METHODS
published: 30 July 2020
doi: 10.3389/fgene.2020.00760

EnACP: An Ensemble Learning Model for Identification of Anticancer Peptides

Ruiquan Ge[1], *Guanwen Feng*[2], *Xiaoyang Jing*[3], *Renfeng Zhang*[4], *Pu Wang*[5]* and *Qing Wu*[1]*

[1] Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China, [2] Xi'an Key Laboratory of Big Data and Intelligent Vision, School of Computer Science and Technology, Xidian University, Xi'an, China, [3] Toyota Technological Institute at Chicago, Chicago, IL, United States, [4] Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, China, [5] Computer School, Hubei University of Arts and Science, Xiangyang, China

OPEN ACCESS

Edited by:
Yungang Xu,
University of Texas Health Science Center at Houston, United States

Reviewed by:
Sotiris Kotsiantis,
University of Patras, Greece
Tao Huang,
Shanghai Institute for Biological Sciences (CAS), China

***Correspondence:**
Pu Wang
nywangpu@yeah.net
Qing Wu
wuqing@hdu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: 20 April 2020
Accepted: 26 June 2020
Published: 30 July 2020

Citation:
Ge R, Feng G, Jing X, Zhang R, Wang P and Wu Q (2020) EnACP: An Ensemble Learning Model for Identification of Anticancer Peptides. Front. Genet. 11:760. doi: 10.3389/fgene.2020.00760

As cancer remains one of the main threats of human life, developing efficient cancer treatments is urgent. Anticancer peptides, which could overcome the significant side effects and poor results of traditional cancer treatments, have become a new potential alternative these years. However, identifying anticancer peptides by experimental methods is time consuming and resource consuming, it is of great significance to develop effective computational tools to quickly and accurately identify potential anticancer peptides from amino acid sequences. For most current computational methods, feature representation plays a key role in their final successes. This study proposes a novel fast and accurate approach to identify anticancer peptides using diversified feature representations and ensemble learning method. For the feature representations, the information is encoded from multidimensional feature spaces, including sequence composition, sequence-order, physicochemical properties, etc. In order to better model the potential relationships of peptides, multiple ensemble classifiers, LightGBMs, are applied to detect the different feature sets at first. Then the obtained multiple outputs are used as inputs of the support vector machine classifier, which effectively identifies anticancer peptides. Experimental results on cross validation and independent test sets demonstrate that our method can achieve better or comparable performances compared with other state-of-the-art methods.

Keywords: anticancer peptides, feature representation, ensemble learning, pseudo amino acid composition, system biology

INTRODUCTION

Cancer has become a common disease in humans, and it often leads to a higher mortality rate, especially in developing and developed countries (Ortega-Garcia et al., 2020). The complexity and heterogeneity of cancer are major obstacles for anticancer therapy development (Kasak and Laan, 2020; Umbreit et al., 2020). Traditional cancer treatments, such as radiation therapy, targeted therapy and chemotherapy, often fail to distinguish cancer cells from normal cells. Traditional surgery could not guarantee the precise removal of the diseased part, which is seriously harmful to the patient's body (An et al., 2019). At the same time, the risk of recurrence after surgery is high. In addition, cancer cells have developed resistance to traditional anticancer drugs due to their overuse. Overall, traditional treatment methods have obvious side effects and poor results. In view of these problems, there is an urgent to discover and design novel cancer treatments and anticancer agents to fight against this deadly disease (Esfandiari Mazandaran et al., 2019; Sima et al., 2019; Bahuguna et al., 2020).

In recent years, peptide-based therapy has become a potential method of cancer treatments. This method can target and kill cancer cells while do not impair the normal cells (Harris, 2020). Anticancer peptides (ACPs) with short amino acid sequences can avoid the disadvantages of traditional cancer treatments. They generally have the characteristics of high specificity, high tissue penetration, low production cost, toxic under normal physiological functions, ease of synthesis and modification, etc. And natural ACPs are safer than synthetic drugs (Feng and Wang, 2019). The electrostatic interactions between ACPs and cancer cell membranes are considered to be one of the main factors for the selective killing of cancer cells (Lin et al., 2018; Naguib et al., 2018). They are believed to play a vital role in the selective toxicity of ACPs to cancer. Currently, many approved peptide-based drugs are being evaluated in various stages of clinical trials (Tesauro et al., 2019; Brunetti et al., 2020). As more and more ACPs are identified and verified by experiments, it is found that most ACPs are derived from protein sequences (Tyagi et al., 2013). However, the discovery of novel ACPs from wet-lab experimentation is laborious, time-consuming and expensive. So, it is essential to develop efficient computational methods to rapidly identify potential ACPs from the peptide sequences.

In the past decade, the accurate identification of ACPs from peptide sequences remains an open research topic in the field of bioinformatics and immunoinformatic. Machine learning methods have been widely used to identify ACPs in many researches. It mainly includes two key techniques which are feature representation and classifier. For feature representation, if the features of peptide sequences are well-extracted, it will be easier to precisely predict the ACPs (Jing et al., 2019). At present, some tools in the prediction of ACPs have been developed. The first computational tool is called Anti-CP (Tyagi et al., 2013), which encoded peptides with sequence-based features and binary profiles to predict ACPs based on Support Vector Machine (SVM). In another work, Hajisharifi et al. considered two kinds features from the local correlation and Chou's pseudo acid amino composition (PseAAC) to improve the prediction of ACPs (Hajisharifi et al., 2014). ACPP used an improved feature encoding method via three type of protein relatedness measure, integrating compositional information, centroidal and distributional information of amino acids (Vijayakumar and Lakshmi, 2015). iACP has referred that membrane interactions are related to their conformation or the order of amino acids. And, it can get better results through cross validation and optimizing the g-gap dipeptide components method compared to the previous predictors (Chen W. et al., 2016). Li et al. indicated that the different types of feature combinations can improve the prediction for ACPs (Chen W. et al., 2016). MLACP constructed features using amino acid composition, atomic composition, dipeptide composition, and physicochemical properties and developed SVM and random forest (RF) methods to predict ACPs (Manavalan et al., 2017). SAP employed 400D features with g-gap dipeptide information and feature selection to identify ACPs (Xu et al., 2018). ACPred-FL can orderly extract effective features from sequence-based feature and a group of SVM models (Wei et al., 2018). mACPpred explored seven feature encodings and a two-step feature selection method to exclude irrelevant features (Ge et al., 2016; Boopathi et al., 2019). Then, the obtained features are input into SVM classifier to gain the predicted result. In addition, a special repository named CancerPPD was collected and created with the manually verified ACPs from the published literature, patents and other databases (Tyagi et al., 2015). It provides a wealth of information related to the peptide for research and experimental personnel to use for reference such as its origin, the nature of the peptide, anticancer activity, terminal modification, conformation, etc. The information is helpful to understand the comprehensive properties of ACPs. And it also provides a reference for the design and identification of ACPs (Lin et al., 2015).

In this paper, we propose a novel two-step prediction model EnACP to accurately identify the ACPs. At first, feature representation is composed of four categories: amino acid composition, autocorrelation, pseudo amino acid composition and profile-based features (Chen et al., 2018). Each type includes a few modes. Finally, 19 kinds of feature patterns are generated. For each feature pattern, LightGBM (Light Gradient Boosting Machine) classifier is employed to generate the initial prediction (Ke et al., 2017). The former predicted results as the new features are input to SVM classifier to get the final prediction. Cross validation results showed that the proposed EnACP model performed better than the previous methods. Furthermore, EnACP achieved comparable performances compared with the existing methods on a new independent dataset. EnACP is available at https://github.com/greyspring/EnACP.

MATERIALS AND METHODS

Dataset

In this study, we use two groups of ACP datasets from the existed literatures to evaluate the performance of the proposed method. For them, one dataset is used to test the cross-validation performance compared with the existing models (Hajisharifi et al., 2014). The other with an independent test dataset can better measure the generalization capability of the model (Boopathi et al., 2019).

