Artificial Intelligence In Medicine 98 (2019) 35–47

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Accurate prediction of potential druggable proteins based on genetic algorithm and Bagging-SVM ensemble classifier

Jianying Lin[a,b,c], Hui Chen[a,b,c], Shan Li[a,b], Yushuang Liu[a,b], Xuan Li[c], Bin Yu[a,b,d,e,*]

[a] College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China
[b] Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China
[c] Key Laboratory of Synthetic Biology, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China
[d] School of Mathematics and Statistics, Changsha University of Science and Technology, Changsha 410114, China
[e] School of Life Sciences, University of Science and Technology of China, Hefei 230027, China

ARTICLE INFO

Keywords:
Druggable proteins
Feature extraction
Genetic algorithm
Support vector machine
Bagging
Ensemble classifier

ABSTRACT

Discovering and accurately locating drug targets is of great significance for the research and development of new drugs. As a different approach to traditional drug development, the machine learning algorithm is used to predict the drug target by mining the data. Because of its advantages of short time and low cost, it has received more and more attention in recent years. In this paper, we propose a novel method for predicting druggable proteins. Firstly, the features of the protein sequence are extracted by combining Chou's pseudo amino acid composition (PseAAC), dipeptide composition (DPC) and reduced sequence (RS), getting the 591 dimension of drug target dataset. Then, the feature information of druggable proteins dataset is selected by genetic algorithm (GA). Finally, we use Bagging ensemble learning to improve SVM classifier to get the final prediction model. The predictive accuracy rate reaches 93.78% by using 5-fold cross-validation and compared with other state-of-the-art predictive methods. The results indicate that the method proposed in this paper has a high reference value for the prediction of potential drug targets, which will successfully play a key role in the drug research and development. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/GA-Bagging-SVM.

1. Introduction

Drug target refers to the role of drug binding sites in the body, including gene loci [1], receptors [2,3], enzymes, ion channels, nucleic acids and other biological macromolecules [4]. Studies have shown that druggable proteins are closely related to the immune system, cardiovascular, hypertension and other diseases [5]. The key to the successful development of new drugs is the discovery and accurate positioning of drug targets [6]. Therefore, selecting new and effective drug targets is the top priority for the research and development of new drugs [7]. However, the traditional methods require analyzing the three-dimensional structure of the protein, resulting in a long development cycle and high cost [8]. So the traditional methods have limited effect. Since the successful implementation of the Human Genome Project [9], protein sequence data has grown exponentially. We can use machine learning algorithms to study drug targets [10], integrate the genome and mining protein sequences information deeply, and provide theoretical support for drug development to make up for the deficiency of traditional computing methods [11,12]. This is great importance for the development of highly selective drugs [13].

The key to studying drug targets is to find the right druggable proteins and to extract the protein feature sequences [14]. At present, the most popular feature extraction algorithms for protein sequences [15,16] are (1) Feature extraction algorithm based on amino acid composition and position, such as amino acid composition and n-order coupling composition, etc. [17–19]. (2) Feature extraction algorithm based on physicochemical properties of amino acids [20]. Typical algorithms include pseudo amino acid composition [21], Zp curves and hydrophobic patterns [22]. (3) Feature extraction algorithm based on database information [23,24]. Typical methods include functional domain composition (FunD) and gene ontology (GO) feature extraction algorithm [25]. At present stage, using machine learning algorithms to predict druggable proteins faces opportunities and challenges [26,27]. How to design a prediction algorithm with high accuracy has been the

* Corresponding author at: College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China.
E-mail addresses: ljy6366399@126.com (J. Lin), 1094685847@qq.com (H. Chen), lishan5600@163.com (S. Li), qustlys@126.com (Y. Liu), lixuan@sibs.ac.cn (X. Li), yubin@qust.edu.cn (B. Yu).

https://doi.org/10.1016/j.artmed.2019.07.005
Received 23 June 2018; Received in revised form 3 March 2019; Accepted 18 July 2019

key issue of drug target prediction research. In recent years, commonly used methods such as statistical and machine learning were random forest [28], K-nearest neighbor [29], neural network [30], Bayesian networks [31,32], hidden Markov model [33], and support vector machine [34,35]. At the same time, the accuracy can effectively improve through ensemble learning algorithms [36,37], and the generalization ability is also enhanced [17,38–40].

So far, academic circles have made great progress in the prediction of drug target. Yu et al. [41] used the PROFEAT software (http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi) to process 1080 feature vectors. Then they used random forest (RF) and support vector machine (SVM) method for drug target prediction. By using a 5-fold cross-validation test, a sensitivity of 81.33% and a specificity of 93.62% were obtained. Based on the feature of the supervised learning, Chen et al. [42] combined the three basic features of the protein's basic sequence, secondary structure, and subcellular localization, and used support vector machine to analyze and predict the potential drug targets in ion channel. Through the 10-fold cross-validation method, the accuracy of the two datasets were 85.13% and 76.44% respectively, and the potential druggable proteins were successfully discovered from ion channel proteins. Li and La [43] and Han et al. [44] obtained the overall prediction accuracy of 84% by using support vector machine method in the human drug targets dataset through the 10-fold cross-validation method. And it predicted the human potential drug targets from Swiss-Prot successfully. At the same time, this method has been widely used for the prediction of the pharmaceutical protein [45], antioncogene [46], high concentration target protein families [47,48] and various functional and structural classes of proteins [49]. Jamali et al. [50] combined with the physicochemical properties of protein sequences, the amino acid and dipeptide composition to extract feature. Support vector machines were used as feature selection methods and the neural network was used as a prediction classifier. The overall prediction accuracy achieved 92.1% by the 5-fold cross-validation test. Based on similarity of the supervised learning, Yamanishi et al. [51] selected and extracted drug structural similarity, protein sequences similarity, and druggable proteins interaction networks, and used nuclear regression methods to predict the drug target through chemical and genomic space integration. Finally, they proved the practicability of the proposed method for predicting drug-target interaction networks. Then Bleakley and Yamanishi [52] used the same dataset, and proposed the bipartite local models (BLM) learning methods to improve and optimize the algorithm. The prediction accuracy rate was further improved, which showed great potential for development.

However, there are still many problems to be solved for the prediction of druggable proteins [53]. Firstly, the higher the prediction accuracy of druggable proteins is, the stronger the ability to predict drug proteins. But with the age of big data, massive amounts of protein sequences have increased the computational complexity. How to further improve the prediction accuracy of druggable proteins becomes the key to researching drug target. Secondly, because the druggable proteins only account for a small part of the proteins, the numbers of positive and negative samples are imbalanced. This problem will interfere with the decision tree, SVM and other machine learning algorithms, impacting the accuracy of prediction. How to build reasonable datasets becomes a breakthrough in improving accuracy. Finally, the prediction of drug targets based on machine learning is the result of the two-class of drug proteins. It is impossible to predict the specific types of drug targets. Therefore, for the weaknesses of existing methods in the prediction of druggable proteins, this article uses machine learning algorithms to improve accuracy and solve problems [54–57].

In this paper, according to the effects of druggable proteins in different machine learning algorithms, we propose a drug target prediction method based on genetic algorithm and Bagging-SVM ensemble classifier. First of all, the features are extracted from protein sequences by combining Chou's PseAAC, dipeptide composition and reduced sequence algorithms. We compare different information fusion methods and build the original datasets. Second, the features vectors of the original dataset are selected by the genetic algorithm, which makes the features of proteins more prominent. Then we input the optimal feature vectors after feature selection to the SVM classifier for prediction. Finally, we use the self-sampling method of Bagging algorithm to ensemble learning the multiple SVM classifiers. By using a majority ballot way with the learning results, we choose the most suitable feature extraction algorithm, the best predictive classifiers and kernel functions for the prediction model. And the druggable proteins dataset obtain the highest overall prediction accuracy of 93.78% by using the 5-fold cross-validation method. The experimental results show that the algorithms we proposed improve the accuracy of druggable proteins prediction. The potential drug targets tested by this model can provide references for discovering new drug targets.

2. Materials and methods

2.1. Dataset

Drug target datasets can be downloaded from the DrugBank database (https://www.drugbank.ca/) [58–62]. For research, we need to divide the datasets into the training and testing data. Both of them should contain druggable proteins (positive samples) and non-druggable proteins (negative samples). However, there are highly similar protein sequences in DrugBank database, which generate noise and data redundancy during training. We use the dataset proposed by Jamali et al. [50], which was originally constructed by combining the dataset proposed by Li et al. [43] and Bakheet et al. [62]. And 1224 druggable protein sequences were selected to comprise the positive samples set and 1319 protein sequences were considered as nondrug targets representing the negative samples. The datasets file exists at: https://github.com/QUST-AIBBDRC/GA-Bagging-SVM.

