Identification of Cancerlectins By Using Cascade Linear Discriminant Analysis and Optimal g-gap Tripeptide Composition

Article *in* Current Bioinformatics · October 2019

DOI: 10.2174/1574893614666190730103156

RESEARCH ARTICLE

Identification of Cancerlectins by Using Cascade Linear Discriminant Analysis and Optimal g-gap Tripeptide Composition

Liangwei Yang[1], Hui Gao[1,*], Keyu Wu[2], Haotian Zhang[2], Changyu Li[2] and Lixia Tang[2]

[1]Center for Informational Biology, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, P.R. China; [2]Center for Informational Biology, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, P.R. China

ARTICL EHISTORY

Received: November 27, 2018
Revised: January 20, 2019
Accepted: April 06, 2019

DOI:
10.2174/1574893614666190730103156

Abstract: ***Background***: Lectins are a diverse group of glycoproteins or glycoconjugate proteins that can be extracted from plants, invertebrates and higher animals. Cancerlectins, a kind of lectins, which play a key role in the process of tumor cells interacting with each other and are being employed as therapeutic agents. A full understanding of cancerlectins is significant because it provides a tool for the future direction of cancer therapy.

Objective: To develop an accurate and practically useful timesaving tool to identify cancerlectins. A novel sequence-based method is proposed along with a correlative webserver to access the proposed tool.

Method: Firstly, protein features were extracted in a newly feature building way termed, g-gap tripeptide composition. After which a proposed cascade linear discriminant analysis (Cascade LDA) is used to alleviate the high dimensional difficulties with the analysis of variance (ANOVA) as a feature importance criterion. Finally, support vector machine (SVM) is used as the classifier to identify cancerlectins.

Results: The proposed method achieved an accuracy of 91.34% with sensitivity of 89.89%, specificity of 92.48% and an 0.8318 Mathew's correlation coefficient based on only 13 fusion features in jackknife cross validation, the result of which is superior to other published methods in this domain.

Conclusion: In this study, a new method based only on primary structure of protein is proposed and experimental results show that it could be a promising tool to identify cancerlectins. An open-access webserver is made available in this work to facilitate other related works.

Keywords: Cancerlectin, Cascade LDA, g-gap tripeptide composition, ANOVA.

1. INTRODUCTION

Lectin is a kind of glycoprotein or glycoconjugate protein, which will specifically recognize and bind to diverse sugar structures [1]. It can mediate cell-cell interactions by combining with complementary carbohydrates on opposing cells. It plays a key role in the control of various normal and pathological processes in living organisms [2]. More than 300 lectins have been discovered from a variety of species, ranging from viruses and bacteria to plants and animals. Based on the monosaccharide for which they exhibit the highest affinity, lectins can be classified into five different categories: mannose, galactose/N-acetylgalactosamine, N-acetylglucosamine, fucose, and sialic acid [3]. They represent a heterogeneous group of oligomeric proteins that vary widely in size, structure, molecular organization, as well as constitution of their combining sites [2]. Due to their ability to recognize cell-surface carbohydrates with high specificity, lectins have been implicated in various essential biological processes, including cell-cell communication, cell proliferation, cell arrest, apoptosis, host-pathogen interactions, tissue development and tumor cell metastasis [4]. Lectins have been found to have great potential value in medicine due to their good attributes [2, 4-6].

Cancer is one of the major health problems with high mortality rate. The new cancer cases worldwide are estimated to increase to 19.3 million per year by 2025, due to the changing lifestyle and increase in longevity [7]. Although survival rates are improving for many types of cancer, it is still an enormous burden on society in more and less economically developed countries alike [8]. Cancerlectins are a group of lectins that are highly related with cancer initiation,

*Address correspondence to this author at the Center for Informational Biology, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, P.R. China; Tel/Fax: +8613550202554, E-mail: huigao@uestc.edu.cn

1574-8936/19 $58.00+.00

survival, growth, metastasis and spread [1, 9-11]. The major advantage of cancerlectins is that they can be found in natural sources, which by large mitigates the risk of side effects when used as a drug [7]. They promise a potential cancer treatment direction and have been widely studied from fundamental research to clinical application [12]. Most lectin studies were found to possess anti-carcinogenic effects. For example, Galactose binding lectin from the Chinese herb (Astragalus membranaceous) was shown to inhibit proliferation and induce apoptosis of human leukemia cells, in vitro [13]. Glycine max, purified from soybean, isolated by ion exchange chromatography was shown to attenuate the proliferation of breast cancer cells and hepatoma cells [14]. Mistletoe-lectin has the function to induce the cell apoptosis and inhibit the telomerase activity [15]. These encouraging study results in turn emphasize the importance of studies on lectins. It has therefore been suggested that an accurate and robust identification model of cancerlectins could be an important step to a full understanding of its behavior for the development of future cancer treatments.

Although being objective and accurate, the traditional wet-experimental methods based on biochemistry are costly and time-consuming. A faster and more accurate method is desirable for the identification of cancerlectins. With the development of protein database, more cancerlectin protein sequences are stored and computational methods are employed to handle the task. Kumar *et al.* proposed the first computational method on this problem using SVM and achieved an accuracy of 69.09% with MCC value of 0.38 [16]. Lin el al. put forward a new feature extraction method called g-gap dipeptide composition and got an optimal feature set in a stepwise way [17]. They achieved an accuracy of 75.19% in jackknife cross validation. Zhang el al. identified cancerlectins through hybrid machine learning technology by combining various kinds of protein features and classifiers [18]. Lai el al. achieved an accuracy of 77.48% with AUC of 85.52% by using the binomial distribution to screen the optimal feature set [19]. Yang *et al.* proposed a two-step feature selection method and got an accuracy of 74.8% based on random forest [20]. The increasing accuracy brings about the confidence in computational method for identification of cancerlectins.

In this study, we introduce a new feature extraction method called g-gap tripeptide composition (g-gap TC) and a new dimension reduction method named cascade linear discriminant analysis (Cascade LDA) to achieve a more accurate and robust model. Firstly, the features were extracted from the protein sequences by g-gap TC to get the sequence information. Then, the F value calculated by ANOVA was used as the feature importance criterion for feature selection. Subsequently, in order to avoid high dimensionality which is common when building protein features, Cascade LDA was performed to reduce the dimensions so as to build a robust identification model. Finally, SVM was used as the classifier to discriminate cancerlectins from non-cancerlectins. Built on only 13 fusion features, the classifier achieved an accuracy up to 91.34% with sensitivity of 89.89%, specificity of 92.48% in jackknife cross-validation. The model outperforms other state-of-the-art model by 13.86% in accuracy. Besides, by using only 13 fusion features, the classifier alleviates overfitting problem as much as possible. The results reveal that our proposed model may be a useful tool for identifying cancerlectins. For the convenience of other scholars, a user-friendly web-server was established and can freely be accessed from the website (http://bigroup.uestc.edu.cn/ services/ldapred/).