For the two datasets, one is called ZH dataset including 138 ACPs and 206 non-ACPs for the 5-fold cross-validation test. The other is from mACPpred for the independent test. In mACPpred dataset, the training dataset consists of 266 ACPs and 266 non-ACPs, and the independent dataset consists of 157 ACPs and 157 non-ACPs. The two group datasets have the low redundancy which were processed to prevent homology bias and high similarity in the related literatures. Amino acid frequency distribution of ACP and non-ACP in the two datasets are shown in **Figure 1**. The sequences containing not 20 natural amino acids are eliminated. From **Figure 2**, most of the peptide sequences are between 5 and 50 in length in the

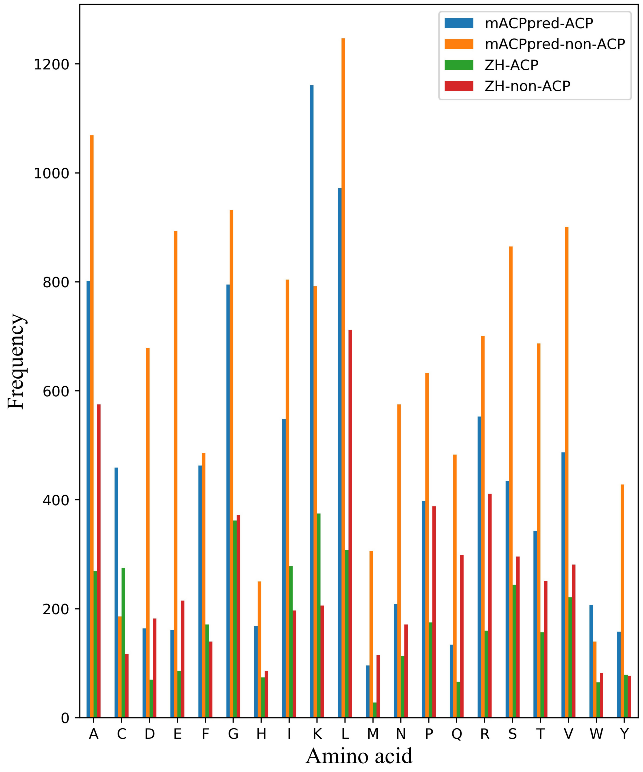

FIGURE 1 | Amino acid frequency distribution on cross-validation and independent datasets. The number of 20 amino acids are counted in mACPpred and ZH datasets. The horizontal axis represents the abbreviation of 20 amino acids. The ordinate represents the number of amino acids.

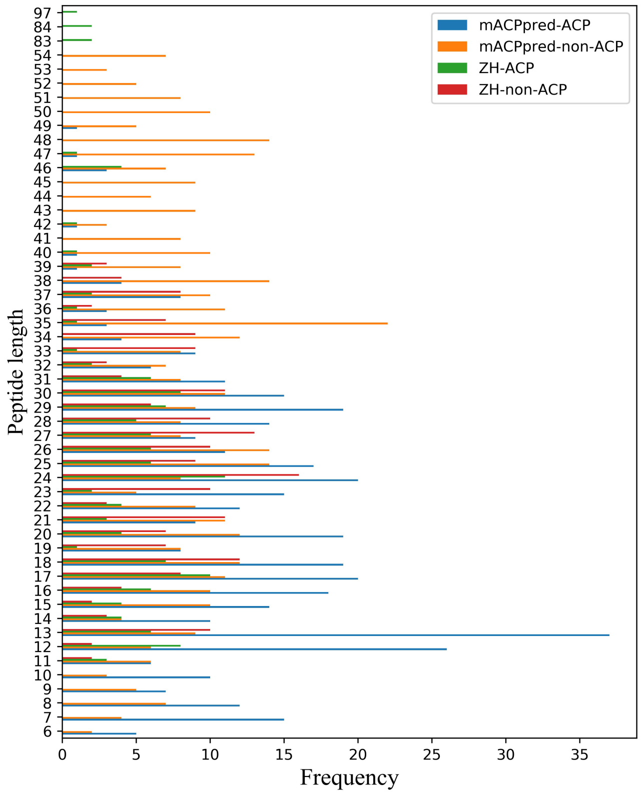

FIGURE 2 | Peptide length distribution of ACP and non-ACP on mACPpred and ZH datasets. The horizontal axis represents the number of statistics. The ordinate represents the length of the peptide sequence.

two datasets especially in mACPpred-ACP and ZH-non-ACP. For mACPpred-non-ACP and ZH-ACP, their ratio is 94.6 and 96.4%, respectively.

Features Representation

There are 19 kinds of features in total used in this study, three of which belong to amino acid composition, four of which belong to autocorrelation features, four of which belong to pseudo amino acid composition, and eight of which belong to profile-based features (Liu et al., 2015, 2017; Liu, 2019).

Amino Acid Composition

Basic kmer (Kmer) (Liu et al., 2008) is a very simple feature extraction method that represents any peptide sequence as a vector consisted of occurrence frequencies of k neighboring amino acids. Distance-based Residue (DR) (Liu et al., 2014b) extracts features from sequence by counting the occurrence frequencies of all possible residue pairs within a certain distance. Just like the DR method, the method of Distance-Pairs and reduced alphabet scheme (Distance Pair) (Liu et al., 2014a) also extracts features from sequence by counting the occurrence frequencies of residue pairs within a certain distance, except that the residue types are reduced by clustering.

Autocorrelation Features

A peptide sequence P is often formulated in the following format, with the N-terminus at the left, and the C-terminus at the right.

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

where R_1 represents the 1st amino acid, R_2 represents the 2nd amino acid, and so forth.

Given a physicochemical index of amino acids, The Auto covariance (AC) (Cao et al., 2013) approach measures the correlation between two residues separated by distance d, which can be calculated as:

$$AC(u,d) = \sum_{i=1}^{L-d} \left(I_u(R_i) - \bar{I}_u\right)\left(I_u(R_{i+d}) - \bar{I}_u\right)/(L-d)$$

where u indicates the physicochemical index, $I_u(R_i)$ means the index value of R_i, and $\bar{I}_u$ is the average index value along the whole sequence:

$$\bar{I}_u = \sum_{i=1}^{L} I_u(R_i)/L$$

The Cross covariance (CC) (Cao et al., 2013) approach measures the correlation between two residues separated by distance d based on two different physicochemical indices, which can be calculated by:

$$\mathrm{CC}(u,v,d)=\sum_{i=1}^{L-d}\left(I_u(R_i)-\bar{I}_u\right)\left(I_v(R_{i+d})-\bar{I}_v\right)/\left(L-d\right)$$

where u and v indicate two different indices, $I_u(R_i)(I_v(R_i))$ means the index value of R_i, and $\bar{I}_u(\bar{I}_v)$ is the average index value along the whole sequence.

Auto-cross covariance (ACC) (Cao et al., 2013) is the combination of AC and CC. Physicochemical distance transformation (PDT) (Liu et al., 2012) is a sequence-based method, in which any peptide sequence is firstly encoded as a series of numbers by amino acid index (AAindex) (Kawashima et al., 2008), and then a fixed length vector is extracted through distance transformation.

