformed into a discrete frequency domain. Then, the frequency information reflecting the periodicity, is merged into a set of discrete components which can be used to identify the distribution of the power contained in a protein sequence over the frequencies [40].

Discrete Cosine Transform (DCT), proposed by Ahmed et al. [41], is a real-valued and quasi-orthogonal transformation approach converting numerical values into frequency domain with lower computational complexities. The strong capability of the DCT to compress energy makes the DCT a good candidate for pattern recognition applications [42].

Based on the hydrophobicity or hydrophilicity of amino acids, the DCT of a given protein sequence with a length of L is formulated as

$$G(k) = a(k)\sum_{n=0}^{L-1} H\left(p_n\right)\cos\left[\frac{(2n+1)k\pi}{2L}\right], \quad k = 0, 1, 2, \cdots, L-1, \tag{7}$$

$$a(k) = \begin{cases} \sqrt{\frac{1}{L}}, k = 0 \\ \sqrt{\frac{2}{L}}, k \ne 0 \end{cases}, \tag{8}$$

where $G(k), k = 0, 1, \cdots, L-1$ represents the spectral characteristic of the sequence. $G(0)$ denotes the constant component and the remaining represent the harmonic components of the sequence.

The low-frequency components of DCT, which preserve the global information along with some sequence order information, contain more biological significance than high frequency noisy ones [39]. As the minimum length of protein sequences in the dataset is 52. For the hydrophobicity or hydrophilicity of amino acids, we use 52 low frequency DCT components to represent protein sequences.

Bi-gram position specific scoring matrix According to molecular evolution, protein sequences stem from a very finite number of ancestral species, which evolves undergoing changes, insertions, and deletions of single or several residues [43]. With the accumulation over a long period of time, many similarities between original and resultant protein sequences are gradually eliminated, but the corresponding sequences may still share some structure similarities and the same functions [44]. It is indicated that protein sequence evolutionary conservations serve as evidence for structural and functional conservations. Therefore, evolutionary conservations can determine important biological functions and are important in biological sequence analysis [45].

The position-specific score matrix (PSSM), derived from the Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) [46], is adopted to obtain the evolutionary conservations. For a given protein sequence with a length of L, the corresponding PSSM profile is composed of $L * 20$ elements defined as:

$$\mathrm{P_{PSSM}} = \begin{bmatrix} E_{1\to1} & E_{1\to2} & \cdots & E_{1\to j} & \cdots & E_{1\to20} \\ E_{2\to1} & E_{2\to2} & \cdots & E_{2\to j} & \cdots & E_{2\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ E_{i\to1} & E_{i\to2} & \cdots & E_{i\to j} & \cdots & E_{i\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ E_{L\to1} & E_{L\to2} & \cdots & E_{L\to j} & \cdots & E_{L\to20} \end{bmatrix}, \quad (9)$$

where the rows and columns of the matrix are indexed by the protein residues and the 20 native amino acids, respectively. The values in the ith row denote the probabilities of the ith residue in the given protein sequence mutating to the 20 native amino acids during the evolution process. PSSM generally contains positive or negative integers. Positive scores indicate that the given amino acid substitution occurs more frequently than expected occasionally while negative scores indicate the opposite [47]. What's more, large positive scores often represent active sites required for other intermolecular interactions [48].

We extract bi-gram features from PSSM [49] to represent protein sequences, which are defined as

$$B_{m,n} = \sum_{i=1}^{L-1} P_{i\to m} P_{i+1\to n}, (m, n = 1, 2, \cdots, 20), \quad (10)$$

where $B_{m,n}$ denotes the frequency of transition from the mth amino acid on the ith position to the nth amino acid on the $(i+1)$th position. These features incorporate neighborhood information of amino acids and evolutionary information from PSSM.

Eventually, 400 frequencies can be obtained and formulated as

$$F_{PSSM} = [B_{1,1}, \cdots, B_{1,20}, B_{2,1}, \cdots, B_{2,20}, \cdots\cdots, B_{20,1}, \cdots, B_{20,20}]. \quad (11)$$

Disorder information Protein segments are defined as unstructured or disordered if they lack stable three-dimensional structures or if they have a large number of conformations under physiological conditions [50]. Such disordered regions of proteins allow for more modification sites and flexible interaction partners. Therefore, the information of disorder regions is of great importance for the functions and structure forming of proteins [51]. In this study, VSL2 [52], which is one of the best disorder predictor and can accurately predict both long and short disordered regions in proteins, is employed to calculate the disorder score for each residue. Disorder score reflects the disorder status of each amino acid in a given protein sequence. The disorder score ranges from 0 to 1. The higher score represents the corresponding residue is more likely to lack fixed structure.

The length of disorder scores for each protein sequence is varying, which is inappropriate to develop a predictor. Auto covariance (AC), depicting the average interactions between two residues, has been successfully adopted to grasp the local discriminative information [53]. To solve the variable dimension problem, AC descriptors are adopted here to acquire more local sequence order information.

For a protein sequence with the length of L, disorder scores are obtained with the same length from VSL2, defined as

$$[d_1, \cdots, d_i, \cdots, d_L], \quad (12)$$

where d_i denotes the disorder score of the residue on the ith position along the given protein sequence.

To extract features from the disorder score, AC is defined as

$$AC_\lambda = \frac{1}{(L-\lambda)} \sum_{i=1}^{L-\lambda} (d_i - \bar{d}) * (d_{(i+\lambda)} - \bar{d}), \quad (\lambda = 1, 2, ..., L_{\min} - 1), \quad (13)$$

where $\bar{d}$ is the average value of the disorder score vector; λ is the distance between two considered amino acid residues, which is closely related to sequence order information and plays an important role in the performance of a predictor. L_{min} is the length of the protein sequence with the minimum length which equals to 52 in this study. From the above equation, 51 order-based features are calculated. To extract more disorder-based feature, the following features can be obtained. (i) mean/standard deviation of all residues disorder scores (2 features); (ii) number of disorder/non-disorder segments (2 features); (iii) minimum/maximum length of disorder/non-disorder segments (4 features). Therefore, 59 disorder-based features can be obtained to represent proteins.