2. MATERIALS AND METHODS

2.1. Benchmark Dataset

Data is the foundation of bioinformatics analysis and plays a key role in the data mining process. A reliable and objective benchmark dataset is a key point in building a powerful classifier. The dataset used in this study consisting of 178 cancerlectin and 226 non-cancerlectin sequences were collected from the work of Lin [17]. The raw cancerlectin sequences were downloaded from CancerlectinDB [21] and non-cancerlectin sequences form UniProt Database [22]. Then, the following rules were considered to obtain a reliable benchmark dataset. Firstly, we removed the duplicated sequences and sequences without experimental evidence, or contain non-standard amino acids. Secondly, we removed the sequences tagged with “similar”, “fragment”, “putative” and “probable” in the non-cancerlectin sequences. Next, the CD-HIT tool was used to remove the high similarity sequences by setting a cutoff threshold of 50% to get rid of the redundant data. After following the previous strict screening procedures, a total of 178 cancerlectin and 226 non-cancerlectin sequences were obtained and used as the benchmark dataset in this study.

2.2. The g-gap Tripeptide Composition (g-gap TC)

The protein sequence is an indefinite length string consisting of 20 English letters which stand for 20 different kinds of amino acids. A protein sequence **P** with L residues with its entire amino acid sequence can be formulated as:

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3 \cdots \mathrm{R}_\mathrm{L} \tag{1}$$

where R_1 represents the 1st residue of protein **P**, R_2 represents the 2nd residue of protein **P**, and so forth. So far, sequence-similarity-search-based methods, such as BLAST algorithm [23] can be used to handle this kind of protein formulation but such methods fail to work when the query sequence doesn’t have significant similarity with any sequence in the database. To cater for this problem, we extracted feature vectors from the formulated sequence, which in effect a better and much easier approach and can be handled by employing machine learning or statistical methods. Many feature extraction methods have been proposed over years.

With the explosive growth of biological sequences in the post-genomic era, extracting features from proteins is not just one of the most important but also most difficult problems in computational biology. Amino acid composition (AAC) proposed by Nakashima and Nishikawa *et al.* [24] represents the protein sequence by a 20-dimensional vector. AAC assumes that protein characteristic is determined by the amino acid composition ratio but the information brought by sequence order is ignored. Chou *et al.* put forward the pseudo amino acid composition (PseAAC) [25] in consideration of the sequence-pattern information. They added physicochemical properties of amino acids when constructing features and have been widely used in nearly all areas of computational proteomics [24-28]. To generate various modes of Chou's special PseAAC, three open access soft-wares, 'PseAAC-Builder', 'propy', and 'PseAAC-General' were set up for the increasing usage of PseAAC [29]. Dubchak *et al.* proposed the composition, transformation and distribution (CTD) feature extraction method [30] which takes a global view of protein sequences. Lin *et al.* proposed the g-gap dipeptide composition (g-gap DC) [17] to obtain sequence-related information and obtained good results on cancerlectin identification. Applied on the same problem, Lai *et al.* used a tripeptide composition consisting of three amino acids in succession and built an 8000-dimensional feature for cancerlectin identification. By classifying 20 amino acids into 7 groups based on amino acid polarity and side chain group masses, Wang *et al.* calculated the frequency of each protein group and built a 343-dimensional feature into SVM for training [31]. Particularly, much more feature extraction methods of protein/peptide and DNA/RNA can be found in a recently updated website called 'Pse-in-One2.0' and related researches can access the website for more detailed information.

Inspired by the above methods, we proposed a feature extraction method called g-gap tripeptide composition to mine the information contained in protein sequences, which can be shown as:

$$\mathrm{F} = \mathrm{R}_1 \overset{gap_1}{\overbrace{\cdots}} \mathrm{R}_2 \overset{gap_2}{\overbrace{\cdots}} \mathrm{R}_3 \tag{2}$$

where R_1, R_2 and R_3 represent the standard amino acids and F is the feature we obtained. gap_1 represents gap between the first two residues and gap_2 represents the gap between the last two residues. There are 20 ×20 ×20 = 8000 kinds of tripeptide composition for 20 kinds of standard amino acids. For a specific gap values combination $gap_1 = g_1$ and $gap_2 = g_2$, the feature vector can be formulated as:

$$\mathbf{P} = \left[f_1^{g_1,g_2}, f_2^{g_1,g_2}, \cdots, f_\xi^{g_1,g_2}, \cdots, f_{8000}^{g_1,g_2}\right]^{\mathrm{T}} \tag{3}$$

where T represents a transpose operation, $f_\xi^{g_1,g_2}$ represents the frequency of the ξ-th $g_1\_g_2\_gap$ tripeptide composition. $f_\xi^{g_1,g_2}$ is calculated by

$$f_\xi^{g_1,g_2} = \frac{n_\xi^{g_1,g_2}}{\sum_{\xi=1}^{8000} n_\xi^{g_1,g_2}} = \frac{n_\xi^{g_1,g_2}}{L - g_1 - g_2 - 2} \tag{4}$$

where $\boldsymbol{n_\xi^{g_1,g_2}}$ represents the occurrence number of the ξ-th $\boldsymbol{g_1\_g_2\_gap}$ tripeptide composition, L is the length of protein P. In this study, the gap values vary from 0 to 9, so 100 8000-dimensional feature vectors were built. We assumed that some important features may hide in different gap value combinations. For that reason, all feature vectors were firstly combined together by simple splicing and then their dimensions were reduced to discover the important features. Before the dimension reduction, a total of $\mathbf{100 \times 8000 = 800000}$ features were obtained as candidates.