Pseudo Amino Acid Composition

Parallel correlation pseudo amino acid composition (PC-PseAAC) (Chou, 2001) is an approach that takes the sequence-order information into account and represents any peptide sequence as:

$$\mathrm{P}=[x_1\ \ x_2\ \ x_3\ \ \cdots\ \ x_{20}\ \ x_{20+1}\ \ \cdots\ \ x_{20+\lambda}]$$

where

$$x_u=\begin{cases}\dfrac{f_u}{\sum_{i=1}^{20}f_i+w\sum_{j=1}^{\lambda}\theta_j} & (1\le u\le 20)\\[2ex] \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20}f_i+w\sum_{j=1}^{\lambda}\theta_j} & (20+1\le u\le 20+\lambda)\end{cases}$$

where $f_i(i=1,2,\ldots,20)$ is the occurrence frequency of the 20 native amino acids in the peptide; the integer λ represents the highest tier of correlation along the sequence; w is the weight factor ranging from 0 to 1; $\theta_j(j{=}1,2,\ldots,\lambda)$ is the j-tier correlation factor that is defined as:

$$\theta_j=\frac{1}{L-j}\sum_{i=1}^{L-j}\Theta\left(R_i,R_{i+j}\right)\qquad\left(1\le j\le\lambda\right)$$

Where the correlation function is given by

$$\Theta\left(R_i,R_j\right)=\frac{1}{3}\left\{\left[H_1(R_i)-H_1(R_j)\right]^2+\left[H_2(R_i)-H_2(R_j)\right]^2+\left[M(R_i)-M(R_j)\right]^2\right\}$$

where $H^1(R_i)$, $H^2(R_i)$, and $M(R_i)$ are the standardized hydrophobicity value, hydrophilicity value, and side-chain mass of R_i, respectively.

Series correlation pseudo amino acid composition (SC-PseAAC) (Chou, 2005) is a variant of PC-PseAAC that represents any peptide sequence as:

$$\mathrm{P}=[x_1\ \ \cdots\ \ x_{20}\ \ x_{20+1}\ \ \cdots\ \ x_{20+\lambda}\ \ x_{20+\lambda+1}\ \ \cdots\ \ x_{20+2\lambda}]$$

where

$$x_u=\begin{cases}\dfrac{f_u}{\sum_{i=1}^{20}f_i+w\sum_{j=1}^{2\lambda}\theta_j} & (1\le u\le 20)\\[2ex] \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20}f_i+w\sum_{j=1}^{2\lambda}\theta_j} & (20+1\le u\le 20+2\lambda)\end{cases}$$

where $f_i(i{=}1,2,\ldots,20)$ is the occurrence frequency of the 20 native amino acids in the peptide; the integer λ represents the highest tier of correlation along the sequence; w is the weight factor ranging from 0 to 1; $\theta_j(j{=}1,\ 2,\ \ldots,\ 2\lambda)$ is the j-tier correlation factor that is defined as:

$$\begin{cases}\theta_1=\frac{1}{L-1}\sum_{i=1}^{L-1}H_{i,i+1}^1\\ \theta_2=\frac{1}{L-1}\sum_{i=1}^{L-1}H_{i,i+1}^2\\ \cdots\\ \theta_{2\lambda-1}=\frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}H_{i,i+\lambda}^1\\ \theta_{2\lambda}=\frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}H_{i,i+\lambda}^2\end{cases}$$

where the correlation functions are given by

$$\begin{cases}H_{i,j}^1=h^1\left(R_i\right)\cdot h^1\left(R_j\right)\\ H_{i,j}^2=h^2\left(R_i\right)\cdot h^2\left(R_j\right)\end{cases}$$

where $h^1(R_i)$ and $h^2(R_i)$ are the standardized hydrophobicity and hydrophilicity values of R_i, respectively.

General parallel correlation pseudo amino acid composition (PC-PseAAC-General) is an enhanced version of PC-PseAAC, in which both the built-in indices extracted from AAindex and the indices provided by users can be incorporated. General series correlation pseudo amino acid composition (SC-PseAAC-General) is an enhanced version of SC-PseAAC, in which both the built-in indices extracted from AAindex and the indices provided by users can be incorporated.

Profile-Based Features

The Top-n-gram (Liu et al., 2014b) approach extracts evolutionary information from the frequency profiles calculated from the multiple sequence alignments outputted by PSI-BLAST (Altschul et al., 1997), and any peptide sequence is represented as a fixed dimension feature vector by counting the occurrence times of each Top-n-gram. Profile-based physicochemical distance transformation (PDT-Profile) is similar with PDT except that the features are extracted from frequency profiles. Distance-based Top-n-gram (DT) extends the original Top-n-gram approach by considering the relative position information of Top-n-gram pairs in peptide sequences, and the feature vector of peptide sequence was calculated by counting the occurrences of all possible Top-n-gram pairs within a certain distance threshold.

Profile-based Auto covariance (AC-PSSM) (Dong et al., 2009) transforms the PSSM of a peptide into fixed-length vector, in which the AC variable measures the correlation of the same property between two residues separated by a distance. Profile-based Cross covariance (CC-PSSM) (Dong et al., 2009) transforms the PSSM of a peptide into fixed-length vector, in which the CC variables measure the correlation of two different properties between two residues separated by a distance. Profile-based Auto-cross covariance (ACC-PSSM) (Dong et al., 2009) represents any peptide sequence as a feature vector consisting of ACC variables that are the combination of AC variables and CC variables. PSSM distance transformation (PSSM-DT) (Xu et al., 2015) extracts features from the PSSM of a peptide which measure the occurrence probabilities of any amino acid pairs separated by a distance. PSSM relation transformation (PSSM-RT) (Zhou et al., 2017) extracts features from the PSSM of a peptide by utilizing the relationships of evolutionary information between residues.

Support Vector Machine and LightGBM

In this study, the dataset has exactly two class labels: anticancer peptides (positive) and non-anticancer peptides (negative). Support vector machines (SVMs) are very suitable for binary classification, and because of the strong generalization ability for small datasets, they are used extensively in biomedical data mining (Chen et al., 2019; Jiang et al., 2020). SVM classifies data by finding the best hyperplane to separate all data points of one class from these of another class. The best hyperplane of SVM is the hyperplane with the largest margin between two classes. SVM is firstly proposed for linearly separable data, and when the data are non-separable, the kernel functions such as radial basis function can be used.

LightGBM (Light Gradient Boosting Machine) is a distributed gradient lifting framework based on decision tree algorithm proposed by Microsoft in 2017 (Ke et al., 2017). In order to shorten the computation time, LightGBM as a good ensemble learning algorithm was designed for two main reasons (Xia et al., 2017). For one thing, it can reduce the use of memory and the communication cost, improves the efficiency when multiple machines are parallel. For another thing, it designs and implements a good strategy for feature selection.

Methodology

To develop an accurate predictor of ACPs, we present a two-step ensemble learning method called EnACP. The framework of the model is shown in **Figure 3**. In the first step, 19 feature encodings of the peptide sequences are extracted in terms of amino acid composition, autocorrelation, pseudo amino acid composition and profile-based features as descripted in section Features Representation. For each group of feature encodings, the initial prediction is obtained separately using an ensemble learning classifier LightGBM. In this way, the complex higher-dimensional features are dispersed to lower dimensions. Then, the outputs of all LightGBMs as combinative nineteen-dimensional feature vector are input into an optimized SVM classifier to capture the hidden relationships. At last, the peptide sequence is identified whether it is ACP or non-ACP.

For a given binary classification problem about a set of sequences $Q(s)$, the class labels $C=\{C_1, C_2, \ldots, C_s\}$, $C_i \epsilon \{0, 1\}$, and each sample q_i has k group features $<F_1(q_i), F_2(q_i),\ldots F_k(q_i)>$, where F_j is the j^{th} group features. Each group has several related features. Firstly, all the features are generated by the 19 kinds of feature representation algorithm for all the sequences. For the train dataset, LightGBM is employed to classify each group features, respectively. The LightGBM classification results of k group features are input SVM to train the model. For the test dataset, the inputs are generated according to the first layer model of the train data set. Finally, ACPs or non-ACPs are identified for the test peptide sequences. The algorithm flow is described in the following pseudocode.