2.2. Methods

2.2.1. Pseudo-amino acid composition

In order to avoid losing sequence information of proteins, Chou [63] proposed pseudo-amino acid composition (PseAAC) approach in 2001. In this study, the pseudo-amino acid composition is the primary structure of the protein encoded method. It is a vector of $(20 + \lambda)$ dimensions, where the first 20 dimension of the feature vector is the amino acid composition and the posterior λ dimension could be represented by

$$\tau_j = \frac{1}{L-j}\sum_{i=1}^{L-j}\Theta(R_i, R_{i+j})(j < L) \tag{1}$$

$$\Theta(R_i, R_{i+j}) = \frac{1}{3}\{[H_1(R_{i+j}) - H_1(R_i)]^2 + [H_2(R_{i+j}) - H_2(R_i)]^2 + [M(R_{i+j}) - M(R_i)]^2\} \tag{2}$$

where L represents the length of the proteins, $H_1(R_i)$, $H_2(R_i)$, $M(R_i)$ respectively represent the hydrophobicity value, hydrophilicity value and side chain mass of the amino acid R. It can get a $(20 + \lambda)$ dimensional unit after normalizing processing.

$$P = [p_1, p_2, p_3, \ldots, p_{20}, p_{20+1}, \ldots, p_{20+\lambda}]^T \tag{3}$$

$$p_\mu = \begin{cases} \dfrac{f_\mu}{\sum_{\mu=1}^{20} f_\mu + \omega \sum_{j=1}^{\lambda} \tau_j}, & 1 \le \mu \le 20 \\ \dfrac{\omega \tau_{\mu-20}}{\sum_{\mu=1}^{20} f_\mu + \omega \sum_{j=1}^{\lambda} \tau_j}, & 20{+}1 \le \mu \le 20{+}\lambda \end{cases} \quad (4)$$

where f_μ is the occurrence frequency of 20 amino acids in the proteins sequence, ω is the weight factor, which was set at 0.05 [63]. τ_k is the k -tier sequence correlation factor, which reflects the sequence-order information and it can also calculate a sequence correlation function. From the above formula, the first 20 dimension of the feature vector P is the amino acid composition and the posterior λ dimension is the sequence correlation factor reflecting the different levels of amino acid sequence information.

It is worth noting that the maximum value of parameters λ (Lambda) must be less than the length of the shortest sequence in all datasets. In this paper, the length of the shortest protein sequence is 8 amino acids, therefore the value of λ in Eqs. (3) and (4) ranges from 1 to 7.

In this study, we use PseAAC online server (http://www.csbio.sjtu.edu.cn) developed by Shen and Chou [64] to extract the feature of protein sequences. The optimal PseAAC parameters λ can be determined from the accuracy of druggable proteins prediction.

2.2.2. Dipeptide composition

Dipeptide composition, also known as residue pair composition, is a $2-gram$ method, which can extract and calculate the probability of occurrence of two consecutive amino acid residues from the sequence string [65]. This sequence encoding method extracts various pattern features consisting of n -th contiguous amino acid residues from the protein sequences in the form of a sliding window, and counts the number of these occurrences in the sequences. Compared with the amino acid composition, the dipeptide composition takes into account the coupling effect between adjacent residues. That is to say, the dipeptide composition contains not only the composition information of amino acids, but also the sequence information of some amino acids. Therefore, the dipeptide composition is considered to be a better feature extraction method, and usually defines as a 400 dimension feature vector:

$$\bar{q} = [q_1, q_2, ..., q_{400}]^T \quad (5)$$

where $q_i (i = 1, 2, ..., 400)$ is the probability of residue pairs i, defined as

$$q_i = \frac{m_i}{M}, \; i = 1, 2, \; ..., 400 \quad (6)$$

where m_i is the number of residue pair i and M is the number of all possible residue pairs.

In this study, we use the DPC algorithm in the PROFEAT online service system [66] (http://bidd2.nus.edu.sg/cgi-bin/profeat2016/main.cgi) to extract feature for protein sequences. Each protein sequence in the dataset generates a 400 dimension feature vector.

2.2.3. Reduced sequence and index-vectors

In order to achieve the desired accuracy of druggable proteins prediction, in this paper, we refer to the protein sequence feature extraction algorithm proposed by Xu et al. [67] and propose a method for proteins sequence that reduced sequence and index-vectors. In this section, we generate five types of reduced protein sequences based on their physicochemical properties, including polarity, acidity, charge, secondary structure and DHP [68]. Because of the difference in physicochemical properties of the 20 amino acid side chains, the type of amino acids is determined. Different combinations of protein sequences have different structures to adapt to various environments and complete their specific physiological functions. Therefore, the classifications of reduced protein sequences according to five properties are shown in Table 1 [67,69].

Table 1
Amino acid classification.

Property	Classifications
Polarity/acidity	DE RHK WYF SCMNQT GAVLIP
Acidity	DE KHR ACFGILMNPQSTVWY
Secondary structure	EHALMQKR VTIYCWF GDNPS
Charge	KR AVNCQGHILMFPSTWY DE
DHP	PALVIFWM QSTYCNG HKR DE

Note: DHP is the detailed HP model [70].

As can be seen from Table 1, the polarity & acidity of protein sequences may serve as an example. 20 amino acids can be classified into 5 groups [67]: acidic amino acids (A): D, E; basic amino acids (K): R, H, K; aromatic amino acids (R): W, Y, F; polar neutral amino acids (P): S, C, M, N, Q, T; non-polar amino acid (N): G, A, V, L, I, P.

For example, **Protein I** = {MATRTQARGAVVELLYAFESGNEEIKKIASSMLEE}, the above 20 amino acids based on the polarity & acidity of protein sequences can be represented as:

Reduced sequence I = {PNPKPPNKNNNNANNRNAPNPAANKKNNPPPNAA}.

Considering a reduced protein sequence, we count the occurrences of characters and di-characters by $|R_i|$ and $|R_iR_j|$ in this paper (where R_i and R_j represent i -th and j -th kind of character of the reduced sequence, respectively). We also use amino acid's frequency $|a_{ki}|$ to represent the k -th amino acid of i -th character.

Take protein I as an example. The occurrence of character 'A' in its corresponding reduced sequence is 6; 'K' is 4; 'R' is 2; 'P' is 9; and 'N' is 14. The di-character 'AA' in its corresponding reduced sequence is 2; 'AK' is 0; …; 'AN' is 2; … and so on.

In this paper, we propose two index-vectors V_1 and V_2. The vector V_1 is defined as

$$V_1 = \left(\frac{|a_{k1}|}{|R_1|}, \frac{|a_{k2}|}{|R_2|}, ..., \frac{|a_{km}|}{|R_m|}\right) \quad (7)$$

Take protein I and the polarity & acidity of protein sequences as an example. According to formula (7), the index-vectors V_{1A} of character 'A' in its corresponding reduced sequence is $V_{1A} = (0, \frac{1}{6})$; the index-vectors V_{1K} of character 'K' is $V_{1K} = (0, \frac{1}{6})$; the index-vectors V_{1R} of character 'R' is $V_{1R} = (\frac{1}{2}, 0, \frac{1}{2})$; the index-vectors V_{1P} of character 'P' is $V_{1P} = (\frac{1}{3}, 0, \frac{2}{9}, \frac{1}{9}, \frac{1}{9}, \frac{2}{9})$ and the index-vectors V_{1N} of character 'N' is $V_{1N} = (\frac{1}{7}, \frac{5}{14}, \frac{1}{7}, \frac{3}{14}, \frac{1}{7}, 0)$.

The other vector V_2 is defined as

$$V_2 = \left(\frac{|R_1R_j|}{|R_1|}, \frac{|R_2R_j|}{|R_2|}, ..., \frac{|R_mR_j|}{|R_m|}\right) \quad (8)$$

Where m represents the number of amino acids groups of the reduced sequence; $j = 1, 2, ..., m$. The overall dimension of V_1 is $20 \times 5 = 100$. And the overall dimension of V_2 equals to $5^2 + 3^3 + 3^2 + 3^2 + 4^2 = 68$.

Therefore, using the RS feature extraction algorithm to reduce sequence and index-vectors for protein sequences. Each protein sequence in the dataset generates a 168 dimension feature vector. This feature extraction algorithm can convert the different lengths protein sequences in the dataset into dimension-unified index feature vectors. It is convenient to implement the algorithm below.

2.2.4. Genetic algorithm

Genetic algorithm (GA) is a parallel stochastic search optimization method developed by John Holland [71] in 1975 to simulate the natural genetic mechanism and biological evolution theory. Its main idea is to generate initial populations randomly, and to continuously update the population through genetic operators, such as selection, crossover, and mutation, finally obtain a better solution [72]. The feature selection method of the genetic algorithm is: Firstly, code the parameters of feature and give a population of initial size determination. Each individual in the population corresponds to a possible solution. Then calculate the fitness value of each individual according to the fitness value function. It sets the probability of crossover and mutation of the population. The next generation is formed by using appropriate genetic strategies such as crossover rate, mutation rate and selection rate until the population performance meets a certain index or has completed a predetermined number of iterations [73]. The basic algorithm steps are as follows:

Step1: Each feature is defined as a gene and all features are equivalent to a chromosome, which length is the same to the features. Each chromosome represents a different subset of features. If the k -th bit of the chromosome is 1, it means that the k -th feature is selected. If it is 0, it means not selected.

Step2: Set the maximum evolution algebra T and create initial population that included M –th individual. Each individual is showed as $X_k^1, X_k^2, ..., X_k^M$, $k = 0$. Randomly generate the number of "1" contained in each chromosome, and then assign these "1" to the corresponding chromosomes randomly to obtain chromosome populations that represent different numbers of features.