2.3. Feature Importance Criterion

As mentioned in the g-gap TC section, a high-dimension feature vector was obtained, which will lead to curse of dimensionality, low efficiency and mar the prediction accuracy of the model due to the redundant information emanated from g-gap TC. Thus, a wise strategy is to use feature selection techniques to judge the importance of each feature and ignore those features that are not relevant to the study. This will not only gain deeper insights into the intrinsic properties of protein sequences, but economize runtime and computational resource [32, 33]. ANOVA is a simple and effective way to test the difference between groups, the purpose of which is to find out the factors that have a significant impact on the data through data analysis. For the following advantages, ANOVA is widely used in feature selection and yielded good results [34-37]. Firstly, it obtains good results even if the data does not satisfy its theoretical hypothesis. Secondly, it can analyze the interaction between two features more intuitively and it is effective to use this method especially when each observation value is different in the group.

The principle of ANOVA is to calculate the ratio (F value) of features between groups and within groups for measuring feature variances. Then, the F value of the *u*-th feature (*F(u)*) in benchmark dataset is defined by:

$$F(u) = \frac{s_B^2(u)}{s_W^2(u)} \tag{5}$$

where $S_B^2(u)$ and $S_W^2(u)$ are the sample variance between groups (also called Means Square Between, MSB) and sample variance within groups (also called Mean Square Within, MSW) respectively, which are given by:

$$\begin{cases} s_B^2(u) = \dfrac{SS_B(u)}{df_B} \\ s_W^2(u) = \dfrac{SS_W(u)}{df_W} \end{cases} \tag{6}$$

where $df_B = K - 1$ and $df_W = M - K$ are degrees of freedom for MSB and MSW, respectively. *K* and *M* represent the number of groups (here *K* = 2) and total number of samples (here *M* = 404), respectively. $SS_B(u)$ and $SS_W(u)$ are sum of squares between groups and sum of squares within groups, respectively, which can be calculated by:

$$\begin{cases} SS_B(u) = \displaystyle\sum_{i=1}^{K} m_i \left(\frac{\sum_{j=1}^{m_i} \psi_u^g(i,j)}{m_i} - \frac{\sum_{i=1}^{K} \sum_{j=1}^{m_i} \psi_u^g(i,j)}{\sum_{i=1}^{K} m_i} \right)^2 \\ \square S_W(u) = \displaystyle\sum_{i=1}^{K} \sum_{j=1}^{m_i} \left(\psi_u^g(i,j) - \frac{\sum_{j=1}^{m_i} \psi_u^g(i,j)}{m_i} \right)^2 \end{cases} \tag{7}$$

where $\psi_u^g(i,j)$ denotes the frequency of the *u*-th g-gap tripeptide of the *j*-th sample in the *i*-th group; m_i denotes the number of samples in the *i*-th group (here $m_1 = 178$, $m_2 = 226$). Obviously, a large value of $F(u)$ means that the *u*-th feature has a better discriminative capability. Thus, the F value obtained from ANOVA was used as the feature importance criterion to judge the discriminative capability of each feature in this study. Feature importance criterion combined with Cascade LDA was used for dimension reduction, which is discussed in proceeding section.

2.4. Dimension Reduction

As mentioned in the previous section, a total of 800000 features were extracted from the protein sequences as candidates for identifying cancerlectins. However, using such number of features would definitely lead to the curse of dimensionality, resulting in discrepancies in accuracy and an inefficient model. When facing high dimensional data, dimension reduction will not only ease the computing problem but also extract the important information brought by the features. Dimension reduction has been successfully applied to image classification [38-44] in situations where dimensions are large. These methods can be broadly categorized into unsupervised methods and supervised methods depending on the completeness of the data. Due to the good performance of supervised methods and that, the data used in this study were all labeled, we considered supervised dimension reduction methods. There have been lots of supervised dimension reduction methods proposed, such as: linear discriminant analysis (LDA) [45], marginal Fisher analysis (MFA) [39], adaptive slow feature discriminant analysis [43], locality preserving projection (LPP) [46], neighborhood preserving embedding (NPE) [47] and local Fisher discriminant analysis (LFDA) [48]. Among all the algorithms mentioned above, LDA is the most commonly used dimension reduction method for classification. The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible and it seeks to find directions along which the classes are best separated. LDA best fits for this study since is promises good performance and also widely applicable for dimension reduction.

LDA does dimension reduction by taking into consideration the scatter within and between classes respectively. Suppose there are C classes (C = 2 in this study), let μ_i and M_i be the mean vector and sample number of class i, $i = 1,2,\cdots,C$, separately. Then $M = \sum_{i=1}^{C} M_i$ is the total number of samples (M = 404 in this study). The within-class scatter matrix S_w and the between-class scatter matrix S_b are defined by:

$$\begin{cases} \mathrm{S}_w = \sum_{i=1}^{C}\sum_{j=1}^{M_i}(x_j-\mu_i)(x_j-\mu_i)^T \\ \mathrm{S}_b = \sum_{i=1}^{C}(\mu_i-\mu)(\mu_i-\mu)^T \end{cases} \tag{8}$$

where $\mu = \frac{1}{C}\sum_{i=1}^{C}\mu_i$, stands for the mean of entire dataset. LDA computes a transformation that maximizes the between-class scatter while minimizing the within-class scatter:

$$\text{maximize } \frac{det\ (S_b)}{det\ (S_w)} \tag{9}$$

If we reduce dimension directly using LDA, the problem of calculating distance in high-dimensional space cannot be avoided. Moreover, such operation may increase computational cost with poor results. We therefore propose a new dimension reduction method based on LDA named Cascade LDA to alleviate the aforementioned problem.