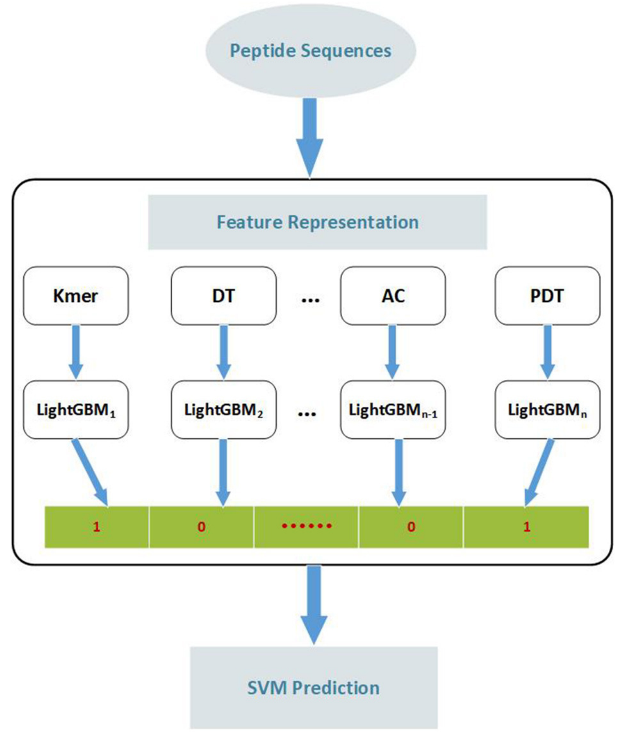

FIGURE 3 | The flow diagram of identifying anticancer peptides. Peptide sequences are represented by 19 feature extraction methods. According to the features obtained by each extraction method, LightGBM is employed to classify the peptide sequences. Then, the outputs of all LightGBMs are input into SVM classifier to predict the peptide sequence as ACP or non-ACP. Kmer: subsequences of length K contained within a peptide sequence; DT, Distance-based Top-n-gram; AC, Auto Covariance; PDT, Physicochemical Distance Transformation; SVM, Support Vector Machine.

As shown from the pseudocode, there are three factors that affect the time complexity of the model EnACP, such as feature extraction, LightGBM and SVM algorithms. Let p and n be the numbers of the most features $F_i(q_i)$ and train samples Q_t, respectively. And the length of the longest sequence is l. Different feature extraction methods are relatively independent, and they can be generated in parallel. So, the most complex feature extraction method determines the time complexity of the feature extraction stage. For the 19 groups of feature extraction methods, the profile-based method with the highest complexity is O(n^*l^3). LightGBM is implemented using three technologies to improve

```
Algorithm: EnACP

Input: a sequences set $Q$: $(q_i,C_i)$, $k$ groups of feature types, class label $C_i=\{0, 1\}$, $q_i$ is a peptide sequence, $Q_t$ is train dataset, $Q_v$ is test dataset.

Begin
for each sequence $q_i$ in $Q$:
        // Initialize all features of $q_i$, each $F_i(q_i)$ represent one group of features
    $F(q_i)=<F_1(q_i), F_2(q_i),\ldots, F_k(q_i)>=\{\}$
        // Initialize second level features
    $L2FK(q_i)[1..k]=\{\}$
        // Feature extract
    for $j = 1$ to $k$
        Generate features $F_j(q_i)$ according to feature representation algorithm $F_j$
    endfor
endfor
for train dataset $Q_t$: $(q_t,C_t)$
    for $m = 1$ to $k$
            // Classify the sequences $Q_t$ in the first level
        $L1Model_m = LightGBM(F(q_t), C_t)$
        $L2FK(q_t)[m]= L1Model_m(F(q_t)-)$
    endfor
        // Train the model in the second level
    $L2Model=SVM (L2FK(q_t) [1..k], C_t)$
endfor
for test dataset $Q_v$: $(q_v,C_v)$
    for $n = 1$ to $k$
            // Classify the sequences Qv in the first level
        $L2FK(q_v) [n]= L1Modeln(F(q_v)-)$
    endfor
        // Predict the peptide sequence $q_v$: ACP or non-ACP
    $FinalPredict(q_v)= L2Model(L2FK(q_v) [1..k])$
endfor
End
```

the model efficiency: gradient-based one-side sampling, exclusive feature bundling, and histogram algorithm. These techniques have resulted in more or less a reduction in the number of samples and features. Moreover, it also supports feature parallel and data parallel processing. So, its worst time complexity will not exceed $O(p^{*}\ n)$. And the computational complexity of an SVM is $O(n^3)$ for the training dataset. So the worst-case time complexity of EnACP is $\max(O(n^{*}l^3), O(p^{*}n), O(n^3)$-). But most of the features will usually be excluded in the first layer. Then the SVM algorithm in the second layer will be significantly speeded up. So the actual calculation time will not reach the upper-bound in the train stage. For the test dataset, the time is mainly consumed in the feature extraction stage after the parameters of LightGBM and SVM are optimized.

Evaluation

The metrics for performance evaluation used in our experiments include Receiver Operating Characteristic curve (ROC), Area Under a ROC Curve (AUC), Sensitivity (Sn), Specificity (Sp), Accuracy (Acc), and the Matthews correlation coefficient (MCC) (Plyusnin et al., 2019). Suppose TP, FP, TN and FN are the abbreviations for true positives, false positives, true negatives, and false negatives respectively, then the evaluation metrics can be calculated as:

$$Sp = \frac{TN}{TN + FP}$$

$$Sn = \frac{TP}{TP + FN}$$

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

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

RESULTS

Performance on Different Feature Representations

In order to find the effective feature coding representation of the peptide sequence, four kinds of feature representation methods including 19 feature encodings were extracted in terms of amino acid composition, autocorrelation, pseudo amino acid composition and profile-based features. Referring to the first step of the model, the ACPs were identified by LightGBM classifier using various feature codes, respectively. From the overall results in **Figure 4**, they were ranked by pseudo amino acid composition, amino acid composition, profile-based features and autocorrelation. In terms of the various feature codes, pseudo amino acid composition worked best according to the value of the performance indexes Acc, AUC, Sp, Sn, and MCC. Its MCC was nearly 14 percentage points higher than the second place. And, its Acc, Sn, and Sp were about 7 percentage points higher than the second-place method from amino acid composition. Among them, autocorrelation encoding was the worst, and its performance indexes were all below 80%.

Performance Comparison on Cross-Validation Dataset

To verify the effect of our model, we compared the results of a few popular methods such as Li method (Li and Wang, 2016), ZH method (Hajisharifi et al., 2014) and iACP (Chen W. et al., 2016) on ZH dataset with 5-fold cross-validation. In order to compare the predictive capability, the predicted results of the four methods were showed in **Table 1**. Judging from the result, our predictor EnACP performed better than other three methods and reached the first place in the evaluation indexes on Sn, Acc, and MCC. In all the evaluation indexes, EnACP only lost to iACP in Sp index. Acc, Sn, and MCC of our method were about 0.6 to 5.7%, 2.2 to 7.6%, and 1.7 to 12.6% higher than the predictive results of other methods, respectively. In terms of Sp index, our method was only 0.9% lower than iACP method, but also much higher than other methods. From the discussion above, it can be seen that our method may automatically learn representative features from the numerous feature codes. The two step combined classifiers with LightGBM and SVM may improve the accuracy of prediction and achieve better identification efficiency between ACPs and non-ACPs.

Furthermore, for the stability of the model, 5-fold cross validation experiment was executed 30 times randomly.