Step3: Test the evaluation value of fitness, the purpose of feature selection is to use less features to achieve the same or better classification effect. In this paper, the fitness is evaluated from two aspects: the accuracy of the classification and the number of feature subset input by classifier. Therefore, the fitness function for each individual is:

$$F(X_k^i) = \begin{cases} R_{RA} \\ F_{Fe(\min)} \end{cases}, i = 1, ..., M \tag{9}$$

where R_{RA} represents the recognition accuracy and $F_{Fe(\min)}$ represents the number of feature. If the accuracy of the two feature subsets is the same, then the subset with fewer features is selected. When training classifiers, we only use the features identified in the feature subset and use the classification results to evaluate the performance of the classifier. This is used to guide the further search for the genetic algorithm.

Step4: Use roulette to select operators, it means selecting the chromosomes in each generation of population based on fitness ratio selection strategies. The probability of selection is

$$p_i = F_i / \sum_{i=1}^{M} F_i, \quad i = 1, 2, ..., M \tag{10}$$

where F_i is the reciprocal of the fitness value, and M is the population size.

Step5: Use single-point cross method and select two individuals with the same probability from $X_k'^1, X_k'^2, ..., X_k'^M$. The two individuals with pre-probability P_c get on crossover operations, which create two new individuals. Repeat this process, until a new group $X_k^{''1}, X_k^{''2}, ..., X_k^{''M}$ is formed.

Step6: According to certain mutation probability P_m, change the value of each individual randomly and create a new generation of groups $X_{k+1}^1, X_{k+1}^2, ..., X_{k+1}^M$.

Step7: Verify whether the termination condition is satisfy, if the condition is satisfy, the operation stops and the optimal solution with maximum fitness is output during evolution. Otherwise, let $X_k^i = X_{k+1}^i$, $i = 1, ..., M$ and go to Step2.

2.2.5. Support vector machine

Support vector machine (SVM) is a machine learning method based on statistical learning theory proposed by Vapnik et al. [74]. In recent years, support vector machines have been widely used in many fields of bioinformatics [75–86], and have achieved better prediction results on drug target prediction [35]. Its concrete principle and solution process are as follows:

For a two-category problem, given training set $T = \{(x_i, y_i) | i = 1, 2, ..., n\}$, where $x_i \in X$ is n dimension feature vectors in the real number field $y_i \in \{-1, +1\}$, whose value is -1 or 1. If the training set is linearly separable, in the sample space, we use the linear equation $\omega^T x + b = 0$ to represent the hyperplane, where $\omega = (\omega 1; \omega 2; ...; \omega d)$ denotes normal vector which determines the direction of the hyperplane and constant b denotes the distance between the hyperplane and the origin. Therefore, the distance from any point x in the sample space to the hyperplane can be written as $\gamma = \frac{|\omega^T x + b|}{||\omega||}$.

The kernel function $K(x_i, x_j)$ is used to replace the dot product in the optimal classification hyperplane to optimize the classification performance of the SVM. Where α_i represents Lagrange multiplier, the classification threshold of the sample is $b^* = y_i - \sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x_j)$. Finally, the ultimate discriminate function is given by: $f(x) = \text{sgn}(\omega \cdot x + b) = sgn\{\sum_{i=1}^{n} y_i \alpha_i^* K(x_i, x_j) + b^*\}$.

The selection of the kernel function is very important for the support vector machine. On the same dataset, different kernel functions about SVM algorithms may produce different prediction effects exceedingly. Generally, the appropriate kernel function can improve the prediction accuracy of the model, such as linear kernel function, polynomial kernel function, radial basis function (RBF) and Sigmoid kernel function.

This experiment uses LIBSVM package developed by Chang and Lin [87], which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

2.2.6. Bagging ensemble learning

In the classification algorithm that fuses multiple sensor data, the single classifier in traditional support vector machine (SVM) cannot directly learn incremental sample on the sensor data streams. It needs to use the new sample to adjust the parameters to exist single classifier and internal structure. For large datasets, it consumes huge time and space resources. And when setting internal structure adjustment parameters, it is prone to experience adaptation problems easily. Therefore, to solve these problems, in order to improve the accuracy and diversity of the learner and enhance the generalization performance of the classifier, we propose an ensemble learning algorithm based on Bagging-SVM in this paper.

Ensemble Learning refers to using datasets to obtain multiple training subsets. According to each base classifier trained by subsets of data, it combines these base classifiers in some way to create a new ensemble classifier. SVM-based Bagging algorithm extracts training set by using self-sampling method in incremental data [88]. It constructs an ensemble classifier that can reflect the changes of new information, making the new sample set significantly different. Then we use multiple support vector machine (SVM) classifiers to learn each subsample set, and use the majority vote method for the learning results to implement ensemble incremental learning [89].

The specific solution process of Bagging-SVM algorithm is as follows:

Given an aggregate as S, and carry out the T -th round of self-help sampling for aggregate S. T -th subsets S_t are extracted from aggregate S ($t = 1, 2, \cdots, T$). And each subsets S_t contains N -th sample.

Given a base classifier, learn the new training sample set S_t by using the SVM algorithm. Then each subset S_t will produce a weak classifier $\varphi(x, S_k)$, calculating the error rate of weak classifiers $\varphi(x, S_k)$:

$$\varepsilon_t = \sum_{(x_i, y_i) \in S_t} [\varphi(x, S_k) \neq y_i]/|S_t| \tag{11}$$

Extract the training set S_{t+1} independently again according to some distribution, after circulating, T -th weak classifier is integrated into a strong classifier $\Phi(x, S)$. And the final decision function is obtained.

When entering the test sample, strong classifier $\Phi(x, S)$ will output voting results for t -th weak classifiers $\varphi(x, S_k)$. It means that the majority categories in the votes will be the test sample categories. The steps of Bagging-SVM algorithm are as follows:

Algorithm for the Bagging ensemble.

Input:	training set $S = \{(X_i, y_i)\}$, $i = 1, 2, ..., m$; leaning machine L; the number of base classifiers T;
Process:	
01:	for $t = 1, 2, ..., T$
02:	Extracting m -th samples randomly from S, designated as S_t;
03:	Using S_t to learn L: $N_t = L(S_t)$;
04:	Combining base classifiers using majority voting: $N^*(x) = \arg\max_{y \in Y} \sum_{t: N_t(x) = y} 1$;
05:	end for
Output:	ensemble N^*

According to comprehensive prediction result of each base classifier, the Bagging algorithm achieves higher performance than the traditional single classifier in predicting classification performance when training the sample set.

2.2.7. Prediction assessment

The classifier performance assessment indicators are crucial for evaluating the performance of each machine learning algorithm [50,86]. There is numerous assessment indicators in the field of protein structure and function prediction. Firstly, we define the following standard assessment:

True positive (TP) is the correct number of positive samples predicted; false positive (FP) is the number of positive samples that are wrongly predicted; true negative (TN) is the number of negative samples correctly predicted; False Negative (FN) is the wrong number of negative samples predicted. From the above four indicators, we can further derive the other six commonly used indicators to evaluate the performance of the classifier:

Accuracy (ACC): percentage of target and nontarget proteins correctly predicted, which reflects the ability of classifier to determine the entire sample—The drug target proteins can be judged as positive, and the nontarget proteins can be judged as negative.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{12}$$

Sensitivity (SN): percentage of drug target proteins that are predicted correctly:

$$SN = \frac{TP}{TP + FN} \times 100\% \tag{13}$$

Specificity (SP): percentage of nontarget proteins that are correctly predicted:

$$SP = \frac{TN}{TN + FP} \times 100\% \tag{14}$$

Area under the curve (AUC): This parameter is the area of receiver operating characteristic (ROC) curve, which intuitive meaning is the probability that the positive sample score is greater than the negative sample, when taking a positive sample and a negative sample randomly. It is a logical assessment for model performance. Its value ranges from 0 to 1, where 1 represents the optimal performance and 0 represents the worst performance. And when the sequence is randomly sorted, the value of AUC is 0.5.

Matthews correlation coefficient (MCC): MCC is used in machine learning as a measure of the quality of binary (two-class) classification. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. This parameter is association of feature classification in dataset. This parameter is association of feature classification in dataset. Its value ranges from -1 to 1, where -1 represents a completely opposite prediction and 1 represents a completely correct prediction. The formula is:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \tag{15}$$

PPV: Precision or positive predictive value

$$PPV = \frac{TP}{TP + FP} \times 100\% \tag{16}$$

NPV: Negative predictive value

$$NPV = \frac{TN}{TN + FN} \times 100\% \tag{17}$$

F-Measure (F1-score): This parameter is the harmonic mean weight of precision and recall. The formula is:

$$F - score = \frac{Precision \times recall}{precision + recall} \times 100\% \tag{18}$$

The general framework of feature extraction and classification prediction model for druggable proteins is shown in Fig. 1. We have implemented it in MATLAB R2014a in windows Server 2012R2 running on a PC with system configuration Intel (R) Xeon (TM) CPU E5-2650 @ 2.30 GHz 2.30 GHz with 32.0 GB of RAM.