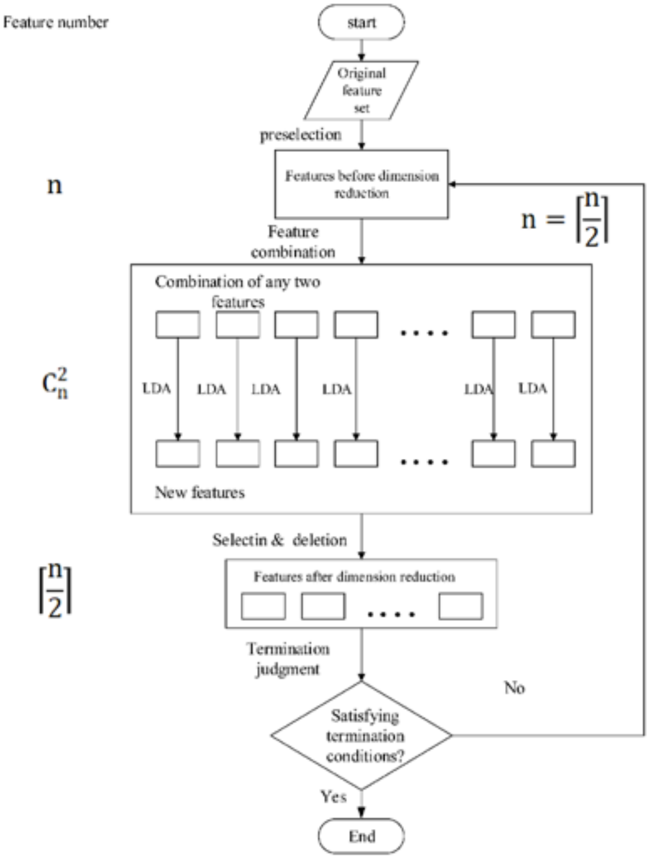

Feature number
start
Original feature set
preselection
n
Features before dimension reduction
$n = \left\lceil \frac{n}{2} \right\rceil$
Feature combination
Combination of any two features
C_n^2
LDA
LDA
LDA
LDA
LDA
LDA
....
New features
Selectin & deletion
$\left\lceil \frac{n}{2} \right\rceil$
Features after dimension reduction
Termination judgment
No
Satisfying termination conditions?
Yes
End

Fig. (1). Flowchart of Cascade LDA. *(*A higher resolution / colour version of this figure is available in the electronic copy of the article*)*.

The dimension reduction process is shown in Fig. (**1**). The process can be divided into 4 steps, namely, preselection, feature combination, selection & deletion and termination judgment. We will clarify the steps one by one in detail in the following paragraphs.

Step1: preselection, rank all the original features in descending order based on evaluating indicator such as ANOVA and information gain and select the best n features. In this study, we ranked all the 800000 original features in descending order based on F value obtained from ANOVA. Taking computational complexity and sample size (M = 404) into consideration, we selected the initial 800 original features in the preselection step since the quantity was nearly twice the sample size and these selected features were sufficient to capture the important information in the dataset.

Step2: feature combination, combine any two different features and perform LDA on each 2-dimension feature space obtained from the combination; hence, the number of features before step2 n will be C_n^2 after it. At this strategy, LDA was performed, $C_{800}^2 = 319600$ times in the first round and each time we obtained a new fusion feature.

Step3: selection & deletion, based on evaluating indicator; repeatedly select the best fusion feature and delete the fusion features whose original 2 features overlap with the selected one's until there is no fusion feature left. Thus, after step3, features have been reduced from n to $\left\lceil\frac{n}{2}\right\rceil$ and can be used for classification. We selected ANOVA as our evaluating indicator and obtained 400 fusion features after the first round.

Step4: termination judgment; decide when to stop the dimension reduction procedure. If it is not terminated, the fusion features obtained from the last round will be used as the input of step2 and a new round begins. Theoretically, the feature space can be reduced to 1 dimension as the cycle proceeds but scholars may terminate the procedure based on their own needs. To capture the effect of this newly proposed dimension reduction technique, we reduced the dimension from 800 to 1. Detailed results can be found in the results and discussion section.

2.5. Support Vector Machine (SVM)

SVM has been successfully applied to a number of applications, such as character recognition, face identification, speaker verification, image classification [49], bioinformatics [50-53], *etc*. The success of SVM in solving real-life problems made it not only a tool for theoretical analysis but also a tool for creating practical algorithms for real-world problems [54]. By seeking a hyperplane which makes the separation interval between classes maximal, SVM relieves the overfitting problem in a great deal. The formulation embodies the Structural Risk Minimization (SRM) principle, which has been shown to be superior to traditional Empirical Risk Minimization (EMR) principle employed by conventional neural networks [55]. The most important character of SVM is that it performs well when dealing with high dimensional feature space, which is common in bioinformatics. For the reasons mentioned above, we adopted SVM as our classification algorithm. A grid search method was used to optimize the regularization parameter C and kernel parameter γ through 5-fold cross-validation. The search spaces for C and γ are $[2^4, 2^8]$ and $[2^{-6}, 2^{-9}]$, respectively. The final exact values of the two parameters after grid search in this study were 64 for C and 2.762×10^{-3} for γ. Note that before using the SVM to train the data, the experimental data was normalized. Zero-mean normalization is the most common standardized method, also known as standard deviation standardization. The method is based on the mean μ and standard deviation σ of the original data to standardize the data. The distribution of processed data meets the standard normal distribution, which means the mean is 0 and the standard deviation is 1. Its conversion method is:

$$\mathrm{x}^* = \frac{x-\mu}{\sigma} \tag{10}$$

2.6. Performance Evaluation

To objectively evaluate the performance of a model, researches usually adopt one of the three kinds of test methods, namely, independent dataset test, K-fold cross-validation and jackknife test [17, 56-63]. Independent dataset test applies the model trained from training set on mutually independent test set, which are split from the original dataset at first. Although this method seems simple, it cannot guarantee a stable result for different splits. Besides, the resulting model will also be affected by the partially use of data. K-fold cross-validation divides the dataset into k un-overlapping subsets, each of which is used as the test set in turn and the rest k-1 ones are used to train the classifier. The results of k tests are combined to measure the performance of model. Jackknife test can be seen as a special case for K-fold cross-validation when K is exactly the same as sample numbers. All samples in the benchmark dataset will be chosen one by one and tested by the predictor trained from the remaining samples. We adopted the jackknife test to examine the quality of the proposed model since among the three methods it is deemed as the least arbitrary that can always yield a unique result for a given benchmark dataset [64]. Besides, the jackknife test has been widely used to examine the quality of the predictor. There are several kinds of evaluation metrics used to estimate the model performance. In the function identification of protein sequences, researchers usually use accuracy (Acc), sensitivity (S_n), specificity (S_p), and Mathew's correlation coefficient (MCC), which is calculated by

$$\begin{cases} S_n = 1 - \dfrac{N_-^+}{N^+} \\ S_p = 1 - \dfrac{N_+^-}{N^-} \\ Acc = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} \\ MCC = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} \end{cases} \tag{11}$$

where N^+, N^- represents the number of positive and negative samples respectively. N_-^+, N_+^- represents the number of samples in which the positive sample is mistakenly classified into negative samples and the negative samples are mistakenly divided into positive samples. Of the aforementioned indicators, the most important are the Acc and MCC. Acc reflects the overall accuracy and MCC represents the reliability of the algorithm results. S_n, S_p can

Table 1. Results of dimension reduction procedure.