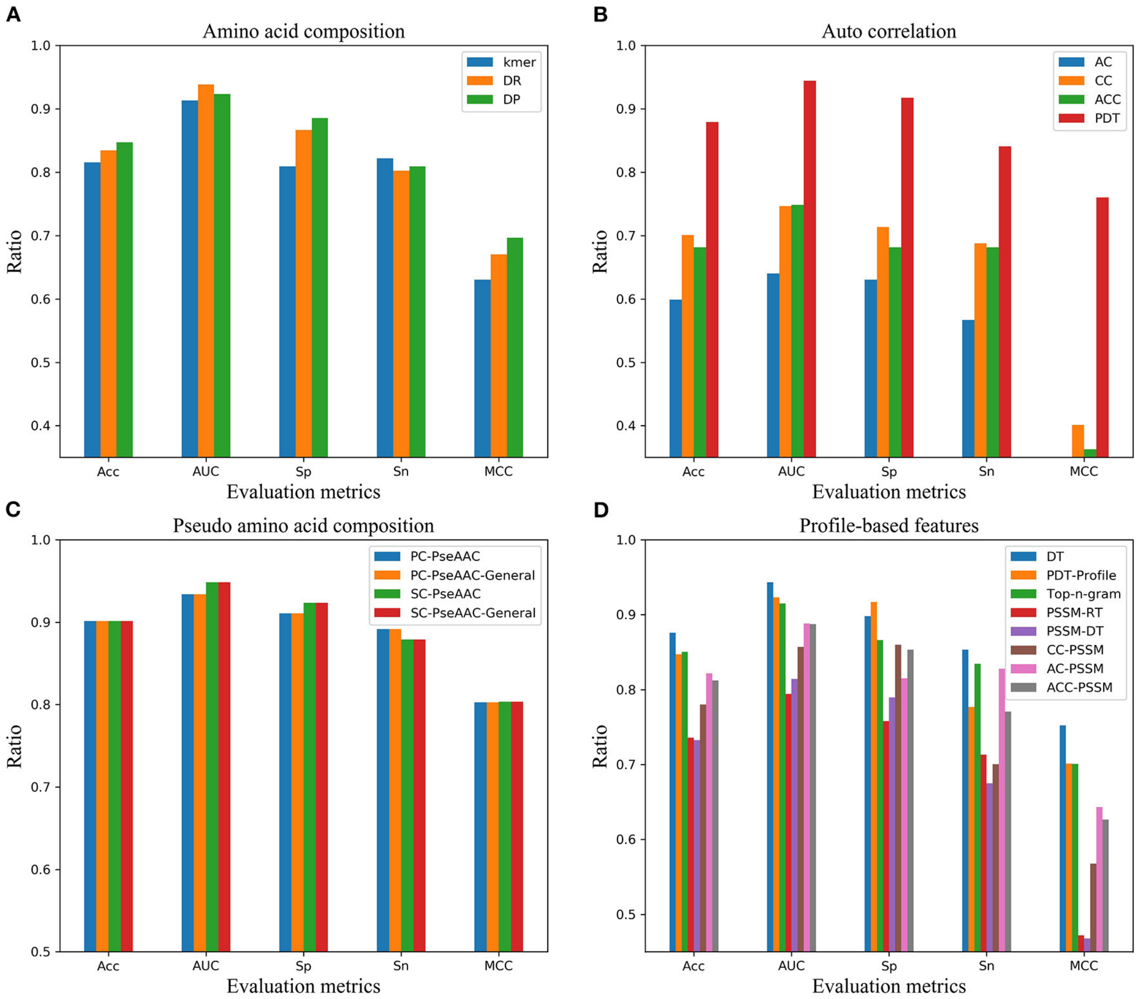

FIGURE 4 | Performance of four kinds of feature encodings on independent dataset in AUC, Sn, Sp, Acc, and MCC. **(A)** Performance of feature encodings based on amino acid composition**, (B)** Performance of feature encodings based on autocorrelation**, (C)** Performance of feature encodings based on pseudo amino acid composition, **(D)** Performance of feature encodings based on profile.

According to the statistical results of various evaluation metrics shown in **Figure 5**, several indicators fluctuate little. And the standard deviation of Acc, MCC, Sn, and Sp is 0.0005, 0.0012,0.0012, and 0.0011, respectively. Therefore, the cross-validation analysis showed the stability and robustness of our model EnACP.

Performance Comparison on Independent Test Datasets

To further verify the power of the current predictor, three independent datasets are analyzed from mACPpred (Boopathi et al., 2019), ACPP (Vijayakumar and Lakshmi, 2015), and Tyagi's paper (Tyagi et al., 2013) named mACP_Ind, ACPP_Ind, and Tyagi_Ind, respectively. For the independent test dataset mACPpred_Ind, SVMACP and RFACP belong to MLACP algorithm based on RF and SVM method, respectively. For this dataset, we refer to the experimental results from the literature mACPpred (**Table 2**). And for the independent test datasets ACPP_Ind and Tyagi_Ind, we compare our algorithms EnACP with mACPpred and iACP (**Table 3**). Experimental results on independent tests show that this proposed EnACP predictor is quite more effective and promising for identification of ACPs compared with the previous methods.

TABLE 1 | Performance comparison of different methods on 5-fold cross-validation dataset.

Methods	Acc	Sn	Sp	MCC
EnACP	0.954	0.928	0.981	0.910
Li method	0.942	0.906	0.967	0.879
ZH method	0.897	0.852	0.927	0.784
iACP	0.948	0.884	0.990	0.893

Compared with mACPpred method, our model EnACP had achieved excellent results, among which, MCC, Acc, Sn, and Sp were all about 2, 1, 0.7, and 1.2% higher, respectively, AUC was

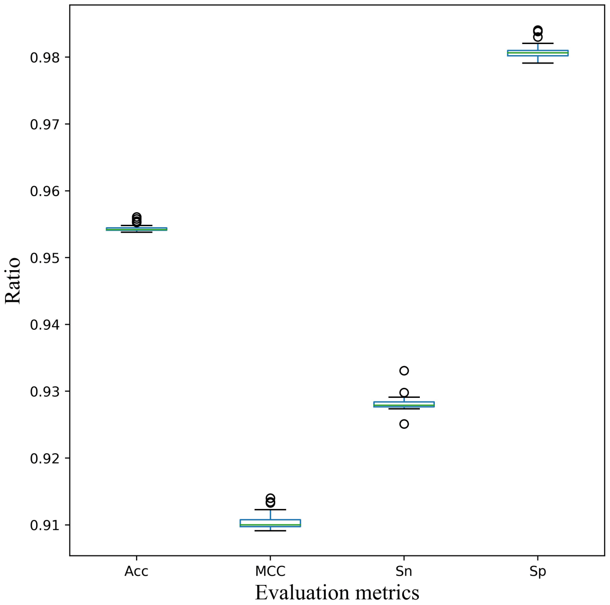

FIGURE 5 | Stability of the EnACP model on 5-fold cross-validation dataset. Five-fold cross validation experiment was executed 30 times randomly. And the metrics of Acc, MCC, Sn and Sp were plotted and analyzed.

TABLE 2 | Performance comparison of different methods on the independent test dataset mACPpred_Ind.

Methods	Acc	Sn	Sp	MCC	AUC
EnACP	0.924	0.892	0.955	0.849	0.968
mACPpred	0.914	0.885	0.943	0.829	0.967
SVMACP	0.768	0.554	0.981	0.592	0.896
RFACP	0.707	0.414	1.000	0.511	0.891
iACP	0.667	0.580	0.753	0.338	0.747

TABLE 3 | Performance comparison of different methods on the independent test datasets ACPP_Ind and Tyagi_Ind.

Datasets	Methods	Acc	Sn	Sp	MCC	AUC
ACPP_Ind	EnACP	0.948	1	0.9	0.901	0.992
	mACPpred	0.948	0.973	0.925	0.898	0.989
	iACP	0.74	0.919	0.575	0.558	0.875
Tyagi_Ind	EnACP	0.853	1	0.708	0.739	0.996
	mACPpred	0.884	0.957	0.813	0.777	0.948
	iACP	0.8	0.894	0.708	0.612	0.905

TABLE 4 | Pairwise comparison of ROC curves in three datasets.

Datasets	P(A, B)	EnACP	mACPpred	iACP
mACP_Ind	EnACP	—	0.9705	<0.0001
	mACPpred	—	—	<0.0001
	iACP	—	—	—
ACPP_Ind	EnACP	—	0.6612	0.0036
	mACPpred	—	—	0.0076
	iACP	—	—	—
Tyagi_Ind	EnACP	—	0.0384	0.0015
	mACPpred	—	—	0.2381
	iACP	—	—	—

The comparison P(A, B) is defined the statistical significance P-value of ROC curves between algorithm A and algorithm B.

TABLE 5 | The comparison triplets between algorithm pairs from EnACP, mACPpred and iACP.