The steps of druggable proteins dataset for feature extraction and classification prediction are described as follows:

(1) Use the dataset constructed by Jamali et al. [50]. Enter the protein sequence and the corresponding class label of the druggable proteins into the druggable proteins dataset;
(2) Use PseAAC network service system developed by Shen and Chou to carry out feature extraction on the protein sequences and generate $20 + \lambda$ dimension feature vector. Then use the DPC algorithm in PROFEAT online service system to generate 400 dimension feature vector. Finally, the reduce sequence algorithm is used to generate 168 dimension feature vector. Three methods combine to generate $20 + \lambda + 400 + 168$ dimension feature vectors as druggable proteins dataset;
(3) Genetic algorithm is used to select the druggable proteins dataset extracted in (2), and the effective information in the sequence is extracted;
(4) In order to predict the druggable proteins, the optimal feature vectors of feature selection are input to the SVM classifier. The Bagging algorithm is used to randomly resample the sample set, and Bagging-SVM ensemble incremental algorithm is used for positive and negative sample sets;
(5) According to the prediction accuracy of druggable proteins, select the optimal parameters of the model, including the λ values of pseudo amino acid composition, different parameter value in genetic algorithm, the optimal ways of the information fusion, different kernel functions in SVM and the optimal predictive classifier;
(6) According to the optimal parameters of the predicted model in (5), use 5-fold cross-validation test to calculate SE, SP, MCC, ACC, AUC, PPV, NPV, and F1-score. Compare the accuracy of the classification, we evaluate the prediction performance of the model.

3. Results and discussion

3.1. Selection of parameter λ of PseAAC

In this paper, because we use the PseAAC algorithm to extract the feature information of the protein sequences, it is inevitable to discuss the choice of λ value. In the PseAAC algorithm, the value of λ is used to

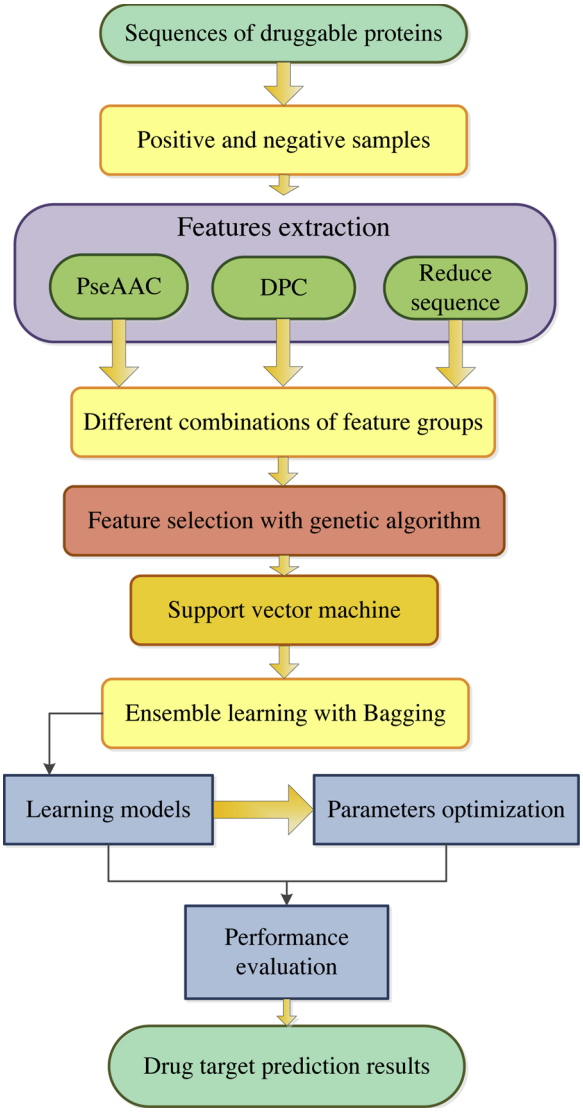

Fig. 1. Druggable proteins prediction model.

represent the amino acid sequences information of the protein sequences. If the value λ is set too large, it will result in dimension disaster and over-fitting phenomenon, thereby cause the classification recognition error. If the value λ is set too small, the features of the protein sequences on the datasets cannot be extracted completely, leading to an erroneous prediction.

Since the maximum value of the PseAAC parameter λ must be less than the length of the shortest sequence in all the protein datasets. The positive and negative samples are selected as the study object of the druggable protein datasets. The shortest length of the nontarget protein's sequence is 8, so select the value of parameter λ from 1 to 7 in turn. The SVM algorithm uses the RBF kernel function to classify the dataset, and the prediction results are tested by 5-fold cross-validation method. We obtain the prediction accuracy of each class of druggable proteins to find the optimal parameter value in feature extraction. The prediction results are shown in Table 2.

As can be seen from Table 2, different predictions results will be got by selecting the value of λ. Comparing and analyzing the obtained results, we can see that: When $\lambda = 3$ and the dimension feature vector is 23, the values of SP, MCC, PPV, F1-score reach the highest values, which are 93.81%、0.8118、92.41% and 89.64%, respectively. And $ACC = 90.61\%$, which obtained the highest prediction accuracy.

When $\lambda = 4$ and the dimension feature vector is 24, the values of SN, AUC, and NPV obtain the highest prediction accuracy of 89.68%, 0.9615 and 89.59% respectively. And the remaining values of λ don't exceed the above results.

In summary, the change of λ value affects the prediction accuracy of the druggable protein in the prediction model. As can be seen from Table 3, the optimal λ value is 3, so each protein sequence generates a $20 + \lambda = 20+3 = 23$ dimensional feature vector, when using the PseAAC algorithm to perform feature extraction on protein sequences.

Table 3
The three group of protein sequences based on feature extraction for drug target prediction.

Feature set	Description	Feature
Group 1	Pseudo amino acid composition of protein sequence	23
Group 2	Dipeptide composition of protein sequence	400
Group 3	Reduced sequence and index-vectors	168

3.2. Feature extraction

How to extract feature information from a protein sequence and use appropriate mathematical methods to describe and represent it is crucial for protein classification research. According to these two types of feature extraction algorithm, we consider the composition, position and physicochemical properties of protein sequences comprehensively in this study, and integrate three feature extraction algorithms for the prediction of druggable proteins. It is shown in Table 3.

As shown in Table 3, In order to achieve the desired accuracy of druggable proteins prediction, we combine three algorithms for feature extraction of protein sequences. Group 1 and Group 2 are feature extraction algorithms based on amino acid composition and location, and Group 3 is a feature extraction algorithm based on amino acid physicochemical properties.

Using the above three feature extraction algorithms, we extract 23, 400 and 168 dimension feature vectors respectively. By arranging and combining different multi-information fusion methods, the SVM algorithm of RBF kernel function is selected to classify and predict the

Table 2
Prediction results of druggable proteins by selecting different values of λ.

Number of λ	Feature	Evaluation index							
		SN (%)	SP (%)	MCC	ACC (%)	AUC	PPV (%)	NPV (%)	F1-score (%)
1	21	89.6	89.3	0.79	89.39	0.9531	86.05	88.94	88.75
2	22	87.35	92.22	0.7969	89.8	0.956	91.7	88.1	89.47
3	23	86.98	93.81	0.8118	90.61	0.9547	92.41	89.18	89.64
4	24	89.68	90.6	0.7913	89.59	0.9615	89.58	89.59	89.03
5	25	84.26	91.61	0.763	88.21	0.9522	89.59	87.15	86.84
6	26	87.6	89.55	0.7719	88.63	0.9526	88.33	88.89	87.97
7	27	87.12	89.49	0.7664	88.41	0.9571	86.05	88.97	89.16

Table 4
The results of different multi-information fusion methods for druggable proteins prediction.

Fusion methods	Feature	Evaluation index							
		SN (%)	SP (%)	MCC	ACC (%)	AUC	PPV (%)	NPV (%)	F1-score (%)
G1	23	86.98	93.81	0.8118	90.61	0.9547	92.41	89.18	89.64
G2	400	85.57	90.66	0.7641	88.31	0.9537	89.22	87.64	87.84
G3	168	89.49	90.47	0.7993	89.98	0.9481	89.88	90.14	89.69
G1 + G2	423	88.08	92.38	0.8045	90.16	0.9589	92.35	88.21	90.19
G1 + G3	191	87.21	89.76	0.7695	88.64	0.9583	87.17	89.79	87.19
G2 + G3	568	86.85	93.12	0.8029	90.18	0.9552	91.58	89.18	89.14
G1 + G2 + G3	591	87.88	94.11	0.8212	90.98	0.9634	93.77	88.54	90.97

dataset. Under the 5-fold cross-validation method, the results are shown in Table 4.

As can be seen from Table 4, we can see that different ways of information fusion will have different effects on the outcome of the prediction results. By comparing seven different methods of multi-information fusion, we find that the results of G1 and G1 + G2 + G3 are better, and the accuracy of prediction exceeds 90.5%. Therefore, we finally choose the information fusion method of G1 + G2 + G3 in this paper.

By analyzing and comparing the values of ACC, we can find that the 591 dimension feature vectors obtained by the three feature extraction algorithms PseAAC + DPC + RS, is 0.37% higher than when only extracted 23 dimension feature vectors by the PseAAC algorithm. The range of precision improvement is small. The possible reason is that the fusion of the datasets obtained by the three feature extraction algorithms, not only brings rich information to the drug target prediction, but also brings more redundancy, resulting in increasing the machine learning training time and decreasing the accuracy of prediction. Therefore, it is necessary to perform feature selection on the drug target dataset that fuses the three feature extraction algorithms, which is discussed in the following.

3.3. Feature selection

In machine learning, too many features will lead to dimension disaster and over-fitting phenomenon, which results in the error of classification recognition. In this paper, the 591 dimension feature vectors we extracted have no significant improvement over the 23 dimension feature vectors extracted by PseAAC. Therefore, it is necessary for us to extract the effective information of the protein sequences with feature selection algorithm.