Dimension	S_n(%)	S_p(%)	Acc (%)
800	89.33	90.71	90.10
400	88.20	90.71	89.60
200	89.89	89.38	89.60
100	87.08	92.04	89.85
50	87.64	94.25	91.34
25	88.76	92.04	90.59
13	89.89	92.48	91.34
7	89.89	92.04	91.09
4	89.33	92.48	91.09
2	90.45	91.59	91.09
1	91.01	90.71	90.84

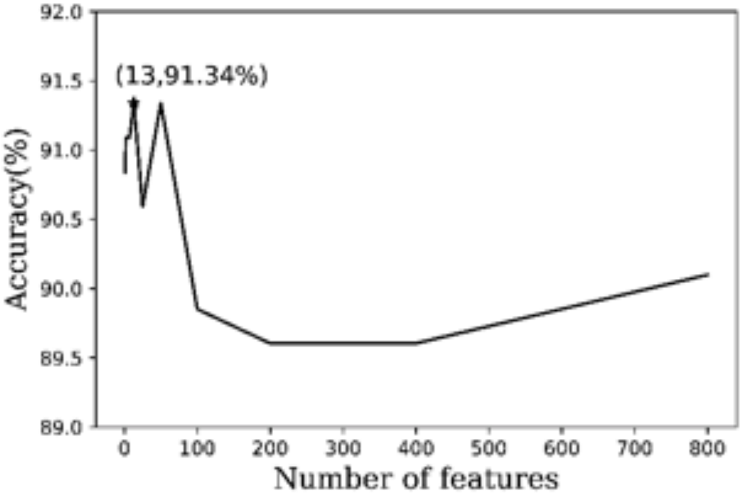


Fig. (2). Changes in accuracy with dimension reduction procedure. (*A higher resolution / colour version of this figure is available in the electronic copy of the article).*

Table 2. The comparison with other methods.

Method	S_n(%)	S_p(%)	Acc (%)	MCC
g-gap DC	69.10	80.10	75.19	0.5499
g-gap DC + Cascade LDA	80.34	84.51	82.67	0.679
g-gap TC	81.46	94.70	88.86	0.7856
g-gap TC + Cascade LDA	89.89	92.48	91.34	0.8318

be seen as the recall rates of positive and negative categories respectively. Besides, the receiver operating characteristic (ROC) curves and area under curve (AUC) are also used for an intuitive comparison between methods. The ROC curve takes the true positive rate as the vertical axis and the false positive rate as the horizontal axis. We can get the ROC curve by connecting the points got by setting different classification thresholds. AUC is the area under the ROC curve. The greater the value of AUC, the stronger the predictability of the model.

3. RESULTS AND DISCUSSION

3.1. Dimension Reduction by Using Cascade LDA

As mentioned in the dimension reduction section, Cascade LDA was applied in this study to ease the problem caused by high-dimensional feature space. We first chose the best 800 features from the original feature set based on F value got from ANOVA. Then the cyclic dimensionality reduction procedure was applied and reduced the dimension by half at a time. The detailed results are shown in Table **1** and the changes in accuracy are also plotted in Fig. (**2**).

From Table **1**, we can see that all the metrics vary around 90%. A continuous high Acc with a balanced $\boldsymbol{S_n}$ and $\boldsymbol{S_p}$ demonstrates the model proposed is insensitive to the dimension reduction procedure and can always yield an impressive result. It can also be seen that there is a slight increase in all the metrics with the reduction of dimensions. The result shows that the dimension reduction procedure used in this study could maintain and even improve the prediction accuracy. In Fig. (**2**), it shows that the accuracy rises gradually after a slight fall at first and reaches the best Acc when the dimension is 13 and 50. However, in the study of cancerlectin identification, $\boldsymbol{S_n}$ is more important than $\boldsymbol{S_p}$ since researchers usually yearn for all the cancerlectins identified. Therefore, we take the 13-dimensional fusion feature space as our final result.

3.2. Comparison with Other Methods

In the study of cancerlectin identification, some computational models have been developed in varied approaches as mentioned in introduction section, among which Lin's method is the most similar to ours. Therefore, we make a comparison with his method named g-gap DC since our proposed methods stands superior over existing state-of -the-art classifiers. Two more contrast experiments were made to explicit the advantage of the proposed g-gap TC and Cascade LDA dimensionality reduction. The g-gap DC + Cascade LDA method adds a Cascade LDA procedure as explicated in the dimension reduction section and g-gap TC method eliminates the Cascade LDA procedure compared to the proposed method named g-gap TC + Cascade LDA. Detailed metrics of results are shown in Table **2**.

It can be seen from Table **2** that the proposed method achieves the best sensitivity, Acc, and MCC with specificity only slightly worse than g-gap TC. With sensitivity and specificity all around 90%, the model proposed shows robustness on both positive and negative samples. An Acc up