Feature selection

After carrying out the above feature extraction methods, all the aptamer-protein interacting pairs with various lengths are converted into numerical feature vectors with the same dimension. However, not all the extracted features can contribute equally to classification. There may have some uncorrelated and redundant information among the extracted features, which can affect the speed and prediction performance of a predictor [54]. Feature selection techniques are essential to pick out informative features and gain deeper insights into intrinsic properties of protein sequences, which can prevent overfitting, improve the prediction quality, and build a

robust prediction model [55]. In this study, the Relief algorithm combined with Incremental Features Selection(IFS) is employed to acquire more discriminative features for predicting aptamer-protein interacting pairs.

Relief The Relief algorithm, originally proposed by Kira [56], is considered one of the most successful algorithms for depicting the relevance between the features and class labels. It is noise-tolerant and requires only linear time. The Relief algorithm can be used to estimate feature weights according to the ability of the feature to distinguish the near samples [57]. The Relief algorithm is executed iteratively. During each iteration process, the Relief algorithm endows each feature with a weight as formulated by

$$W_p^{i+1} = W_p^i - \frac{diff(Y, x_i, H(x_i))}{m} + \frac{diff(S, x_i, M(x_i))}{m}, \tag{14}$$

$$diff(*, x, y) = \begin{cases} \|x - y\|, & x \neq y \\ 0, & x = y \end{cases}, \tag{15}$$

where W_p^i, W_p^{i+1} denote the current and next weight values, respectively; p represents a given feature; x_i stands for the ith sample; $H(x_i)$ represents the nearest neighbor samples from the same class label against x_i (termed the nearest hit); $M(x_i)$ stands for the nearest neighbor samples from different class labels against x_i, (termed the nearest miss). Y and S denote the sample sets with the same and different class labels against x_i, respectively; m is the number of random samples; The function of $diff(*, x, y)$ is used for calculating the distance between the random samples to find the nearest neighbor one.

Relief endows each feature a weight value within range [0, 1]. The feature with a larger weight value indicates that it is a more highly relevant one for the target to be predicted. In other words, predicted targets have a stronger correlation with the jth feature than that with the ith feature if $W_{f_j} > W_{f_i}$.

The ranked feature list can be obtained based on feature weight values, represented as

$$F = \{f_1, f_2, \cdots, f_N\}, \tag{16}$$

where f_1 is the feature with the highest value of W, f_2 with the second highest value of W,...f_N with the lowest value of W.

The WEKA (Waikato Environment for Knowledge Analysis) software package [58] is used for the feature selection algorithm of relief, where default parameters are employed.

Incremental feature selection Base on the ranked feature list according to the relevance to the class evaluated by the relief algorithm, the incremental feature selection (IFS), one of the well-known searching strategies of feature selection, is employed to determine the optimal features [59]. The IFS procedure starts with an empty subset, and adds features one by one from higher to lower rank into the feature subset. A new feature subset is generated when another feature has been added. The ith feature subset can be formulated as

$$F_i = \{f_1, f_2, \cdots, f_i\} (1 \leq i \leq N). \tag{17}$$

For each feature subset F_i, an ensemble predictor is constructed and evaluated using 10-fold cross validation test. The IFS curve can be drawn with Youden's index values as the y-axis and index i of F_i as the x-axis. The feature subset that yields the best prediction performance is determined as the final input of the classification system.

Ensemble learning method

As illustrated in 'Data collection', the data imbalance problem exists in predicting aptamer-protein interacting pairs. Previous research has shown that imbalanced datasets are problematic when constructing classifiers [60], which would result in a high prediction accuracy for the majority class but a poor prediction accuracy for the minority class [61, 62]. For example, the predictor in [20] yields serious imbalance performance, with a high specificity of 0.922, but a very low sensitivity of 0.488, and even a relatively high accuracy of 0.813. Many researches [63, 64] extract a very small fraction of the negative samples randomly as the training data, which has changed the distribution of positive and negative samples. This method can't take full advantage of the most information in the original data, which will lead to a biased estimate of the accuracy. Therefore, the ensemble learning method is used to resolve the imbalanced problem.

An ensemble classifier is a collection of multiple basic individual classifiers with diverse learning policies, which is supposed to significantly improve the performance of a prediction method due to the fact that ensemble classifier is able to make use of the different decision boundaries generated from the individual classifiers to strategically combine the classification results [65]. Hansen [66] has demonstrated why an ensemble method gives a much better performance than its component individual classifiers in theory. In order to improve the prediction performance and deal with the data imbalance problem, we employ an ensemble classifier to predict aptamer-protein interacting pairs. The negative samples are divided into N parts, and N is determined by

$$N = {}^{N_{negative}}/_{N_{positive}}, \tag{18}$$

where $N_{negative}$ and $N_{positive}$ are the numbers of negative and positive samples, respectively. Here, N equals to 3. Then, each negative part with the same number of positive

samples is combined with the positive samples to construct a sub-training dataset. Three RF models are trained by the 3 sub-training datasets, respectively. The ultimate prediction result of the ensemble RF classifier is determined by the average probability of the outputs of the 3 RF models. This method takes advantage of the information available in the non-aptamer-protein interacting pairs as much as possible to construct the prediction model, so the prediction result is more objective. The whole procedure of the construction of the ensemble RF classifier is shown in Fig. 2.