Feature selection is one of the most important methods for data processing in the field of pattern recognition and data mining. From the optimization point of view, the feature selection problem is actually a combinatorial optimization problem. As a random search method, genetic algorithm has a wide range of applications in combinatorial optimization problems [90–92]. It usually evaluates feature subsets based on classifiers and gives individual assessment indicators and fitness levels based on classification accuracy, to effectively remove redundant information in protein sequences and extract the information for each type of protein. At the same time, in order to get the best performance of the model in data processing, the parameters of the genetic algorithm need to be set. In this paper, according to the empirical parameters, we set the crossover probability p_c and mutation probability p_m to 0.75 and 0.01 respectively. The feature selection algorithm has the same coding length l with feature dimension. Considering the efficiency of the algorithm, we set the population size $popsize = 50$ and evolutionary generations $gen = 3$ after the experimental determination.

In this paper, we assess the suitability of each chromosome (gene subset) based on the classification accuracy of the SVM algorithm. Through 5-fold cross-validation method to test the accuracy of the classification of gene subsets on the training set, we divide the dataset into 5 parts, taking 4 of them as training data and 1 part as test data. The average of the 5 test correct results is used to estimate the accuracy of the algorithm.

The SVM with RBF is used as the base classifier to find the optimal model value through the fitness function. The prediction results of the druggable proteins under different feature selection dimensions are obtained and shown in Table 5.

It can be seen from Table 5 that it can't get the best predictive effect when using all protein features. So it is a correct decision to remove the features with less contribution. Comparing the data in the Table 6, the SP and F1-score values of the 138 features, the AUC and NPV values of the 124 features, the SP, PPV and F1-score values of the 154 features are slightly higher than the 143 features. According to the most important assessment indicators MCC and ACC, however, 143 features have obvious advantage. In the comparison of MCC, the prediction value of the 143 features is 0.87% higher than the 154 features, 0.28% higher than the 138 features and 2.69% higher than the 154 features. In the comparison of ACC, the prediction accuracy of the 143 features are respectively 0.24%, 0.15% and 1.08% higher than the 154 features, the 138 features, and the 124 features. The optimal dimension of the feature selection by genetic algorithm is 143.

In addition, in order to verify the rationality of the genetic algorithm, we compare the protein prediction results of the optimal dimension under different algorithms and select the feature selection algorithm with the best performance. Among them, taking the traditional feature selection algorithm Relief and principal component analysis (PCA) as examples, the prediction result is compared with the GA feature selection algorithm and is shown in Table 6.

It can be seen from Table 6 that we select three feature selection algorithms GA, Relief, and PCA to reduce the dimension, based on the fusion of the 591 dimension feature vectors obtained by the three feature extraction algorithms. The selected best dimensions are 143, 236 and 59 respectively and the accuracy of the final prediction is different. For the feature-extracted druggable proteins dataset, when the feature selection method is not used, the SVM classifier is used for calculation directly, and the predicted accuracy of the druggable proteins is 90.98%, which is lower than the 93.42% of the genetic algorithm and the 91.57% of the Relief algorithm, but higher than the 90.59% of the PCA algorithm. It can be seen that not all of the feature selection methods can improve the prediction accuracy of the druggable proteins, and the genetic algorithm is the best feature selection method.

We use receiver operating characteristic (ROC) curves to compare the robustness of the prediction model under different feature extraction algorithms. As can be seen from Fig. 2, the coverage area of the ROC curve of the drug target dataset after feature selection by the genetic algorithm is the largest, and the area under curve (AUC) value is the largest. It shows that the genetic algorithm is better than other algorithms. In the practical application of the test system, the prediction results will have higher confidence.

As can be seen from Tables 5, 6 and Fig. 2, using the GA algorithm

Table 5
The prediction results of druggable protein assessed from the GA feature selection algorithm.

Number of feature	Evaluation index							
	SN (%)	SP (%)	MCC	ACC (%)	AUC	PPV (%)	NPV (%)	F1-score (%)
591	87.88	94.11	0.8212	90.98	0.9634	93.77	88.54	90.97
515	87.87	93.33	0.8149	90.77	0.9579	92.11	89.68	89.94
474	88.15	92.46	0.8157	90.78	0.9586	91.89	89.64	89.57
412	88.93	94.34	0.8353	91.75	0.9646	93.53	90.25	91.18
347	89.4	92.36	0.8219	91.16	0.9677	90.91	91.37	90.32
308	88.51	93.8	0.8262	91.36	0.9572	92.44	90.49	90.43
252	89.49	92.95	0.8255	91.28	0.9631	92.19	90.48	90.82
216	91.3	92.58	0.8389	91.94	0.9613	92.4	91.51	90.57
186	88.1	94.33	0.8369	91.75	0.9651	94.87	89.09	91.36
168	89.25	94.49	0.8417	92.16	0.962	93.42	91.15	91.52
154	91.55	95.24	0.8624	93.18	0.9742	95.22	91.26	93.24
143	**92.21**	**94.56**	**0.8711**	**93.42**	**0.9752**	**95.17**	**91.93**	**93.07**
138	91.16	95.4	0.8683	93.27	0.9724	94.88	91.87	93.23
124	91.82	92.73	0.8442	92.34	0.9774	90.58	93.71	91.2
102	89.26	92.88	0.8228	91.16	0.9559	91.91	90.51	90.57
83	88.11	92.83	0.8114	90.57	0.9596	91.88	89.45	89.96
78	87.33	92.36	0.7997	90.18	0.9502	91.34	89.21	89.41
62	86.95	93.42	0.7994	89.98	0.9471	92.8	87.55	89.66
46	85.77	92.11	0.7692	88.61	0.9575	92.41	85.96	87.17

Note: The bold values represent the optimal dimensions after feature selection by the genetic algorithm, which have the highest MCC and ACC values.

Table 6
The comparison of drug target prediction results based on different feature selection algorithms.

Algorithm	Number of feature	Evaluation index					
		SN (%)	SP (%)	MCC	ACC (%)	AUC	F1-score (%)
GA	143	**92.21**	**94.56**	**0.8711**	**93.42**	**0.9752**	**93.07**
Relief	236	91.2	91.92	0.8313	91.57	0.9664	91.38
PCA	59	87.64	93.63	0.8135	90.59	0.9636	90.44

Note: The bold values represent the optimal feature selection results.

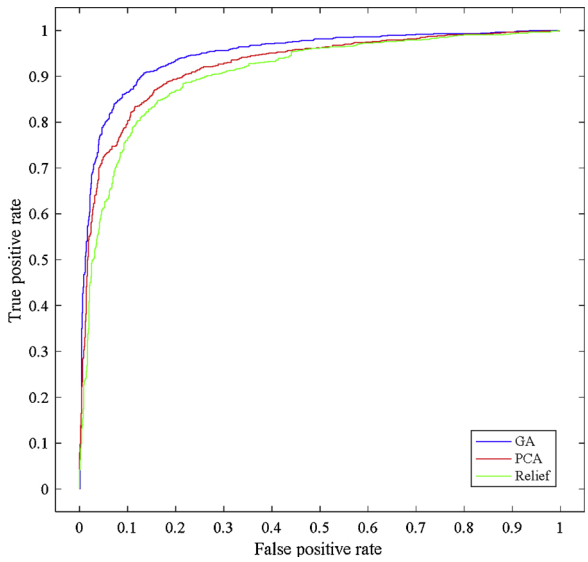

Fig. 2. The ROC curves for different feature selection algorithms.

to select the features of the dataset, the prediction accuracy is 2.44% higher than before. And the prediction accuracy is significantly higher than the other two methods. In practical applications, we use the genetic algorithm as a feature selection algorithm in this paper. It can effectively reduce the scope of the experiment, improve the efficiency of the experiment, and reduce the cost of the experiment.

3.4. Implemented machine-learning predictors

In this paper, we evaluate different machine learning algorithms and use five different learners: support vector machine (SVM), K-nearest neighbor (kNN), Naïve Bayes, Random Forest (RF), and Decision Tree (DT) to conduct drug target prediction experiments. By comparing the prediction results of drug targets, the optimal learner is selected as the machine learning algorithm. The basic assumptions of these machine learning algorithms are to minimize the empirical risk and to reduce errors in the training set.

In this study, SVM algorithm uses RBF, KNN algorithm using Euclidean distance, the number of neighbors is 3, the number of decision trees selected in RF is 200, Naïve Bayes algorithm and DT algorithm adopt default parameters. For the druggable proteins dataset, we compare the different classification algorithms with the 143 dimension feature vector under the 5-fold cross-validation test. As shown in Table 7.

As can be seen from Table 7, on the one hand, the prediction accuracy of each classifier is generally improved when the genetic algorithm is used for feature selection. On the other hand, the SVM classifier is used for the 143 dimension feature vector after using the genetic algorithm. The value of MCC, ACC, and AUC are 0.8711, 93.42% and 0.9752, respectively. The results of the assessment indicators are better than the other four classification algorithms. This further validates the excellent generalization performance of the support vector machine algorithm.

As can be seen from Fig. 3, the value of AUC is the largest in the support vector machine classification algorithms, which is 0.9752. It is not only higher than the value before the feature selection, which is 0.9634, but also significantly higher than the other four classification algorithms methods. The AUC values using Naïve Bayes, RF, KNN, and DT classifiers for the 143 dimension feature vector are 0.8874, 0.9434, 0.9218 and 0.8427, respectively. By comparing the effects of different classifiers on the prediction results and comparing the robustness of the five classifiers from the ROC curves, we select the SVM algorithm as a classifier for druggable proteins prediction.