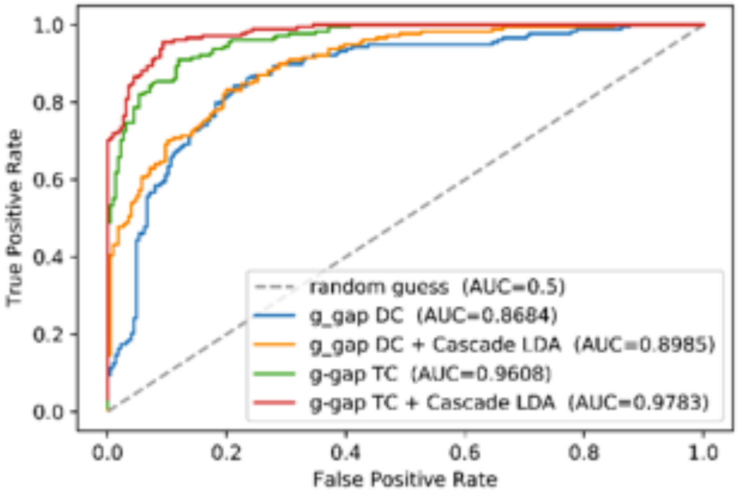
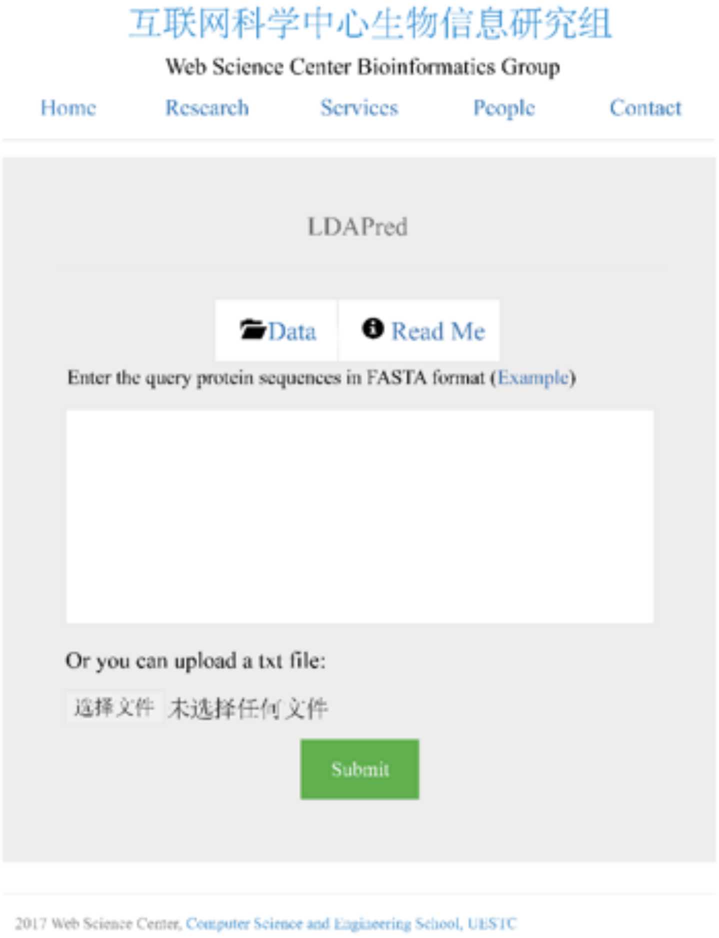


Fig. (3). ROC curves and AUC values of the contrast methods. *(A higher resolution / colour version of this figure is available in the electronic copy of the article).*



互联网科学中心生物信息研究组

Web Science Center Bioinformatics Group

Enter the query protein sequences in FASTA format (Example)

Or you can upload a txt file:

2017 Web Science Center, Computer Science and Engineering School, UESTC

Fig. (4). A semi-screenshot of the LDAPred webserver. *(A higher resolution / colour version of this figure is available in the electronic copy of the article).*

to 91.34% with an MCC over 0.8 indicate that our method could be an accurate and reliable tool for identification of cancerlectins. By comparing the two methods with Cascade LDA and the other two without, we can see that the methods with Cascade LDA have a better or comparative performance on all metrics than the others when the feature is fixed. The result reveals that when dealing with a high-dimensional feature space, we could consider inducing the Cascade LDA dimension reduction procedure to enhance the performance of prediction. It can also be seen that the methods utilizing g-gap TC as features yield a better performance on all metrics than using g-gap DC. Our proposed method surpasses Lin's by 16.15% in Acc and 0.2819 in MCC, which is a significant improvement in cancerlectin identification. For a more intuitive comparison, we also plotted the four method's ROC curves with their AUC value as shown in Fig. (**3**).

In Fig. (**3**), our method's ROC curve wraps up all the other curves with the largest AUC value of 0.9783. The methods with Cascade LDA procedure have a slightly larger AUC value compared with those without when utilizing the same feature and methods using g-gap TC as features have a significant improvement than g-gap DC. The experimental results of ROC curves agree with the previous analysis in Table **1**, indicating that our proposed method could be a promising tool in the work of cancerlectin identification.

4. WEB-SERVER

As demonstrated in a series of recent publications [65-75], user-friendly and open access web-servers demonstrate the future direction of developing practical and more useful forecasting methods and computing tools. We have also established a web-server named LDApred as shown in Fig. 4. to facilitate related works. Researches can access the web-server at http://bigroup.uestc.edu.cn/services/ldapred/ and upload protein sequences in FASTA format either as a single file or copied/pasted into the input box. The identification results will be shown in a new interface soon after clicking the submit button.

CONCLUSION

Cancer is one of the most serious health problems which threatens mortality rate among all human races. Cancerlectins, which can be extracted from natural sources, are highly related with the physiological processes of cancer cells. They shed light on the potential therapy direction of cancer and more and more researches are focusing on their role or microarray analysis in cancer prevention, detection, treatment and diagnosis. It is therefore significant to have an accurate and open-accessed identification tool for related works. In this study, we propose a new method based on g-gap TC feature extraction method and Cascade LDA dimension reduction procedure. The proposed method achieves a high accuracy up to 91.34% with an MCC of 0.8318 in jackknife cross-validation, which outperforms all the other state-of-the-art models. The impressive results show that the proposed model could be adopted as an improved method of identifying cancerlectins. Researchers of interest may access the website for related works.

In the future, we will seek to find the influence caused by different dimensionality reduction strategies and apply it on high-dimensional related problems [76, 77]. Besides, we are eager to build high accurate computational prediction models for bioinformatics problems in order to provide more practical useful tools in this area of study.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are base of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

Not applicable.

FUNDING

This research was funded by the National Natural Science Foundation of China (No. 26531771471, No. 21673034, No. 61673085 and No. 71661167005) and the authors declare no interest of any potential conflict.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Conceptualization, Hui Gao, Keyu Wu and Lixia Tang; Methodology, Liangwei Yang, Hui Gao and Haotian Zhang; Software, Songtao Li; Investigation, Keyu Wu, Changyu Li, Haotian Zhang; Writing-Original Draft Preparation, Liangwei Yang and Hui Gao; Writing-Review & Editing, Liangwei Yang, Hui Gao and Lixia Tang; Supervision, Mensah Dennis N.A, Hui Gao, Lixia Tang; Funding Acquisition, Hui Gao, Duanbing Chen and Lixia Tang.

REFERENCES

[1] Lotan R, Raz A. Lectins in Cancer Cells. Ann N Y Acad Sci. 2010;551(1):385-98.

[2] Sharon N, Lis H. Lectins as cell recognition molecules. Science. 1989;246(4927):227-34.

[3] Hu S, Wong DT. Lectin microarray. PROTEOMICS - Clinical Applications. 2009;3(2):148-54.

[4] Sharon N. Lectins: Properties, Functions and Applications in Biology and Medicine. Kitasato Medicine. 1986;18:109-10.