Performance measures

In the statistical prediction, there are 3 cross-validation methods often used for examining the accuracy, including independent dataset test, sub-sampling test (e.g. 5-fold or 10-fold cross validation), and jackknife test [67]. Among these three methods, the jackknife test is deemed the most objective and rigorous one that can exclude the memory effects during the entire testing process and can always yield a unique result for a given benchmark dataset, as elucidated in [68] and demonstrated by Eq. 50 of Chou and Shen [69]. Therefore, the jackknife test has been increasingly and widely adopted by investigators to test the power of various prediction methods [70, 71]. To reduce the computational complexity and compare with the existing method objectively, 10-fold cross validation is implemented in this study.

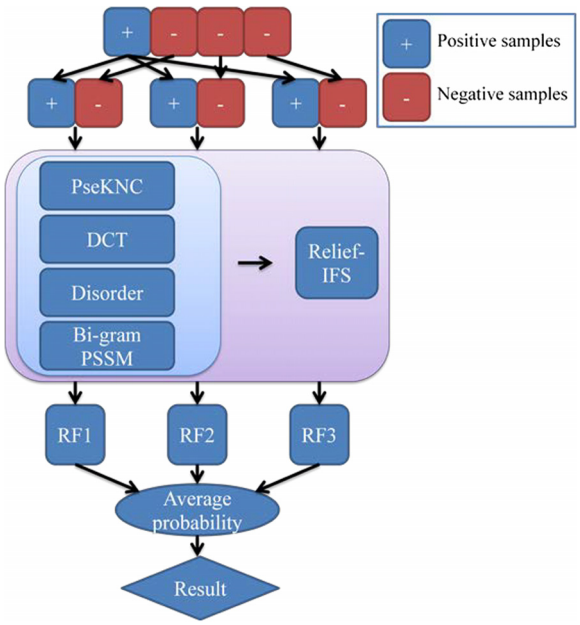

Fig. 2 The whole procedure of the ensemble RF classifier. PseKNC: Pseudo K-tuple Nucleotide Composition; DCT: Discrete Cosine Transform; Bi-gram PSSM: Bi-gram Position Specific Scoring Matrix; IFS: Incremental Feature Selection; RF: Random Forest

During the procedure, the training dataset is randomly separated into 10 equally-sized parts. Each time, one part is for testing and the other nine parts form the training dataset. This process is repeated ten times to test each part. The ultimate result is the average of the 10 prediction results. To assess the performance of the predictor intuitively, sensitivity (*Sn*), specificity (*Sp*), accuracy (*Acc*), and Matthew's Correlation Coefficient (*MCC*) are employed, which are defined as

$$Sn = \frac{TP}{FN + TP}, \tag{19}$$

$$Sp = \frac{TN}{FP + TN}, \tag{20}$$

$$Acc = \frac{TN + TP}{TN + FP + FN + TP}, \tag{21}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}, \tag{22}$$

where *TP*, *FP*, *TN* and *FN* represent true positive (correctly predicted aptamer-protein interacting pairs.), false positive (non aptamer-protein interacting pairs incorrectly predicted as aptamer-protein interacting pairs), true negative (correctly predicted non aptamer-protein interacting pairs), and false negative (aptamer-protein interacting pairs incorrectly predicted as non aptamer-protein interacting pairs), respectively.

Additionally, due to the distinct numbers of positive samples and negative samples in the training dataset, Youden's Index [72] is used for gaining insights into the relative performance of predictors in general, defined as

$$J = Sn + Sp - 1. \tag{23}$$

Youden's index gives the probability of an informed decision and is advantageous as it offers comparison of aptamer-protein interacting pair prediction quality by means of a single informative parameter [73].

Results and discussions

Performance analysis of ensemble learning method using different feature spaces

In order to explore the effectiveness of various feature spaces, the prediction results obtained by hybrid feature spaces using 10-fold cross validation are listed in Table 1. The feature space of DCT and PseKNC identifies aptamer-protein interacting pairs with a sensitivity of 0.621 and a Youden's index of 0.261. DCT, incorporating global information along with some sequence order information, results in an acceptable discrimination power. The feature space of PseKNC and bi-gram PSSM discriminates aptamer-protein interacting pairs with the best performance among the first 3 feature spaces with

Table 1 Prediction performance of the ensemble RF models using various feature spaces by 10-fold cross validation

Features	*Sn*	*Sp*	*Acc*	*MCC*	Youden's index
PseKNC+DCT	0.621	0.641	0.636	0.229	0.261
PseKNC+Bi-gram PSSM	0.660	0.673	0.670	0.293	0.333
PseKNC+Disorder	0.103	0.321	0.267	-0.499	-0.575
PseKNC+DCT+Bi-gram PSSM	0.693	0.666	0.673	0.315	0.359
PseKNC+DCT +Disorder	0.597	0.616	0.611	0.186	0.213
PseKNC+Bi-gram PSSM+Disorder	0.671	0.675	0.674	0.304	0.345
PseKNC+DCT+Bi-gram PSSM+Disorder	0.7	0.680	0.685	0.334	0.380

a sensitivity of 0.660 and a Youden's index of 0.333. Bi-gram PSSM, considering evolution and order information of the protein sequences, yields a satisfactory prediction performance. The bi-gram PSSM information also has shortcomings. The generation of PSSM of a protein depends largely on the searching dataset. If no homologous sequence is found in the searching dataset, the PSSM can not be obtained [74]. In the implementation process of our proposed method, when there is no homologous of a given protein in search dataset, we assign a zero matrix to the PSSM of the protein. As a minority of sequences have no homologous sequences in the benchmark dataset, the overall prediction performance of the ensemble method will not be affected. In Table 1, the discrimination power of disorder is weaker compared to that of the other two feature spaces, due to the fact that the sequence order information based on disorder along the sequence may not have enough information for identifying aptamer-protein interacting pairs.