T(A,B)	EnACP	mACPpred	iACP
EnACP	—	1/2/0	3/0/0
mACPpred	0/2/1	—	2/1/0
iACP	0/0/3	0/1/2	—

The comparison triplet T(A, B) is defined to be the numbers of the three datasets where algorithm A performs better, equally well and worse, compared with algorithm B in terms of P-value.

basically flat. MCC, Acc, Sn, and AUC obtained from our model EnACP were about 25.7 to 51.1%, 15.6 to 25.7%, 31.2 to 47.8%, 7.2 to 22.1% higher, respectively, compared with SVMACP, RFACP, and iACP. Additionally, it can also be seen from the results of **Figure 4** and **Table 2** that the EnACP method has an advantage over the pseudo amino acid composition method with one step prediction. Sn is only slightly lower less than a percentage point. And, MCC, Sp, Acc, and AUC obtained from EnACP model were about 4, 4, 2, 2% higher than the pseudo amino acid composition method with one step prediction. For ACPP_Ind and Tyagi_Ind datasets, EnACP achieves the similar performance advantages on AUC and Sn.

The statistical significance is evaluated using rank-based ROC curves comparison to determine whether EnACP performs better than, similarly to or worse than the other algorithms (DeLong et al., 1988; Hanley and Hajian-Tilaki, 1997). The results are shown in the following **Table 4**. For a confidence level of 0.95, EnACP perform statistically significantly better than iACP on all datasets. EnACP performs similarly or slightly better than mACPperd algorithms on mACP_Ind and ACPP_Ind. And mACPpred performs better than iACP on the previous two datasets. The algorithms EnACP and mACPperd perform better than iACP with statistical significance. The comparison triplets are also statistically tabulated between algorithm pairs from EnACP, mACPpred and iACP which show that one algorithm performs better, equally well and worse, compared with another algorithm in **Table 5**.

Comparison of Different Classification Methods

Based on many previous studies, using SVM classifier for task of peptide classification outperforms most of other classical classifiers such as AdaBoost, decision tree (DT), logistic regression (LR), Naïve Bayes (NB), random forest (RF) (Becker et al., 2011). We also conducted a comparative study on the two datasets and obtained the similar conclusion in the second step of

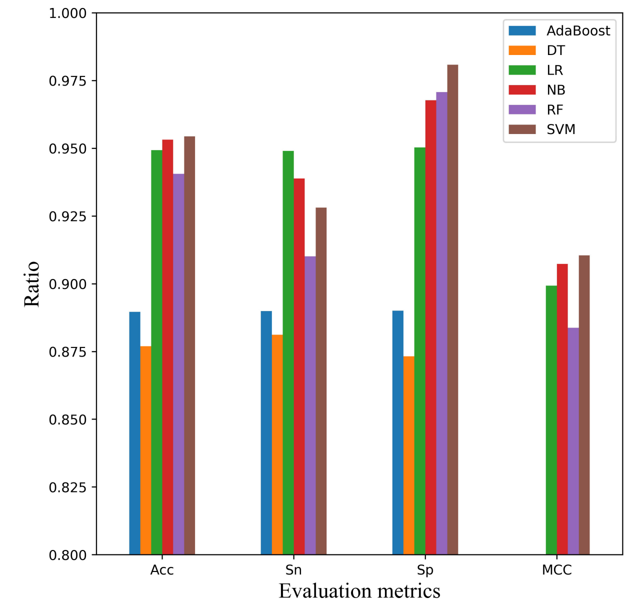

FIGURE 6 | Comparison of SVM with other classifiers on 5-fold cross-validation dataset. Four performance indicators which are Sn, Sp, Acc, and MCC are compared using six classifiers that are AdaBoost, decision tree (DT), logistic regression (LR), Naïve Bayes (NB), random forest (RF), and Support Vector Machine (SVM), respectively.

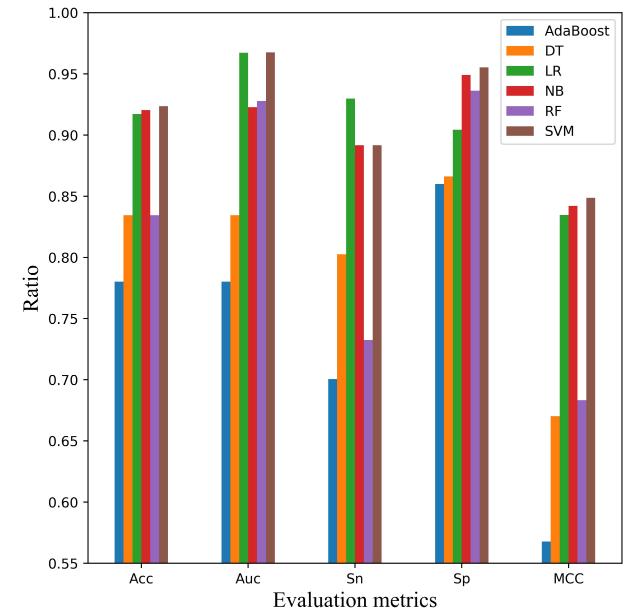

FIGURE 7 | Comparison of SVM with other classifiers on independent test dataset mACPpred_Ind. Five performance indicators which are AUC, Sn, Sp, Acc, and the Matthews correlation coefficient (MCC) are compared using six classifiers that are AdaBoost, decision tree (DT), logistic regression (LR), Naïve Bayes (NB), random forest (RF), and Support Vector Machine (SVM), respectively.

the model EnACP. Experimental results on both the 5-fold cross-validation and independent test showed that SVM, NB and LR were relatively stable, and SVM has the best overall effect.

In order to verify the performance of SVM classifier, we randomly selected scrambled data before 5-fold cross-validation. Finally, the average result of six classifiers were obtained after 30 times of 5-fold cross validation, as shown in **Figure 6**. Each classifier performed well, but in the comprehensive comparison, SVM, LR, and NB classifiers were better. On the whole, SVM classifier worked best. SVM achieved the first place in the three indexes of Acc, MCC, and Sp. For the Sn index, it was only about 1 and 2% lower than the classifier of NB and LR, respectively.

In addition, independent test dataset mACPpred_Ind was used to measure the performance and categorization capabilities of the optimal model in **Figure 7**. Compared with the cross-validation experiment, the AUC evaluation metric was added into this experiment except Acc, Sn, Sp, and MCC. Except for Sn, SVM classifier ranked the first place in Acc, AUC, Sp, and MCC, which was similar to the cross-validation result. But, SVM had better performance relative to cross validation tests. For example, for AUC index, SVM was more than 13 points higher than AdaBoost and DT. For Sp index, SVM is more than 5 points higher than AdaBoost, LR and DT. For MCC, SVM was 16% higher than RF and DT.

DISCUSSION

Even to this day, it is difficult to trace the cause of cancer because of its complex mechanisms. In spite of various treatment strategies, the effect was not ideal. Peptide-based therapy has become a research field of precision medicine. The rapid and accurate identification of ACPs from peptide sequences based on machine learning methods can be better applied to anticancer drug development and other biomedical experiments (Diller et al., 2018).

From the experimental results of the independent test datasets, our model EnACP performs well overall especially the high AUC and sensitivity. The higher the sensitivity is, the better the predicted model of ACPs is. The highly sensitive discovery of anticancer peptides plays an important role in the design of anticancer and anti-tumor synthetic drugs. The innovation of our model mainly includes the following points. The model EnACP is robust and easy to extend. Multi-group feature encodings contain abundant information. For each group of feature encoding, LightGBM as the first layer of EnACP can auto pre-learning and select the key features, respectively. Actually, for the higher-dimensional features, the computation is not very large. Meanwhile, the model implements the multi-layer feature learning strategy. Moreover, the second layer has fewer features and the model is more efficient to identify the ACPs and non-ACPs. The proposed EnACP performs better in identifying whether the peptide sequence is ACP compared with the existing methods. Its accuracy and stability may be attributed to the following reasons.