[5] Beuth J, Ko HL, Pulverer G, Uhlenbruck G, Pichlmaier H. Importance of lectins for the prevention of bacterial infections and cancer metastases. Glycoconj J. 1995;12(1):1-6.

[6] Bevilacqua MP, Nelson RM. Lectins. J Clin Invest. 1993;91(2):379-87.

[7] Jamal S, Lavanya V, Adil AM, Ahmed N. Lectins-the promising cancer therapeutics. Oncobiology & Targets. 2014;1(1):12.

[8] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics, 2012. CA Cancer J Clin. 2011;61(2):69-90.

[9] Sherwani AF, Mohmood S, Khan F, Khan RH, Azfer MA. Characterization of lectins and their specificity in carcinomas—An appraisal. Indian J Clin Biochem. 2003;18(2):169.

[10] Liu FT, Rabinovich GA. Galectins as modulators of tumour progression. Nature Reviews Cancer. 2005;5(1):29-41.

[11] Gorelik E, Galili U, Raz A. On the role of cell surface carbohydrates and their binding proteins (lectins) in tumor metastasis. Cancer Metastasis Rev. 2001;20(3-4):245-77.

[12] Young LS, Searle PF, Onion D, Mautner V. Viral gene therapy strategies: from basic science to clinical application. J Pathol. 2010;208(2):299-318.

[13] Huang LH, Yan QJ, Kopparapu NK, Jiang ZQ, Sun Y. Astragalus membranaceus lectin (AML) induces caspase‐dependent apoptosis in human leukemia cells. Cell Prolif. 2012;45(1):15-21.

[14] Peng, Xiujuan. Purification of melibiose-binding lectins from two cultivars of Chinese black soybeans. Acta Biochimica Et Biophysica Sinica. 2010;40(12):1029-38.

[15] Choi SH, Lyu SY, Park WB. Mistletoe lectin induces apoptosis and telomerase inhibition in human A253 cancer cells through dephosphorylation of Akt. Arch Pharm Res. 2004;27(1):68-76.

[16] Kumar R, Panwar B, Chauhan JS, Raghava GP. Analysis and prediction of cancerlectins using evolutionary and domain information. BMC Res Notes. 2011;4(1):237.

[17] Hao L, Liu WX, Jiao H, Liu XH, Hui D, Wei C. Predicting cancerlectins by the optimalg-gap dipeptides. Sci Rep. 2015;5:16964.

[18] Zhang J, Ju Y, Lu H, Xuan P, Zou Q. Accurate Identification of Cancerlectins through Hybrid Machine Learning Technology. International Journal of Genomics,2016,(2016-7-13). 2016;2016(4):1-11.

[19] Lai H-Y, Chen X-X, Chen W, Tang H, Lin H. Sequence-based predictive modeling to identify cancerlectins. Oncotarget. 2017;8(17):28169.

[20] Yang R, Zhang C, Zhang L, Gao R. A Two-Step Feature Selection Method to Predict Cancerlectins by Multiview Features and Synthetic Minority Oversampling Technique. BioMed Research International,2018,(2018-2-7). 2018;2018(1):1-10.

[21] Damodaran D, Jeyakani J, Chauhan A, Kumar N, Chandra NR, Surolia A. CancerLectinDB: a database of lectins relevant to cancer. Glycoconj J. 2008;25(3):191-8.

[22] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S*, et al.* UniProt: the Universal Protein knowledgebase.

[23] Lobo I. Basic Local Alignment Search Tool (BLAST). J Mol Biol. 2012;215(3):403-10.

[24] Nakashima H, Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. Journal of Molecular Biology. 1994;238(1):54.

[25] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins Structure Function & Bioinformatics. 2010;43(3):246-55.

[26] Mei J, Zhao J. Analysis and prediction of presynaptic and postsynaptic neurotoxins by Chou's general pseudo amino acid composition and motif features. Journal of Theoretical Biology. 2018;447:147.

[27] Muthukrishnan S. Using Chou's general PseAAC to analyze the evolutionary relationship of receptor associated proteins (RAP) with various folding patterns of protein domains. J Theor Biol. 2018;445:62.

[28] Rahman MS, Shatabda S, Saha S, Kaykobad M, Rahman MS. DPP-PseAAC: A DNA-binding Protein Prediction model using Chou’s general PseAAC. Journal of Theoretical Biology. 2018;452.

[29] Chou KC. Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology. Current Proteomics. 2009;6(4):-.

[30] Dubchak I, Muchnik I, Holbrook SR, Kim SH. Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci U S A. 1995;92(19):8700-4.

[31] Wang H, Hu X. Accurate prediction of nuclear receptors with conjoint triad feature. Bmc Bioinformatics. 2015;16(1):402.

[32] Zou Q, Zeng J, Cao L, Ji R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing. 2016;173:346-54.

[33] Zou Q, Wan S, Ju Y, Tang J, Zeng X. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. Bmc Systems Biology. 2016;10(4):114.

[34] Ding H, Feng PM, Chen W, Lin H. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Mol Biosyst. 2014;10(8):2229-35.

[35] Ding H, Guo SH, Deng EZ, Yuan LF, Guo FB, Huang J*, et al.* Prediction of Golgi-resident protein types by using feature selection technique. Chemometrics & Intelligent Laboratory Systems. 2013;124(6):9-13.

[36] Ding H, Li D. Identification of mitochondrial proteins of malaria parasite using analysis of variance. Amino Acids. 2015;47(2):329-33.

[37] Lin H, Chen W, Ding H. AcalPred: A Sequence-Based Tool for Discriminating between Acidic and Alkaline Enzymes. PLoS One. 2013;8(10):e75726.

[38] Ling Y, Yin X, Bhandarkar SM, editors. Sirface vs. Fisherface: recognition using class specific linear projection. International Conference on Image Processing, 2003 ICIP 2003 Proceedings; 2003.

[39] Yan S, Xu D, Zhang B, Zhang HJ, Yang Q, Lin S. Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Transactions on Pattern Analysis & Machine Intelligence. 2007;29(1):40.

[40] Yang J, Zhang L, Yang J-y, Zhang D. From classifiers to discriminators: A nearest neighbor rule induced discriminant analysis. Pattern Recognition. 2011;44(7):1387-402.

[41] Jin Z, Yang JY, Hu ZS, Lou Z. Face recognition based on the uncorrelated discriminant transformation. Pattern Recognition. 2001;34(7):1405-16.