As shown in Table 1, the hybrid feature space of PseKNC, DCT and bi-gram PSSM achieves a better prediction performance compared to that of PseKNC+DCT and that of PseKNC+bi-gram PSSM. The same result occurs in the hybrid feature space of PseKNC, bi-gram PSSM and disorder. However, the hybrid feature space of PseKNC, DCT and disorder obtains a sensitivity of 0.597 and a Youden's index of 0.213, worse than those of PseKNC+DCT, but better than those of PseKNC+disorder. This phenomenon may be due to that disorder introduces some redundancy features in the hybrid feature space of PseKNC, DCT and disorder. Furthermore, the hybrid feature space of PseKNC incorporating DCT, bi-gram PSSM and disorder yields the highest sensitivity of 0.7 and the highest Youden's index of 0.380, indicating the powerful discriminant ability of the ensemble method using the hybrid feature space. Other measures also show the case. These results reveal that different feature spaces extract diverse types of information from different sources and contribute to the prediction accuracy differently. Any feature spaces that may show a poor performance on certain protein attribute prediction cannot be declared as non-discriminative features. They may contain some important information that might be missed by other powerful feature extraction techniques. The hybrid feature spaces can complement each other to enhance the prediction performance of a predictor. Therefore, this study uses the hybrid feature space of PseKNC combining DCT, bi-gram PSSM and disorder to construct the prediction model.

Solving imbalanced dataset problem

Based on the results of individual RF modules, the ensemble RF classifier attempts to combine different models into a consensus classifier by the average probability of the outputs of the 3 RF models. To evaluate the effectiveness of our ensemble method to overcome the imbalanced problem, Table 2 shows the prediction results with or without the ensemble method by means of the hybrid feature space of PseKNC combining DCT, bi-gram PSSM and disorder.

As shown in Table 2, without the ensemble method, the accuracy and specificity achieve as high as 0.819 and 0.993, respectively. But the sensitivity and Youden's index are only 0.3 and 0.293, respectively. The ensemble method achieves a more balanced sensitivity of 0.7 and specificity of 0.680. The value of Youden's index is 0.380, better than that without ensemble method. The accuracy and *MCC* obtained with the ensemble method are lower than those without ensemble method, which may be due to the imbalanced data size. For the classification of imbalanced data, accuracy and *MCC* are both not appropriate measures because they may be still high when the sensitivity is very low. These results suggest that the ensemble method can solve the imbalanced problem effectively.

Feature selection results

The ranked feature list of the hybrid feature space of PseKNC combining DCT, bi-gram PSSM and disorder is

Table 2 Prediction results with or without the ensemble method

Method	*Sn*	*Sp*	*Acc*	*MCC*	Youden's index
With ensemble	0.7	0.680	0.685	0.334	0.380
Without ensemble	0.3	0.993	0.819	0.465	0.293

obtained according to their relevance to the classes based on the Relief method. Within the list (see Additional file 2), a feature with a smaller index represents a more important one for aptamer-protein interacting pair prediction. The feature list is utilized to select the optimal feature subset in the following IFS procedure. Based on the ranked feature list, adding the ranked features one by one, individual predictors for different feature subsets are constructed using the ensemble FR classifier and evaluated by 10-fold cross validation. The IFS results are given in Additional file 3. Then, the IFS curve is plotted in Fig. 3, which shows the relationships of feature indices against Youden's index. The curve reaches its peak at 0.479 when the top 304 features in Additional file 2 are selected. Thus, these 304 features are regarded as the optimal features for the ensemble RF classifier.

To investigate the influence of feature selection on the performance of the ensemble RF classifier, the prediction performance of the ensemble method with and without feature selection based on hybrid feature space of PseKNC combining DCT, bi-gram PSSM and disorder is shown in Table 3. As can be seen from Table 3, the ensemble method with feature selection achieves a sensitivity of 0.753, a specificity of 0.725, an accuracy of 0.732, a *MCC* of 0.424, and a Youden's index of 0.479 based on the 304 features, which are all superior to those of the ensemble method without feature selection. These results demonstrate that the original feature set really contains redundant information or noise. The Relief-IFS method can significantly remove these useless features to greatly improve the performance of the ensemble model. The ensemble learning method with feature selection is determined as the final predictor for aptamer-protein interacting pair prediction.

Analysis of the optimal features

The feature type distributions of the original features and the optimal features are investigated and shown in Fig. 4. There are 57 PseKNC features, 57 disorder features, 104 DCT features, and 86 bi-gram PSSM features in the optimal feature set, indicating that all kinds of features contribute to the prediction of aptamer-protein interacting pairs. The percentages of the optimal features accounting for the corresponding feature types are also investigated, which are 0.626 for PseKNC, 0.966 for disorder, 1.00 for DCT and 0.215 for bi-gram PSSM. It is interesting to note that all DCT features are in the optimal feature set, indicating that DCT based features play a crucial role in predicting aptamer-protein interacting pairs. This is the first attempt to employ DCT based features for aptamer-protein interacting pair prediction, which may help provide new annotations for the properties of these interaction pairs. An overwhelming majority of disorder features (0.966) are selected as the optimal features. It is suggested that disorder based features act an irreplaceable role in the prediction of aptamer-protein interacting pairs. These results indicate that disordered regions of a protein are closely related with the formation of the interaction of an aptamer and a protein, which is in accordance with the statement that disorder information of proteins are of great importance for the functions and structure forming. More than half of features are selected from PseKNC (0.626). This implies that the nucleotide composition and order information play some roles in predicting aptamer-protein interacting pairs. It is noted that a minority of bi-gram PSSM features (0.215) are selected from the original bi-gram PSSM features, due to the fact that the number of this feature type in the original feature set is the most of all those of other feature types. Results in