3.5. Determination of SVM kernel function

The key to supporting vector machine classification algorithms is the selection of kernel functions. The appropriate kernel functions can improve the prediction accuracy of the classification model. In this

Table 7
Prediction results of the 143 dimension feature vector dataset under different classification algorithms.

Classifier	Evaluation index							
	SN (%)	SP (%)	MCC	ACC (%)	AUC	PPV (%)	NPV (%)	F1-score (%)
Naïve Bayes	90.66	66.27	0.5879	78.59	0.8874	73.27	87.43	81.04
RF	85.66	92.08	0.7803	89	0.9434	90.87	87.46	88.19
KNN	92.73	65.61	0.6097	80.35	0.9218	73.23	90.22	81.93
SVM	**92.21**	**94.56**	**0.8711**	**93.42**	**0.9752**	**95.17**	**91.93**	**93.07**
DT	82.4	79.92	0.6232	79.84	0.8427	82.47	81.1	81.14

Note: The bold values represent the optimal prediction results under different classification algorithms.

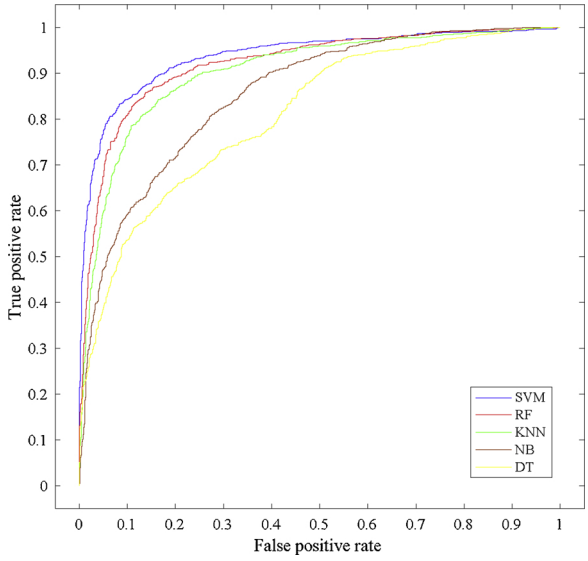

Fig. 3. The ROC curves of the 143 dimension feature vector under different classification algorithms.

paper, the positive and negative druggable protein datasets are selected as the study object. We use PseAAC, DPC and RS algorithms to extract the feature of protein sequences. When $\lambda = 3$ is chosen, the $20 + \lambda + 400 + 168 = 23 + 400 + 168 = 591$ dimension feature vector is generated. The SVM algorithm is used to classify and the 5-fold cross-validation method is used to test the results. We compare the classification results of the four kernel functions: linear kernel function, cubic polynomial kernel function, radial basis kernel function and Sigmoid kernel function, respectively, as shown in Table 8 showing the variation of the prediction accuracy of the druggable proteins dataset under different kernel functions.

As can be seen from Table 8, the prediction effect by RBF kernel function is better than that of the other three kernel functions when using druggable proteins dataset. And the prediction accuracy of the druggable proteins using the radial basis function reaches 90.98%. The MCC reaches 0.8212, F1-score reaches 90.97%, and the value of ACC is 3.55% higher than the Sigmoid kernel function, 6.11% higher than the linear kernel function, and 28.7% higher than the polynomial kernel function. And there will be no case where the individual index (SN, NPV) of the polynomial kernel function is extremely high, but the important indicators MCC, ACC are extremely low. In summary, we use radial basis function (RBF) as the kernel function of SVM in this paper.

3.6. Ensemble learning algorithm

In theory, the SVM is a strong classifier, which has good generalization ability, and is more stable than learning machines such as decision trees and random forests. However, SVM have obvious drawbacks in solving the problem of approximate solutions. In order to reduce the time and space complexity of solving the quadratic programming, the optimization problem needs to be decomposed into several subproblems. According to a cyclic iteration strategy, the subproblems are solved repeatedly, and the results converge to the optimal solution of the original problem finally. However, the solution obtained by approximation algorithm usually isn't the optimal solution, and the generalization ability of the support vector machine is reduced, which makes the ensemble learning based on the support vector machine has space for performance improvement. Therefore, it is necessary to improve the SVM prediction classifier.

Compared with the traditional classifier method, in order to achieve the effect of improving the classification prediction performance of the learning system, ensemble learning achieves the final decision results by integrating the prediction effects of each base classifier. For the drug target dataset studied in this paper, there are 1224 positive samples and 1319 negative samples. A portion of each sample set is taken out as a test set randomly, and the remaining portion is used as training set without incremental learning. 5-fold cross-validation method is used to randomly extract part of the dataset as a test set, and the remaining part as a training set. If incremental learning is performed, it is divided into a number of increments.

The experiments are conducted from three aspects. They are the performance comparison between the ensemble classifier and the single classifier, the incremental learning and the non-incremental learning, and the influence of the base classifier scale of ensemble learning on the accuracy. In the incremental process, the training set is divided into three parts randomly. We take one of them as the incremental training

Table 8
Prediction results of druggable proteins under different kernel functions.

Kernel function	Evaluation index							
	SN (%)	SP (%)	MCC	ACC (%)	AUC	PPV (%)	NPV (%)	F1-score (%)
liner kernel	86.02	83.88	0.6975	84.87	0.9257	82.19	87.4	84.06
polynomial kernel	98.78	28.41	0.3771	62.28	0.894	56.15	96.15	71.6
RBF	87.88	94.11	0.8212	90.98	0.9634	93.77	88.54	90.97
Sigmoid	87.44	87.41	0.746	87.43	0.947	84.42	89.93	85.9

Table 9
The prediction of drug target proteins with different ensemble learning algorithms.

Method	Evaluation index							
	SN (%)	SP (%)	MCC	ACC (%)	AUC	PPV (%)	NPV (%)	F1-score (%)
SVM	92.21	94.56	0.8711	93.42	0.9752	95.17	91.93	93.07
Bagging-SVM	92.86	94.45	0.8781	93.78	0.9787	94.22	93.73	93.58
Boosting-SVM	87.12	89.49	0.7664	88.41	0.9537	87.5	89.17	87.31

set each time and use ensemble classifier to test performance. Till then, a round of incremental learning is over, and then starts the next round of incremental learning until all incremental set training is completed. We select feature subsets that have the higher ROC value for ensemble learning, and to obtain better classification performance.

In order to compare the superiority-inferiority of the ensemble learning SVM algorithm, we use the single SVM classifier above to train all sample sets. On this basis, algorithm 1 uses Bagging algorithm to integrate the base classifiers of SVM (the number of classifiers is 10), which is denoted as Bagging-SVM. Algorithm 2 uses Boosting algorithm to integrate the base classifiers of SVM, which is denoted as Boosting-SVM. The prediction results are shown in Table 9.

It can be seen from Table 9 that on the one hand the accuracy of Bagging-SVM ensemble algorithm has been significantly improved, compared with a single SVM classifier. The accuracy of the prediction reaches 93.78%, AUC value reaches 0.9787. It effectively proved that the Bagging algorithm can be used to independently sample the sample set. The difference of the base classifier is improved and the entire ensemble classifier has better generalization ability. Therefore, the classifier prediction performance can be improved. On the other hand, the Boosting-SVM ensemble algorithm reduces the accuracy of drug target prediction by 5.01%. The probable reason is that the integration of the Boosting algorithm is different from the Bagging algorithm when integrating several weak classifiers into one strong classifier, resulting in a high degree of model bias, low variance, and strong correlation. It does not significantly reduce variance and can lead to a decrease in prediction accuracy.

The Bagging ensemble improves the prediction performance of the SVM classifier. Because the number of protein sequences is large and the hidden protein information is difficult to reveal, there is a high requirement for the performance of the prediction algorithm. In order to test the effect of the Bagging algorithm on the prediction performance of different classifiers after ensemble, we further use the other four kinds of classifiers as the base classifier to integrate Bagging ensemble in this paper. The prediction results of the druggable proteins are obtained and shown in Table 10.

As can be seen from Table 10, according to the ensemble learning of the Bagging algorithm under different classification algorithms, the prediction results of the druggable proteins have been improved gradually. Compared with the results before ensemble algorithm, the accuracy of the Naive Bayes algorithm is improved by 0.15%, the random forest algorithm is improved by 0.28%, the KNN algorithm is improved by 2.98%, the SVM algorithm is improved by 0.36%, and the decision tree algorithm is improved by 1.69%. By comparing the accuracy of the ensemble classifier, the Bagging-SVM algorithm is still the optimal drug target prediction classifier with the accuracy of 93.78%. Therefore, the Bagging-SVM algorithm is finally selected as the best algorithm for the prediction model in this paper.

3.7. Performance of predictive models and comparison with other methods

In this study, a drug target protein prediction model is established. We use a series of feature extraction and feature selection algorithms for the druggable datasets to calculate a prediction accuracy of 93.78% by 5-fold cross-validation method. In order to show the superiority of the prediction model, we compare the same dataset prediction accuracy with other methods.