At first, how to effectively extract the valuable information of ACPs is a major challenge for all the predicted methods. It has been proved that the membrane interaction and insertion of membrane-active peptides could be related to the order of amino acids. Systematic analysis revealed that some physiochemical properties of peptides are not clearly sufficient to predict their selectivity for example net positive charge, hydrophobicity, and hydrophobic moments (Chen W. et al., 2016). Some methods also are developed using amino acid composition and binary profiles as input features (Lin et al., 2015). Therefore, in order to find a suitable feature representation, EnACP extracts 19 kinds of features from four aspects, including amino acid composition, auto correlation, pseudo amino acid composition and profile-based features.

Then, in purpose to accurately identify the ACPs quickly, LightGBM classifier is applied to detect the peptide sequences with the 19 kinds of features. As an ensemble learning method, LightGBM can automatically optimize to achieve dimension reduction and effectively prevent overfitting. On the other hand, it can better discover the relationship of peptides and select the representative feature description from the integrating multiple groups features (Huang et al., 2010). In addition, the secondary structure and tertiary structure prediction characteristics of peptides can be added into this model as a part of basis feature description, which may further improve the performance of the model (Ma et al., 2015). Furthermore, neural network method can also be explored for the identification of ACPs with the increase of datasets (Hashemifar et al., 2018).

Finally, in terms of the used classifiers, many prediction tools have demonstrated the effectiveness of the SVM method. As a two-step prediction model, SVM finally outputs the identified results with grid search to optimize its parameters. Besides, in order to expedite the identification of ACPs, we called LightGBM with the default parameters in the scikit-learn package library. Better model parameters may be obtained by modern optimization methods to improve the prediction performance.

CONCLUSION

In order to effectively identify ACPs from amino acid sequences, a novel hybrid predicted model EnACP is proposed in this paper. EnACP involves two-step strategy based on ensemble learning method. Firstly, multi-type and multi-group feature descriptions were constructed based on amino acid composition, autocorrelation, pseudo amino acid composition and profile-based features. In purpose to find a suitable feature representation and accurately classify quickly, the ensemble classifier LightGBM was applied to detect the peptide sequences. Secondly, multiple groups of results from the output of LightGBMs were integrated as the input of SVM model to enhance the final prediction accuracy of ACP as well as non-ACP. To validate the performance of EnACP, two group experiments were performed on cross validate dataset and independent dataset. The experimental results indicated that the proposed EnACP model achieved competitive performance on some performance metrics. On the other hand, our model can be used to solve other protein sequence problems, such as homologous detection of proteins (Chen J. et al., 2016), prediction of various sites (Chou and Shen, 2008, 2010), prediction of protein-protein interaction (Wang et al., 2019), etc.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The datasets can be found in: https://github.com/greyspring/EnACP/tree/master/datasets.

AUTHOR CONTRIBUTIONS

RG and PW designed the method. RG and GF developed the prediction models. GF, XJ, RZ, and QW analyzed the data and results. All authors have read and approved the revised manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China (Nos. 61702146, 61841104, 81602490), National key research and development program (No. 2019YFC0118404), Shandong province key research and development projects (No. 2017GSF218067), Zhejiang Postdoctoral Foundation (No. zj20180025) and China Scholarship Council (No. 201808330081).

ACKNOWLEDGMENTS

RG sincerely thanks Professor Jinbo Xu and his group from Toyota Technological Institute at Chicago (TTIC) for their creative discussions and valuable suggestions.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

An, Z., Flores-Borja, F., Irshad, S., Deng, J., and Ng, T. (2019). Pleiotropic role and bidirectional immunomodulation of innate lymphoid cells in cancer. *Front. Immunol.* 10:3111. doi: 10.3389/fimmu.2019.03111

Bahuguna, A., Singh, A., Kumar, P., Dhasmana, D., Krishnan, V., and Garg, N. (2020). Bisindolemethane derivatives as highly potent anticancer agents: synthesis, medicinal activity evaluation, cell-based compound discovery, and computational target predictions. *Comput. Biol. Med.* 116:103574. doi: 10.1016/j.compbiomed.2019.103574

Becker, N., Toedt, G., Lichter, P., and Benner, A. (2011). Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data. *BMC Bioinform* 12:138. doi: 10.1186/1471-2105-12-138

Boopathi, V., Subramaniyam, S., Malik, A., Lee, G., Manavalan, B., and Yang, D. C. (2019). mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides. *Int. J. Mol. Sci.* 20:1964. doi: 10.3390/ijms20081964

Brunetti, J., Piantini, S., Fragai, M., Scali, S., Cipriani, G., Depau, L., et al. (2020). A new NT4 peptide-based drug delivery system for cancer treatment. *Molecules* 25:1088. doi: 10.3390/molecules25051088

Cao, D. S., Xu, Q. S., and Liang, Y. Z. (2013). propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* 29, 960–962. doi: 10.1093/bioinformatics/btt072

Chen, J., Guo, M., Wang, X., and Liu, B. (2018). A comprehensive review and comparison of different computational methods for protein remote homology detection. *Brief. Bioinform.* 19, 231–244. doi: 10.1093/bib/bbw108

Chen, J., Liu, B., and Huang, D. (2016). Protein Remote Homology Detection Based on an Ensemble Learning Approach. *Biomed Res. Int.* 2016, 5813645. doi: 10.1155/2016/5813645

Chen, T., Zhang, C., Liu, Y., Zhao, Y., Lin, D., Hu, Y., et al. (2019). A gastric cancer LncRNAs model for MSI and survival prediction based on support vector machine. *BMC Genomics* 20:846. doi: 10.1186/s12864-019-6135-x

Chen, W., Ding, H., Feng, P., Lin, H., and Chou, K. C. (2016). iACP: a sequence-based tool for identifying anticancer peptides. *Oncotarget* 7, 16895–16909. doi: 10.18632/oncotarget.7815

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43, 246–255. doi: 10.1002/prot.1035

Chou, K. C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 21, 10–19. doi: 10.1093/bioinformatics/bth466

Chou, K. C., and Shen, H. B. (2008). Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* 3, 153–162. doi: 10.1038/nprot.2007.494

Chou, K. C., and Shen, H. B. (2010). A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PLoS ONE* 5:e9931. doi: 10.1371/journal.pone.0009931

DeLong, E. R., Delong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837–845. doi: 10.2307/2531595

Diller, K. I., Bayden, A. S., Audie, J., and Diller, D. J. (2018). PeptideNavigator: an interactive tool for exploring large and complex data sets generated during peptide-based drug design projects. *Comput. Biol. Med.* 92, 176–187. doi: 10.1016/j.compbiomed.2017.11.016

Dong, Q., Zhou, S., and Guan, J. (2009). A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics* 25, 2655–2662. doi: 10.1093/bioinformatics/btp500

Esfandiari Mazandaran, K., Mirshokraee, S. A., Didehban, K., and Houshdar Tehrani, M. H. (2019). Design, synthesis and biological evaluation of ciprofloxacin- peptide conjugates as anticancer agents. *Iran. J. Pharm. Res.* 18, 1823–1830. doi: 10.22037/ijpr.2019.111721.13319

Feng, P., and Wang, Z. (2019). Recent advances in computational methods for identifying anticancer peptides. *Curr. Drug Targets* 20, 481–487. doi: 10.2174/1389450119666180801121548

Ge, R., Zhou, M., Luo, Y., Meng, Q., Mai, G., Ma, D., et al. (2016). McTwo: a two-step feature selection algorithm based on maximal information coefficient. *BMC Bioinform.* 17:142. doi: 10.1186/s12859-016-0990-0

Hajisharifi, Z., Piryaiee, M., Beigi, M. M., Behbahani, M., and Mohabatkar, H. (2014). Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *J. Theor. Biol.* 341, 34–40. doi: 10.1016/j.jtbi.2013.08.037

Hanley, J. A., and Hajian-Tilaki, K. O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. *Acad. Radiol.* 4, 49–58. doi: 10.1016/S1076-6332(97)80161-4