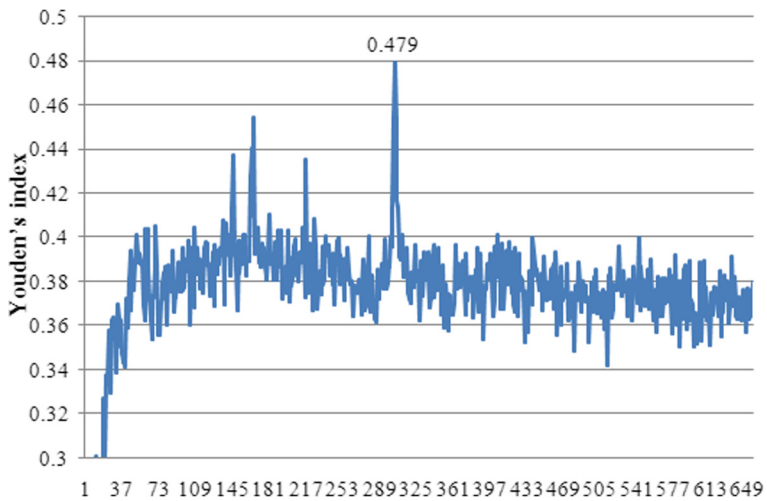

Fig. 3 The IFS(Incremental Feature Selection) curve. The values of Youden's index against the number of features

Table 3 Performance of the ensemble FR classifier with and without feature selection

Method	No. of features	*Sn*	*Sp*	*Acc*	*MCC*	Youden's index
Without feature selection	654	0.7	0.680	0.685	0.334	0.380
With feature selection	304	0.753	0.725	0.732	0.424	0.479

'Performance analysis of ensemble learning method using different feature spaces' indicate that they perform a non-negligible role in improving the prediction performance of the ensemble method.

Comparison with existing method

The existing method for identifying aptamer-protein interacting pairs [20] present prediction results by using the same size dataset (580 positive and 1740 negative samples) and same validation method (10-fold cross validation). To evaluate the prediction performance objectively, we compare our method with reference [20] on the training dataset. The performance comparison based on the same dataset is much more reliable, which can reflect the performance of a predictor more objectively. Table 4 reports the detailed prediction results obtained by the aforementioned 2 methods. As we can see from Table 4, [20] obtains an imbalanced performance with a low sensitivity of 0.488, but a high specificity of 0.922, indicating that positive samples tend to be identified incorrectly as the negative ones. This study achieves a balanced performance, with a sensitivity of 0.753 and a specificity of 0.725. The sensitivity of this study is far better than that of reference [20]. It is noted that due to the large number of negative samples, the negative samples tend to be identified correctly, which will lead to a large Acc value and a large *MCC* value as given in [20]. As mentioned above, *Acc* and *MCC* are not proper and objective indexes for this serious data imbalance problem. In addition, Youden's index of this study is better than that of [20]. Overall, the proposed ensemble method achieves a satisfactory performance and can play a complementary role to identify aptamer-protein interacting pairs.

To further assess the prediction performance of the proposed method, it is essential to compare the performance of the present method with that of the previous predictor on the same independent testing dataset. The prediction results are summarized in Table 5. In Table 5, though the specificity (0.871) yielded by [20] is better than that (0.713) obtained by our method, the sensitivity (0.483) of [20] is far worse than that (0.738) of our method, which indicates that the imbalance between sensitivity and specificity exists in [20]. Our method achieves a balanced performance with sensitivity of 0.738 and specificity of 0.713, which is also proved by the Youden's index of 0.451. It is worth pointing out that the proposed ensemble method has a fairly good prediction performance and prediction robust in predicting aptamer-protein interacting pairs.

Case study

In the case study section, we select two aptamer-protein interacting pairs identified correctly by our proposed method and analyze their physiological functions. For example, 17155909-human interleukin-23-2 interacting with human-interleukin-23 [75], an aptamer-protein interacting pair, can perform functions in congenital immunity and make a response to infection in organisms. It may not only be responsible for autoimmune

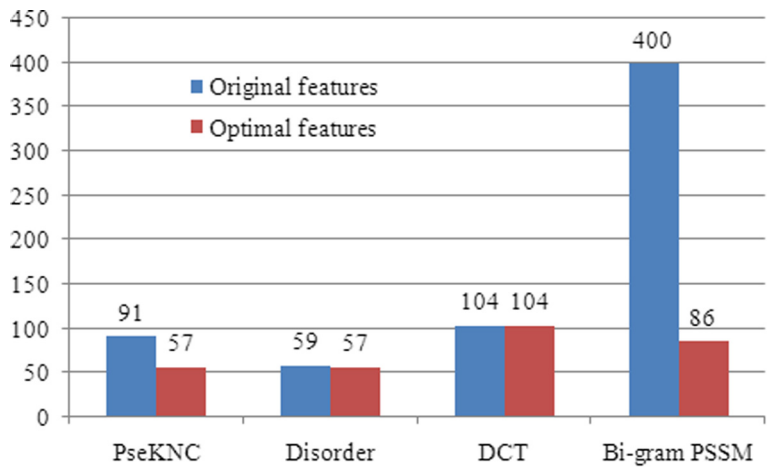

Fig. 4 The feature type distribution of the original features and the optimal features

Table 4 Performance comparison with the existing method on the training dataset by 10-fold cross validation

Method	*Sn*	*Sp*	*Acc*	*MCC*	Youden's index
[20]	0.488	0.922	0.813	0.461	0.410
This study	0.753	0.725	0.732	0.424	0.479

inflammatory diseases but also be important for tumorigenesis [75]. Another aptamer-protein interacting pair, 20387790-PAI-1-2 interacting with plasminogen-activator-inhibitor-1 [76], may function as a major control point in the regulation of fibrinolysis and blood coagulation system, regarded as a key marker for cardiovascular diseases [77]. Our proposed method can effectively identify aptamer-protein interacting pairs annotated and reviewed using experimental methods, which is of great theoretical significance in guiding research on aptamer-protein interacting pairs and relevant therapy.