At present, the machine learning algorithms have received more and more attention in the field of drug development. For the prediction of druggable proteins, bioinformatics experts have proposed multiple algorithms. For example, Jamali et al. [51] extracted the features of protein sequences and fusion with physicochemical properties, amino acid composition and dipeptide composition algorithms. It generates a 443 dimension feature vector. Under the 5-fold cross-validation method, the ACC value of the neural network classifier reaches 89.78%. After feature selection of protein sequences using support vector machines, the optimal dimension is 130 dimensions and the ACC value reaches 92.1%. Table 11 compares the prediction results between the Bagging-SVM method and other prediction methods on the same drug target protein dataset.

As can be seen from Table 11, different feature extraction and feature selection algorithms will result in different features for the same dataset. It can be seen that the effect of the feature extraction algorithm is compared to the premise of using the same drug target dataset. The PseAAC-DPC-RS feature selection algorithm obtains 90.98% of ACC and 0.8212 of MCC under the 5-fold cross-validation, which improve 1.2% of ACC, 2.6% of MCC and 0.0042 of the value of AUC than the method used by Jamali et al. [43]. Through GA feature selection and Bagging ensemble learning algorithm, the prediction accuracy of druggable proteins is steadily improved. After the feature selection with the

Table 10
Prediction results of the Bagging algorithm under different classification algorithms.

Classifier	Evaluation index							
	SN (%)	SP (%)	MCC	ACC (%)	AUC	PPV (%)	NPV (%)	F1-score (%)
Bagging-NB	90.83	66.67	0.5892	78.74	0.8851	74.12	87.43	82.16
Bagging-RF	85.68	90.97	0.7883	89.28	0.9464	90.93	87.62	89.69
Bagging-KNN	89.92	76.59	0.6718	83.33	0.9369	79.73	88.13	84.52
Bagging-SVM	92.86	94.45	0.8781	93.78	0.9787	94.22	93.73	93.58
Bagging-DT	90.55	70.94	0.6322	81.53	0.8772	78.55	86.46	84.12

Table 11
Performance comparison with different methods on the same druggable proteins dataset.

Algorithms	Number of feature	Evaluation index					
		SN (%)	SP (%)	MCC	ACC (%)	AUC	F-score (%)
PseAAC-DPC-RS	591	87.88	94.11	0.8212	90.98	0.9634	90.97
Jamali et al. [50]	443	90.14	89.38	0.7952	89.78	0.9592	90.14
PseAAC-DPC-RS-GA	143	92.21	94.56	0.8711	93.42	0.9752	93.07
Jamali et al. [50]	130	92.8	91.34	0.8417	92.1	0.9728	92.41
GA-Bagging-SVM	143	92.86	94.45	0.8781	93.78	0.9787	93.58

genetic algorithm, the ACC increases by 2.44%, the MCC increases by 4.99% and the AUC increases by 0.0118 under the SVM classifier. In addition, the prediction result of ACC improves 0.36%, MCC improves 0.7% and AUC improves 0.35% by using the Bagging algorithm. By comparison, the prediction accuracy of this model is 93.78%, which is 1.68% higher than the prediction method proposed by Jamali et al., the MCC value is 3.64% higher and the AUC value is 0.59% higher than it.

In summary, the above results fully show that the good feature extraction methods and the effective machine learning algorithms can significantly improve the prediction accuracy of the druggable proteins, and the prediction results of the model constructed in this paper are satisfactory. In addition, the high prediction accuracy of druggable proteins can reduce the cost of drug design in the prediction of human drug targets, and increase the speed of drug targets from discovery to clinical research, and can also be applied to gene expression and signal transduction. Therefore, the research in this paper has a certain application value for tapping potential drug targets.

4. Conclusion

The key to the successful development of new drugs lies in the discovery and accurate positioning of drug targets. With the advent of the post-genomic period, massive amounts of protein sequences have been growing exponentially into the database. It is crucial for the design and application of drugs using machine learning algorithms to analyze drug target protein data deeply and excavate in order to identify potential drug targets. In this paper, we propose a new druggable proteins prediction method. Firstly, we combine Chou's PseAAC, DPC and reduced sequence algorithms to carry out feature extraction on protein sequences. And it gets more detailed protein sequence and evolutionary information. In addition, we use genetic algorithm feature selection to remove redundant information from the protein sequences in the dataset. It screens the significant druggable proteins and is of great significance for ensuring a high accuracy of prediction. In order to further improve the learning performance of the classification prediction model, we propose to integrate the single SVM classifier with Bagging algorithm. This classification algorithm can avoid over-fitting and effectively reject unsupported vectors. Under the 5-fold cross-validation test, the ACC and MCC of the druggable proteins dataset reach 93.78% and 0.8781 respectively. Compared with other works of literature, the experimental results show that the feature extraction, feature selection, ensemble learning and other algorithms make the prediction accuracy of the model significantly improved. Therefore, the druggable proteins prediction model based on genetic algorithm and Bagging-SVM ensemble classifier plays a key role in the success of drug development. The accumulation of various types of data in proteomics, genomics, and the improvement of bioinformatics methods also provide new methods for discovering new drug targets. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/GA-Bagging-SVM for academic use.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgements

The authors sincerely thank the anonymous reviewers for their many valuable comments that have improved this manuscript. This work was supported by the National Natural Science Foundation of China (Nos. 61863010 and 11771188), the Key Research and Development Program of Shandong Province of China (2019GGX101001), the Natural Science Foundation of Shandong Province of China (No. ZR2018MC007), the Project of Shandong Province Higher Educational Science and Technology Program (No. J17KA159), the Scientific Research Fund of Hunan Provincial Key Laboratory of Mathematical Modeling and Analysis in Engineering (No. 2018MMAEZD10), the Key Laboratory Open Foundation of Shandong Province, the College Students' Innovative Practice Training Program of Chinese Academy of Sciences, and the College Students' Innovative Entrepreneurial Training Program (No. 201710426046).

References

[1] Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov 2002;1:727–30.
[2] Chou KC. Prediction of g-protein-coupled receptor classes. J Proteome Res 2005;4:1413–8.
[3] Xiao X, Wang P, Chou KC. GPCR-CA: a cellular automaton image approach for predicting G-protein-coupled receptor functional classes. J Comput Chem 2009;30:1414–23.
[4] Drews J. Drug discovery: a historical perspective. Science 2000;287:1960–4.
[5] Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics 2014;26:i246–54.
[6] Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nat Rev Drug Discov 2006;5:993–6.
[7] Lindsay MA. Finding new drug targets in the 21st century. Drug Discov Today 2005;10:1683–7.
[8] Matter H. Computational medicinal chemistry for drug discovery. Drug Discov Today 2004;9:350–6.
[9] Dai YF, Zhao XM. A survey on the computational approaches to identify drug targets in the postgenomic era. BioMed Res Int 2015;2015:239654.
[10] Kumari P, Nath A, Chaube R. Identification of human drug targets using machine-learning algorithms. Comput Biol Med 2015;56:175–81.
[11] Murakami Y, Tripathi LP, Prathipati P, Mizuguchi K. Network analysis and in silico prediction of protein–protein interactions with applications in drug discovery. Curr Opin Struct Biol 2017;44:134–42.
[12] Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu YZ. A survey of current trends in computational drug repositioning. Brief Bioinform 2016;17:2–12.
[13] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 2001;46:3–26.
[14] Wang YY, Nacher JC, Zhao XM. Predicting drug targets based on protein domains. Mol BioSyst 2012;8:1528–34.
[15] Nakashima H, Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol 1994;238:54–61.
[16] Feng ZP. An overview on predicting the subcellular location of a protein. Silico Biol 2002;2:291–303.