Harris, A. L. (2020). Development of cancer metabolism as a therapeutic target: new pathways, patient studies, stratification and combination therapy. *Br. J. Cancer* 122, 1–3. doi: 10.1038/s41416-019-0666-4

Hashemifar, S., Neyshabur, B., Khan, A. A., and Xu, J. (2018). Predicting protein-protein interactions through sequence-based deep learning. *Bioinformatics* 34, i802–10. doi: 10.1093/bioinformatics/bty573

Huang, T., Shi, X. H., Wang, P., He, Z., Feng, K. Y., Hu, L., et al. (2010). Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. *PLoS One* 5:e10972. doi: 10.1371/journal.pone.0010972

Jiang, H., Gu, J., Du, J., Qi, X., Qian, C., and Fei, B. (2020). A 21gene Support Vector Machine classifier and a 10gene risk score system constructed for patients with gastric cancer. *Mol. Med. Rep.* 21, 347–359. doi: 10.3892/mmr.2019.10841

Jing, X., Dong, Q., Hong, D. C., and Lu, R. (2019). Amino acid encoding methods for protein sequences: a comprehensive review and assessment. *IEEE/ACM Trans Comput Biol Bioinform.* doi: 10.1109/TCBB.2019.2911677. [Epub ahead of print].

Kasak, L., and Laan, M. (2020). Monogenic causes of non-obstructive azoospermia: challenges, established knowledge, limitations and perspectives. *Hum Genet.* doi: 10.1007/s00439-020-02112-y. [Epub ahead of print].

Kawashima, S., Pokarowski, P., Pokarowsika, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 36, D202–D205. doi: 10.1093/nar/gkm998

Ke, G. L., Meng, Q., Finley, T., Wang, T. F., Chen, W., Ma, W. D., et al. (2017). "LightGBM: a highly efficient gradient boosting decision tree," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3149–3157.

Li, F. M., and Wang, X. Q. (2016). Identifying anticancer peptides by using improved hybrid compositions. *Sci. Rep.* 6:33910. doi: 10.1038/srep33910

Lin, M. W., Tseng, Y. W., Shen, C. C., Hsu, M. N., Hwu, J. R., Chang, C. W., et al. (2018). Synthetic switch-based baculovirus for transgene expression control and selective killing of hepatocellular carcinoma cells. *Nucleic Acids Res.* 46:e93. doi: 10.1093/nar/gky447

Lin, Y. C., Lim, Y. F., Russo, E., Schneider, P., Bolliger, L., Edenharter, A., et al. (2015). Multidimensional design of anticancer peptides. *Angew. Chem. Int. Ed Engl.* 54, 10370–10374. doi: 10.1002/anie.201504018

Liu, B. (2019). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* 20, 1280–1294. doi: 10.1093/bib/bbx165

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K. C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 43, W65–W71. doi: 10.1093/nar/gkv458

Liu, B., Wang, X., Chen, Q., Dong, Q., and Lan, X. (2012). Using amino acid physicochemical distance transformation for fast protein remote homology detection. *PLoS ONE* 7:e46633. doi: 10.1371/journal.pone.0046633

Liu, B., Wang, X. L., Lin, L., Dong, Q. W., and Wang, X. (2008). A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis. *BMC Bioinformatics* 9:510. doi: 10.1186/1471-2105-9-510

Liu, B., Wu, H., and Chou, K.-C. (2017). Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nat. Sci.* 9, 67–91. doi: 10.4236/ns.2017.94007

Liu, B., Xu, J., Lan, X., Xu, R., Zhou, J., Wang, X., et al. (2014a). iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *PLoS ONE* 9:e106691. doi: 10.1371/journal.pone.0106691

Liu, B., Xu, J., Zou, Q., Xu, R., Wang, X., and Chen, Q. (2014b). Using distances between Top-n-gram and residue pairs for protein remote homology detection. *BMC Bioinform.* 15(Suppl. 2):S3. doi: 10.1186/1471-2105-15-S2-S3

Ma, J., Wang, S., Wang, Z., and Xu, J. (2015). Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics* 31, 3506–3513. doi: 10.1093/bioinformatics/btv472

Manavalan, B., Basith, S., Shin, T. H., Choi, S., Kim, M. O., and Lee, G. (2017). MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget* 8, 77121–77136. doi: 10.18632/oncotarget.20365

Naguib, A., Mathew, G., Reczek, C. R., Watrud, K., Ambrico, A., Herzka, T., et al. (2018). Mitochondrial complex I inhibitors expose a vulnerability for selective killing of Pten-null cells. *Cell Rep.* 23, 58–67. doi: 10.1016/j.celrep.2018.03.032

Ortega-Garcia, M. B., Mesa, A., Moya, E. L. J., Rueda, B., Lopez-Ordono, G., Garcia, J. A., et al. (2020). Uncovering tumour heterogeneity through PKR and nc886 analysis in metastatic colon cancer patients treated with 5-FU-based chemotherapy. *Cancers* 12:379. doi: 10.3390/cancers12020379

Plyusnin, I., Holm, L., and Toronen, P. (2019). Novel comparison of evaluation metrics for gene ontology classifiers reveals drastic performance differences. *PLoS Comput. Biol.* 15:e1007419. doi: 10.1371/journal.pcbi.1007419

Sima, P., Richter, J., and Vetvicka, V. (2019). Glucans as new anticancer agents. *Anticancer Res.* 39, 3373–3378. doi: 10.21873/anticanres.13480

Tesauro, D., Accardo, A., Diaferia, C., Milano, V., Guillon, J., Ronga, L., et al. (2019). Peptide-based drug-delivery systems in biotechnological applications: recent advances and perspectives. *Molecules* 24:351. doi: 10.3390/molecules24020351

Tyagi, A., Kapoor, P., Kumar, R., Chaudhary, K., Gautam, A., and Raghava, G. P. (2013). In silico models for designing and discovering novel anticancer peptides. *Sci. Rep.* 3:2984. doi: 10.1038/srep02984

Tyagi, A., Tuknait, A., Anand, P., Gupta, S., Sharma, M., Mathur, D., et al. (2015). CancerPPD: a database of anticancer peptides and proteins. *Nucleic Acids Res.* 43, D837–D843. doi: 10.1093/nar/gku892

Umbreit, N. T., Zhang, C. Z., Lynch, L. D., Blaine, L. J., Cheng, A. M., Tourdot, R., et al. (2020). Mechanisms generating cancer genome complexity from a single cell division error. *Science* 368:aba0712. doi: 10.1126/science.aba0712

Vijayakumar, S., and Lakshmi, P. T. V. (2015). ACPP: a web server for prediction and design of anti-cancer peptides. *Int. J. Pept. Res. Ther.* 21, 99–106. doi: 10.1007/s10989-014-9435-7

Wang, X., Yu, B., Ma, A., Chen, C., Liu, B., and Ma, Q. (2019). Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics* 35, 2395–2402. doi: 10.1093/bioinformatics/bty995

Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018). ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016. doi: 10.1093/bioinformatics/bty451

Xia, J., Peng, Z., Qi, D., Mu, H., and Yang, J. (2017). An ensemble approach to protein fold classification by integration of template-based assignment and support vector machine classifier. *Bioinformatics* 33, 863–870. doi: 10.1093/bioinformatics/btw768

Xu, L., Liang, G., Wang, L., and Liao, C. (2018). A novel hybrid sequence-based model for identifying anticancer peptides. *Genes (Basel)* 9:158. doi: 10.3390/genes9030158

Xu, R., Zhou, J., Wang, H., He, Y., Wang, X., and Liu, B. (2015). Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. *BMC Syst. Biol.* 9(Suppl. 1):S10. doi: 10.1186/1752-0509-9-S1-S10

Zhou, J., Lu, Q., Xu, R., He, Y., and Wang, H. (2017). EL_PSSM-RT: DNA-binding residue prediction by integrating ensemble learning with PSSM relation transformation. *BMC Bioinformatics* 18:379. doi: 10.1186/s12859-017-1792-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.