Conclusions

In this paper, an ensemble method has been presented with a combination of sequence descriptors extracted from PseKNC, DCT, disorder information, and bi-gram PSSM to predict the aptamer-protein interacting pairs. To solve the dimension disaster and improve the prediction capability of the model, the Relief-IFS method is adopted to obtain the optimal feature set. By investigating predictive capabilities of various feature spaces, the proposed ensemble method obtains the best sensitivity of 0.7, specificity of 0.680, and Youden's Index of 0.380, with the hybrid feature space of PseKNC, DCT, bi-gram PSSM, and disorder information by 10-fold cross validation. These results indicate that the hybrid feature space can complement each other to enhance the prediction performance and the ensemble method can solve the imbalanced problem effectively. The Relief-IFS method can significantly remove useless features to greatly improve the performance of the ensemble model. Analysis of optimal features reveals that all feature types play roles in the determination of aptamer-protein interacting pairs, which may help understand the mechanism of aptamer-protein interactions and provide guidelines for experimental validation. To evaluate the prediction performance objectively, the proposed method is compared with previous study on the same training dataset and independent testing dataset, respectively. Our method obtains a balanced performance. The sensitivity yielded by our method is far better than that achieved by the previous method. In addition, Youden's index of this study is better than that of the existing method. It is convinced that the proposed method is an effective and powerful approach for predicting aptamer-protein interacting pairs. Since user-friendly and publicly accessible webservers represent the future direction for developing more practically predictors, we will attempt to provide a webserver in our future work for the method presented in this paper.

Table 5 Performance comparison with the existing method on the independent testing dataset

Method	*Sn*	*Sp*	*Acc*	*MCC*	Youden's index
[20]	0.483	0.871	0.774	0.372	0.354
This study	0.738	0.713	0.719	0.398	0.451

Additional files

Additional file 1: The benchmark dataset. The training dataset contains 580 positive and 1740 negative samples while the independent testing dataset consists of 145 positive and 435 negative samples. (XLS 441 kb)

Additional file 2: The ranked feature list given by the Relief algorithm. Within the list, a feature with a smaller index indicates that it is more important for aptamer-protein interacting pair prediction. Such a list of ranked features are used to establish the optimal feature set in the IFS procedure. (XLS 56.5 kb)

Additional file 3: The Incremental Feature Selection (IFS) result. By adding features one by one from higher to lower rank, 654 different feature subsets are obtained. The ensemble predictor is then accordingly built for each feature subset and evaluated by 10-fold cross validation. (XLS 125 kb)

Abbreviations

AMD, age-related macular degeneration; Bi-gram PSSM, bi-gram position specific scoring matrix; DCT, discrete cosine transform; IFS, incremental feature selection; mRMR, maximum relevance minimum redundancy; MCC, Matthew's correlation coefficient; NAC, nucleic acid composition; PseKNC, pseudo K-tuple nucleotide composition; PSI-BLAST, position-specific iterative basic local alignment search tool; RF, random forest; SELEX, systematic evolution of ligands by exponential enrichment

Acknowledgements

This research is supported by the NSFC (National Nature Science Foundation of China) under grant No.61174044, 61473335, and 61174218, Natural Science Foundation of Shandong Province of China under Grant No. ZR2015PG004, and the Doctoral Foundation of University of Jinan under Grant No. XBS1334. We also would like to thank Apatmer Base, WEKA, VSL2, and PSI-BLAST for supplying related data applied in this study.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files.

Authors' contributions

Conceived and designed the experiments: LNZ CJZ. Performed the experiments: LNZ RTY. Analyzed the data: LNZ CJZ RG QS. Contributed reagents/materials/analysis tools: RG RTY. Wrote the paper: LNZ CJZ RG RTY QS. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not Applicable.

Ethics approval and consent to participate

Not Applicable.

Author details

[1]School of Control Science and Engineering, Shandong University, Jingshi Road No.17923, 250061 Jinan, China. [2]School of Mechanical, Electrical and Information Engineering, Shandong University at Weihai, Wenhuaxi Road

No.180, 264209 Weihai, China. [3]School of Electrical Engineering, University of Jinan, Nanxinzhuangxi Road No.336, 250022 Jinan, China.