[17] Mishra A, Pokhrel P, Hoque MT. StackDPPred: a stacking based prediction of DNA-binding protein from sequence. Bioinformatics 2019;35:433–41.
[18] Saini H, Raicar G, Sharma A, Lal S, Dehzangi A, Ananthanarayanan R, et al. Protein structural class prediction via k-separated bigrams using position specific scoring matrix. J Adv Comput Intell 2014;18:474–9.
[19] Sharma A, Dehzangi A, Lyons J, Imoto S, Miyano S, Nakai K, et al. Evaluation of sequence features from intrinsically disordered regions for the estimation of protein function. PLoS One 2014;9:e89890.
[20] Bu WS, Feng ZP, Zhang Z, Zhang CT. Prediction of protein (domain) structural classes based on amino-acid index. Eur J Biochem 2010;266:1043–9.
[21] Lin H, Ding H. Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. J Theor Biol 2011;269:64–9.
[22] Feng ZP, Zhang CT. A graphic representation of protein sequence and predicting the subcellular locations of prokaryotic proteins. Int J Biochem Cell Biol 2002;34:298–307.
[23] Barneh F, Jafari M, Mirzaie M. Updates on drug-target network; facilitating poly-pharmacology and data integration by growth of DrugBank database. Briefings Bioinf 2016;17:1070–80.
[24] Chou KC. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 2000;278:477–83.
[25] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.
[26] Ding H, Takigawa I, Mamitsuka H, Zhu S. Similarity-based machine learning methods for predicting drug-target interactions—a brief review. Brief Bioinform 2014;15:734–47.
[27] Kuang Q, Li Y, Wu Y, Li R, Dong Y. A kernel matrix dimension reduction method for predicting drug-target interaction. Chemometr Intell Lab Syst 2017;162:104–10.
[28] Shi H, Liu SM, Chen JQ, Li X, Ma Q, Yu B. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. Genomics 2018. https://doi.org/10.1016/j.ygeno.2018.12.007.
[29] Shen HB, Chou KC. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 2005;334:288–92.
[30] Byvatov E, Fechner U, Sadowski J, Schneider G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. J Chem Inf Comput Sci 2003;43:1882–9.
[31] Gönen M. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. Bioinformatics 2012;28:2304–10.
[32] Fernandez M, Caballero J, Fernandez L, Sarai A, et al. Genetic algorithm optimization in drug design QSAR: bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM). Mol Div 2011;15:269–89.
[33] González-DíAz H, Cruz-Monteagudo M, Molina R, Tenorio E, Uriarte E. Predicting multiple drugs side effects with a general drug-target interaction thermodynamic Markov model. Bioorg Med Chem 2015;13:1119–29.
[34] Chen H, Zhang R, Chen Z, Jiang Y, Shang Z. Predict potential drug targets from the ion channel proteins based on SVM. J Theor Biol 2010;262:750–6.
[35] Han LY, Zheng CJ, Xie B, Jia J, Ma XH. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. Drug Discov Today 2007;12:304–13.
[36] Ezzat A, Wu M, Li XL, Kwoh CK. Drug-target interaction prediction using ensemble learning and dimensionality reduction. Methods 2007;129:81–8.
[37] Zhang W, Liu F, Luo L, Zhang J. Predicting drug side effects by multi-label learning and ensemble learning. BMC Bioinf 2015;16:365–76.
[38] Zhang W, Zou H, Luo L, Liu Q, Wu W. Predicting potential side effects of drugs by recommender methods and ensemble learning. Neurocomputing 2016;173:979–87.
[39] Won HH, Kim MJ, Kim S, Kim JW. EnsemPro: an ensemble approach to predicting transcription start sites in human genomic DNA sequences. Genomics 2008;91:259–66.
[40] Iqbal S, Hoque MT. PBRpredict-Suite: a suite of models to predict peptide-recognition domain residues from protein sequence. Bioinformatics 2018;34:3289–99.
[41] Yu H, Chen J, Xu X, Li Y, Zhao HH, Fang YP, et al. A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. PLoS One 2012:e37608.
[42] Chen H, Zhang RJ, Chen ZQ, Jiang YS, Shang ZW, Sun P, et al. Predict potential drug targets from the ion channel proteins based on SVM. J Theor Biol 2010;262:750–6.
[43] Li Q, Lai L. Prediction of potential drug targets based on simple sequence properties. BMC Bioinf 2007;8:1–11.
[44] Han LY, Zheng CJ, Xie B, Jia J, Ma XH, Zhu F, et al. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. Drug Discov Today 2007;12:304–13.
[45] Zheng CJ, Han LY, Yap CW, Ji ZL, Cao ZW, Chen YZ, et al. Therapeutic targets: progress of their exploration and investigation of their characteristics. Pharmacol Rev 2006;58:259.
[46] Bao L, Sun Z. Identifying genes related to drug anticancer mechanisms using support vector machine. FEBS Lett 2002;521:109–14.
[47] Bhardwaj N, Langlois RE, Zhao G, Lu H. Kernel-based machine learning protocol for predicting DNA-binding proteins. Nucleic Acids Res 2005;33:6486–93.
[48] Cai CZ, Han LZ, Chen YZ. Enzyme family classification by support vector machines. Proteins: Struct Funct Bioinf 2004;55:66–76.
[49] Han L, Cui J, Lin H, Ji ZZ, Li Y. Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. Proteomics 2010;6:4023–37.
[50] Jamali AA, Ferdousi R, Razzaghi S, Li J, Safdari R. DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. Drug Discov Today 2016;21:718–24.
[51] Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 2008;24:i232–40.
[52] Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics 2009;25:2397–403.
[53] Chen X, Yan CC, Zhang XT, Zhang X, Dai F, Yin J, et al. Drug-target interaction prediction: databases, web servers and computational models. Brief Bioinform 2016;17:696–712.
[54] Niu B, Zhang Y, Ding J, Lu Y, Wang M. Predicting network of drug-enzyme interaction based on machine learning method. Biochim Biophys Acta 2014;1844:214–23.
[55] Ferdousi R, Safdari R, Omidi Y. Computational prediction of drug-drug interactions based on drugs functional similarities. J Biomed Inf 2017;70:54–64.
[56] Chen FS, Jiang ZR. Prediction of drug's anatomical therapeutic chemical (ATC) code by integrating drug–domain network. J Biomed Inf 2015;58:80–8.
[57] Azam SS, Shamim A. An insight into the exploration of druggable genome of streptococcus gordonii for the identification of novel therapeutic candidates. Genomics 2014;104:203–14.
[58] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014;42:D1091.
[59] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 2008;36:901–6.
[60] Knox C, Law V, Jewison T, Liu P, Ly S. DrugBank 3.0: a comprehensive resource for' Omics' research on drugs. Nucleic Acids Res 2011;39:D1035.
[61] Bakhtiarizadeh MR, Moradi-Shahrbabak M, Ebrahimi M, Ebrahimie E. Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology. J Theor Biol 2014;356:213–22.
[62] Bakheet TM, Doig AJ. Properties and identification of human protein drug targets. Bioinformatics 2009;25:451–7.
[63] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Struct Funct Bioinf 2001;43:246–55.
[64] Shen HB, Chou KC. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 2008;373:386–8.
[65] Khan A, Majid A, Hayat M. CE-PLoc: an ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition. Comput Biol Chem 2011;35:218–29.
[66] Zhang P, Tao L, Zeng X, Qin C, Chen SY, Zhu F, et al. PROFEAT update: a protein features web server with added facility to compute network descriptors for studying omics-derived networks. J Mol Biol 2017;429:416–25.
[67] Xu C, Ge L, Zhang Y, Dehmer M, Gutman I. Computational prediction of therapeutic peptides based on graph index. J Biomed Inf 2017;75:63–9.
[68] Han GS, Yu ZG Anh V, Krishnajith AP, Tian YC. An ensemble method for predicting subnuclear localizations from primary protein structures. PLoS One 2013;8:e57225.
[69] Berger JA, Mitra SK, Carli M, Neri A. Visualization and analysis of DNA sequences using DNA walks. J Franklin Inst 2004;341:37–53.
[70] Yu ZG, Anh V, Lau KS. Fractal analysis of measure representation of large proteins based on the detailed HP model. Physica A 2004;337:171–84.
[71] Holland J. Genetic algorithms. Sci Am 1992;267:66–72.
[72] Coit D. Genetic algorithms and engineering design. Eng. Econ 1998;43:379–81.
[73] Chowdhury B, Garai G. A review on multiple sequence alignment from the perspective of genetic algorithm. Genomics 2017;109:419–31.
[74] Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995.
[75] Guo YZ, Yu LZ, Wen ZN, Li ML. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Nucleic Acids Res 2008;36:3025–30.
[76] Zhang SB, Tang QR. Protein-protein interaction inference based on semantic similarity of gene ontology terms. J Theor Biol 2016;401:30–7.
[77] Rahmana MS, Rahman MK, Kaykobad M, Rahman MS. isGPT: an optimized model to identify sub-Golgi protein types using SVM and Random Forest based feature selection. Artif Intell Med 2018;84:90–100.
[78] Yu B, Zhang Y. The analysis of colon cancer gene expression profiles and the extraction of informative genes. J Comput Theor Nanosci 2013;10:1097–103.
[79] Du PF, Jiao YS. Predicting protein submitochondrial locations by incorporating the positional-specific physicochemical properties into Chou's general pseudo-amino acid compositions. J Theor Biol 2017;416:81–7.
[80] Khan M, Hayat M, Khan SA, Iqbal N. Unb-DPC: identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC. J Theor Biol 2017;415:13–9.
[81] Yu B, Lou LF, Li S, Zhang YS, Qiu WY, Wu X, et al. Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising. J Mole Graph Model 2017;76:260–73.
[82] Yu B, Li S, Chen C, Xu JM, Qiu WY, Wu X, et al. Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino composition. Chemometr Intell Lab Syst 2017;167:102–12.
[83] Yu B, Li S, Qiu WY, Chen C, Chen RX, Wang L, et al. Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising. Oncotarget 2017;8:107640–65.
[84] Xiang QL, Liao B, Li XH, Xu HM, Chen J, Shi ZX, et al. Subcellular localization prediction of apoptosis proteins based on evolutionary information and support vector machine. Artif Intell Med 2017;78:41–6.
[85] Wang LN, Shi SP, Xu HD, Wen PP, Qiu JD. Computational prediction of species-specific malonylation sites via enhanced characteristic strategy. Bioinformatics 2017;33:1457–63.

[86] Yu B, Li S, Qiu WY, Wang MH, Du JW, Zhang YS, et al. Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction. BMC Genomics 2018;19:478.
[87] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2011;2:1–27.
[88] Breiman L. Bagging predicators. Mach Learn 1996;24:123–40.
[89] Zhang J, Zhu MC, Chen P, Wang B. DrugRPE: random projection ensemble approach to drug-target interaction prediction. Neurocomputing 2017;228:256–62.
[90] Zhao F, Hou H, Bao Q, Wu J. PGA4genomics for comparative genome assembly based on genetic algorithm optimization. Genomics 2009;94:284–6.
[91] Yang J, Honavar V. Feature subset selection using a genetic algorithm. Springer US 1998;13:44–9.
[92] Anbarasi M, Anupriya E, Iyengar NCSN. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. Int J Eng Sci Res Technol 2010;2:5370–6.