[42] Wang S, Gu X, Lu J, Yang JY, Wang R, Yang J, editors. Unsupervised Discriminant Canonical Correlation Analysis for Feature Fusion. International Conference on Pattern Recognition; 2014.

[43] Gu X, Liu C, Wang S, Zhao C. Feature extraction using adaptive slow feature discriminant analysis. Neurocomputing. 2015;154(C):139-48.

[44] Feng P-M, Chen W, Lin H, Chou K-C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Analytical Biochemistry. 2013;442(1):118-25.

[45] Belhumeur PN, Hespanha JP, Kriegman DJ. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Transactions on Pattern Analysis & Machine Intelligence. 2002;19(7):711-20.

[46] Pami IT, Kingravi HA. Face Recognition Using Laplacianfaces. 2005.

[47] He X, Cai D, Yan S, Zhang HJ, editors. Neighborhood Preserving Embedding. Tenth IEEE International Conference on Computer Vision; 2005.

[48] Sugiyama M. Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis. Jmachlearnres. 2007;8(1):1027-61.

[49] Zhang N, Yu S, Guo Y, Wang L, Wang P, Feng Y. Discriminating Ramos and Jurkat Cells with Image Textures from Diffraction Imaging Flow Cytometry Based on a Support Vector Machine. Current Bioinformatics. 2018;13:50-6.

[50] Li D, Ju Y, Zou Q. Protein Folds Prediction with Hierarchical Structured SVM. Current Proteomics. 2016;13(2):79-85.

[51] Wang SP, Zhang Q, Lu J, Cai YD. Analysis and Prediction of Nitrated Tyrosine Sites with the mRMR Method and Support Vector Machine Algorithm. Current Bioinformatics. 2018;13(1):3-13.

[52] Chen W, Yang H, Feng P, Ding H, Lin H. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics. 2017.

[53] Chen W, Xing P, Zou Q. Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines. Scientific reports. 2017;7:40242.

[54] Wang X, Zhong Y, editors. Statistical Learning Theory and State of the Art in SVM. IEEE International Conference on Cognitive Informatics; 2003.

[55] Brereton RG, Lloyd GR. Support vector machines for classification and regression. Analyst. 1998;135(2):230-67.

[56] Chen XX, Hua T, Li WC, Hao W, Wei C, Hui D*, et al.* Identification of Bacterial Cell Wall Lyases via Pseudo Amino Acid Composition. BioMed Research International,2016,(2016-6-29). 2016;2016(4):1-8.

[57] Chou KC, Zhang CT. Prediction of Protein Structural Classes, Critical Reviews in Biochemistry and Molecular Biology, Informa Healthcare. Informa Uk Ltd Uk. 2008.

[58] Zhu XJ, Feng CQ, Lai HY, Chen W, Lin H. Predicting protein structural classes for low-similarity sequences by evaluating different features. Knowledge-Based Systems. 2018.

[59] Su ZD, Huang Y, Zhang ZY, Zhao YW, Wang D, Chen W*, et al.* iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. Bioinformatics. 2018.

[60] Yang H, Tang H, Chen XX, Zhang CJ, Zhu PP, Ding H*, et al.* Identification of Secretory Proteins in Mycobacterium tuberculosis Using Pseudo Amino Acid Composition. BioMed research international. 2016;2016:5413903.

[61] Tang H, Chen W, Lin H. Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. Molecular bioSystems. 2016;12(4):1269-75.

[62] Feng PM, Lin H, Chen W. Identification of antioxidants from sequence information using naive Bayes. Computational and mathematical methods in medicine. 2013;2013:567529.

[63] Feng PM, Ding H, Chen W, Lin H. Naive Bayes classifier with feature selection to identify phage virion proteins. Computational and mathematical methods in medicine. 2013;2013:530696.

[64] Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol. 2011;273(1):236-47.

[65] Yang H, Qiu WR, Liu GQ, Guo FB, Chen W, Chou KC*, et al.* iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC. Int J Biol Sci. 2018;14(8):883-91.

[66] Yang H, Lv H, Ding H, Chen W, Lin H. iRNA-2OM: A Sequence-Based Predictor for Identifying 2'-O-Methylation Sites in Homo sapiens. Journal of computational biology : a journal of computational molecular cell biology. 2018;25(11):1266-77.

[67] Tang H, Zhao YW, Zou P, Zhang CM, Chen R, Huang P*, et al.* HBPred: a tool to identify growth hormone-binding proteins. Int J Biol Sci. 2018;14(8):957-64.

[68] He WY, Jia CZ, Duan YC, Zou Q. 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. Bmc Syst Biol. 2018;12.

[69] Feng CQ, Zhang ZY, Zhu XJ, Lin Y, Chen W, Tang H*, et al.* iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics. 2018.

[70] Dao FY, Lv H, Wang F, Feng CQ, Ding H, Chen W*, et al.* Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. Bioinformatics. 2018.

[71] Zhang T, Tan P, Wang L, Jin N, Li Y, Zhang L*, et al.* RNALocate: a resource for RNA subcellular localizations. Nucleic acids research. 2017;45(D1):D135-D8.

[72] Yi Y, Zhao Y, Li C, Zhang L, Huang H, Li Y*, et al.* RAID v2.0: an updated resource of RNA-associated interactions across organisms. Nucleic acids research. 2017;45(D1):D115-D8.

[73] Tang H, Zhang CM, Chen R, Huang P, Duan CG, Zou P. Identification of Secretory Proteins of Malaria Parasite by Feature Selection Technique. Lett Org Chem. 2017;14(9):621-4.

[74] Liang ZY, Lai HY, Yang H, Zhang CJ, Yang H, Wei HH*, et al.* Pro54DB: a database for experimentally verified sigma-54 promoters. Bioinformatics. 2017;33(3):467-9.

[75] Chen W, Zhang X, Brooker J, Lin H, Zhang L, Chou K-C. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics. 2014;31(1):119-20.

[76] Tang W, Wan S, Yang Z, Teschendorff AE, Zou Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics. 2018;34(3):398-406.

[77] Jia C, Zuo Y, Zou Q. O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. Bioinformatics. 2018;34(12):2029-36.

DISCLAIMER: The above article has been published in Epub (ahead of print) on the basis of the materials provided by the author. The Editorial Department reserves the right to make minor modifications for further improvement of the manuscript.