Received: 7 January 2016 Accepted: 17 May 2016
Published online: 31 May 2016

References

1. Ellington AD, Szostak JW. In vitro selection of RNA molecules that bind specific ligands. Nature. 1900;346(6287):818–22.
2. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science. 1990;249(4968):505–10.
3. Wang TJ. Function and dynamics of aptamers: A case study on the malachite green aptamer. Graduate Theses and Dissertations. 2008.
4. Keefe AD, Pai S, Ellington A. Aptamers as therapeutics. Nat Rev Drug Discov. 2010;9(7):537–50.
5. Dupont DM, Andersen LM, Botkjaer KA, Andreasen PA. Nucleic acid aptamers against proteases. Curr Med Chem. 2011;18(27):4139–51.
6. Shangguan D, Li Y, Tang Z, Cao ZC, Chen HW, Mallikaratchy P, et al. Aptamers evolved from live cells as effective molecular probes for cancer study. Proc Natl Acad Sci. 2006;103(32):11838–43.
7. Stojanovic MN, Landry DW. Aptamer-based colorimetric probe for cocaine. J Am Chem Soc. 2002;124(33):9678–9.
8. Weigand JE, Suess B. Aptamers and riboswitches: perspectives in biotechnology. Appl Microbiol Biotechnol. 2009;85(2):229–36.
9. Liu MZ, Kagahara T, Abe H, Ito Y. Direct In Vitro Selection of Hemin-Binding DNA Aptamer with Peroxidase Activity. Bull Chem Soc Jpn. 2009;82(1):99–104.
10. Song SP, Wang LH, Li J, Fan CH, Zhao JL. Aptamer-based biosensors. TrAC Trends Anal Chem. 2008;27(2):108–17.
11. McKeague M, Derosa MC. Challenges and opportunities for small molecule aptamer development. J Nucleic Acids. 2012;2012:748913.
12. Wu X, Chen J, Wu M, Zhao JX. Aptamers: active targeting ligands for cancer diagnosis and therapy. Theranostics. 2015;5(4):322–44.
13. Pendergrast PS, Marsh HN, Grate D, Healy JM, Stanton M. Nucleic acid aptamers for target validation and therapeutic applications. J Biomol Tech. 2005;16(3):224–34.
14. Sullenger B, Woodruff R, Monroe DM. Potent anticoagulant aptamer directed against factor IXa blocks macromolecular substrate interaction. J Biol Chem. 2012;287(16):12779–86.
15. Floege J, Ostendorf T, Janssen U, Burg M, Radeke HH, Vargeese C, et al. Novel approach to specific growth factor inhibition in vivo: antagonism of platelet-derived growth factor in glomerulonephritis by aptamers. Am J Pathol. 1999;154(1):169–79.
16. Ostendorf T, Kunter U, Grone HJ, Bahlmann F, Kawachi H, Shimizu F, et al. Specific antagonism of PDGF prevents renal scarring in experimental glomerulonephritis. J Am Soc Nephrol. 2001;12(5):909–18.
17. Walsh TA. The emerging field of chemical genetics: potential applications for pesticide discovery. Pest Manag Sci. 2007;63(12):1165–71.
18. Stoltenburg R, Reinemann C, Strehlitz B. SELEX–a(r)evolutionary method to generate high-affinity nucleic acid ligands. Biomol Eng. 2007;24(4):381–403.
19. Dupont DM, Larsen N, Jensen JK, Andreasen PA, Kjems J. Characterisation of aptamer-target interactions by branched selection and high-throughput sequencing of SELEX pools. Nucleic Acids Res. 2015;43(21):e139.
20. Li BQ, Zhang YC, Huang GH, Cui WR, Zhang N, Cai YD. Prediction of aptamer-target interacting pairs with pseudo-amino acid composition. PLoS ONE. 2014;9(1):e86729.
21. Li T, Du P, Xu N. Identifying human kinase-specific protein phosphorylation sites by integrating heterogeneous information from various sources. PLoS ONE. 2010;5(11):e15411.
22. Wang M, Zhao XM, Tan H, Akutsu T, Whisstock JC, Song J. Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets. Bioinformatics. 2014;30(1):71–80.
23. Li F, Li C, Wang M, Webb GI, Zhang Y, Whisstock JC, Song J. GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. Bioinformatics. 2015;31(9):1411–9.
24. Zhang YN, Yu DJ, Li SS, Fan YX, Huang Y, Shen HB. Predicting protein-ATP binding sites from primary sequence through fusing bi-profile sampling of multi-view features. BMC Bioinformatics. 2012;13:118.
25. Hayat M, Tahir M, Khan SA. Prediction of protein structure classes using hybrid space of multi-profile Bayes and bi-gram probability feature spaces. J Theor Biol. 2014;346:8–15.
26. Li L, Zhang Y, Zou L, Li C, Yu B, Zheng X. An ensemble classifier for eukaryotic protein subcellular location prediction using gene ontology categories and amino acid hydrophobicity. PLoS ONE. 2012;7(1):e31057.
27. Xie HL, Fu L, Nie XD. Using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC. Protein Eng Des Sel. 2013;26(11):735–42.
28. Chen MC, Chen LS, Hsu CC, Zeng WR. An information granulation based data mining approach for classifying imbalanced data. Inf Sci. 2008;178(16):3214C–27.
29. Sanders WS, Johnston CI, Bridges SM, Burgess SC, Willeford KO. Prediction of cell penetrating peptides by support vector machines. PLoS Comput Biol. 2011;7(7):e1002101.
30. Gautam A, Chaudhary K, Kumar R, Sharma A, Kapoor P, Tyagi A, et al. In silico approaches for designing highly effective cell penetrating peptides. J Transl Med. 2013;11:74.
31. Cruz-Toledo J, McKeague M, Zhang X, Giamberardino A, McConnell E, Francis T, et al. Aptamer Base: a collaborative knowledge base to describe aptamers and SELEX experiments. Database (Oxford). 2012;2012:bas006.
32. 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. 2014;46(4):977–93.
33. Chen W, Lei TY, Jin DC, Lin H, Chou KC. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal Biochem. 2014;456:53–60.
34. Li L, Yu S, Xiao W, Li Y, Huang L, Zheng X, et al. Sequence-based identification of recombination spots using pseudo nucleic acid representation and recursive feature extraction by linear kernel SVM. BMC Bioinforma. 2014;15:340.
35. Qiu WR, Xiao X, Chou KC. iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int J Mol Sci. 2014;15(2):1746–66.
36. Zhou X, Li Z, Dai Z, Zou X. Predicting promoters by pseudo-trinucleotide compositions based on discrete wavelets transform. J Theor Biol. 2013;319:1–7.
37. Guo SH, Deng EZ, Xu LQ, Ding H, Lin H, Chen W, et al. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics. 2014;30(11):1522–9.
38. Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH. Local DNA topography correlates with functional noncoding regions of the human genome. Science. 2009;324(5925):389–92.
39. Panda B, Mishra AP, Majhi B, Rout M. Prediction of protein structural class by functional link artificial neural network using hybrid feature extraction method. In: Swarm, Evolutionary, and Memetic Computing. Cham, Switzerland: Springer International Publishing AG; 2013.
40. Sahu SS, Panda G. A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Comput Biol Chem. 2010;34(5-6):320–7.
41. Ahmed N, Natarajan T, Rao KR. Discrete cosine transform. IEEE Trans Comput. 1974;C-23:90–3.
42. Sarhan AM. Iris recognition using the discrete cosine transform and artificial neural networks. J Comput Sci. 2009;5(5):369–73.
43. Chou KC. Structural bioinformatics and its impact to biomedical science. Curr Med Chem. 2004;11(16):2105–34.
44. Li BQ, Hu LL, Chen L, Feng KY, Cai YD, Chou KC. Prediction of protein domain with mRMR feature selection and analysis. PLoS ONE. 2012;7(6):e39308.
45. Niu S, Hu LL, Zheng LL, Huang T, Feng KY, Cai YD, et al. Predicting protein oxidation sites with feature selection and analysis approach. J Biomol Struct Dyn. 2012;29(6):650–8.
46. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
47. Zhang J, Zhao X, Sun P, Ma Z. PSNO: predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC. Int J Mol Sci. 2014;15(7):11204–19.

48. Xu R, Zhou J, Wang H, He Y, Wang X, Liu B. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. BMC Syst Biol. 2015;9 Suppl 1:S10.
49. Sharma A, Lyons J, Dehzangi A, Paliwal KK. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. J Theor Biol. 2013;320:41–6.
50. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry. 2002;41(21):6573–82.
51. Noivirt-Brik O, Prilusky J, Sussman JL. Assessment of disorder predictions in CASP8. Proteins. 2009;77 Suppl 9:210–6.
52. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinforma. 2006;7:208.
53. Yu L, Guo Y, Zhang Z, Li Y, Li M, Li G, et al. SecretP: a new method for predicting mammalian secreted proteins. Peptides. 2010;31(4):574–8.
54. Qian J, Miao DQ, Zhang ZH, Li W. Hybrid approaches to attribute reduction based on indiscernibility and discernibility relation. Int J Approx Reason. 2011;52(2):212–30.
55. Lin H, Ding H, Guo FB, Huang J. Prediction of subcellular location of mycobacterial protein using feature selection techniques. Mol Divers. 2010;14(4):667–71.
56. Kira K, Rendell LA. The feature selection problem: Traditional methods and a new algorithm. In: Proceedings of the Tenth National Conference on Artificial Intelligence. San Jose, CA, United States: AAAI Press; 1992. p. 129–134.
57. Sun Y. Iterative RELIEF for feature weighting: algorithms, theories, and applications. IEEE Trans Pattern Anal Mach Intell. 2007;29(6):1035–51.
58. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics. 2004;20(15):2479–81.
59. Yang R, Zhang C, Gao R, Zhang L. An ensemble method with hybrid features to identify extracellular matrix proteins. PLoS ONE. 2015;10(2): e0117804.
60. Provost F. Machine learning from imbalanced data sets 101. Soft Computing & Pattern Recognition. International Conference of. IEEE, in New York, NY, United States. 2015;435–439.
61. Xu L, Chow MY. A classification approach for power distribution systems fault cause identification. IEEE Trans Power Syst. 2006;21(1):53–60.
62. Zhou ZH, Liu LY. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans Knowl Data Eng. 2006;18(1):66–77.
63. Li S, Li H, Li M, Shyr Y, Xie L, Li Y. Improved prediction of lysine acetylation by support vector machines. Protein Pept Lett. 2009;16(8):977–83.
64. Li ZC, Zhou X, Dai Z, Zou XY. Identification of protein methylation sites by coupling improved ant colony optimization algorithm and support vector machine. Anal Chim Acta. 2011;703(2):163–71.
65. Lo SL, Chiong R, Cornforth D. Using support vector machine ensembles for target audience classification on Twitter. PLoS ONE. 2015;10(4): e0122855.
66. Hansen LK, Salamon P. Neural network ensembles. IEEE Trans Pattern Anal Mach Intell. 1990;12(10):993–1001.
67. Chou KC, Zhang CT. Prediction of protein structural classes. Crit Rev Biochem Mol Biol. 1995;30(4):275–349.
68. Chou KC, Shen HB. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat Protoc. 2008;3(2):153–62.
69. Chou KC, Shen HB. Recent progress in protein subcellular location prediction. Anal Biochem. 2007;370(1):1–16.
70. Lin H, Deng EZ, Ding H, Chen W, Chou KC. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 2014;42(21):12961–72.
71. Ding H, Li D. Identification of mitochondrial proteins of malaria parasite using analysis of variance. Amino Acids. 2015;47(2):329–33.
72. Youden YW. Index for rating diagnostic tests. Cancer. 1950;3(1):32–35.
73. Sukanta M, Priyadarshini PP. Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. Journal of Theoretical Biology. 2014;356:30–35.
74. Lin H, Chen W, Ding H. AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes. PLoS ONE. 2013;8(10):e75726.
75. Parham C, Chirica M, Timans J, Vaisberg E, Travis M, Cheung J, et al. A receptor for the heterodimeric cytokine IL-23 is composed of IL-12Rbeta1 and a novel cytokine receptor subunit, IL-23R. J Immunol. 2002;168(11): 5699–708.
76. Szabo R, Netzel-Arnett S, Hobson JP, Antalis TM, Bugge TH. Matriptase-3 is a novel phylogenetically preserved membrane-anchored serine protease with broad serpin reactivity. Biochem J. 2005;390(Pt 1):231–42.
77. Kohler HP, Grant PJ. Plasminogen-activator inhibitor type 1 and coronary artery disease. N Engl J Med. 2002;342(24):1792–801.