Recent Advances in Computational Methods for Identifying Anticancer Peptides

Pengmian Feng¹,* and Zhenyi Wang²

¹School of Public Health, North China University of Science and Technology, Tangshan, 063000, China; ²Center for Genomics and Computational Biology, School of Life Science, North China University of Science and Technology, Tangshan, 063000, China

Abstract: Anticancer peptide (ACP) is a kind of small peptides that can kill cancer cells without damaging normal cells. In recent years, ACP has been pre-clinically used for cancer treatment. Therefore, accurate identification of ACPs will promote their clinical applications. In contrast to labor-intensive experimental techniques, a series of computational methods have been proposed for identifying ACPs. In this review, we briefly summarized the current progress in computational identification of ACPs. The challenges and future perspectives in developing reliable methods for identification of ACPs were also discussed. We anticipate that this review could provide novel insights into future researches on anticancer peptides.

Keywords: Anticancer peptides, disease, cancer, drug target, machine learning methods, sequence encoding scheme.

1. INTRODUCTION

Cancer is one of the most common causes of high morbidity and mortality [1-3]. Due to their adverse effects on normal cells [4, 5], the conventional radiation therapy and chemotherapy are not effective for cancer treatment. Therefore, there is an urgent need of potential new drugs for cancer treatment [6].

In recent years, researchers have discovered a kind of short peptide, called anticancer peptide (ACP), which exhibits the ability to kill cancer cells, destroys primary tumors, and prevents metastasis without damaging normal cells [7-9]. Therefore, ACP has been receiving the attention of the scientific community [10-14]. ACP is naturally occurring biologics and usually contains 5-30 amino acid residues [15]. Owing to their advantages such as short time-frame of interaction, low toxicity, high tissue penetration, ease of modifications, mode of action, specificity and good solubility [16], ACPs have been selected as one of the alternative candidates for cancer therapy [17-20]. Accordingly, ACPs have been pre-clinically used for cancer treatment [21]. However, the clinical usage of ACPs is still under development and under observation.

In order to promote its clinical application, it is necessary to distinguish ACP from other peptides. Although experimental techniques can identify ACPs, they are still cost-ineffective and time-consuming [8]. The development of machine learning approaches provides us with an opportunity to computationally identify biological molecules [22-44]. Therefore, it is urgent to develop computational methods to identify potential ACPs. Consequently, a series of machine learning based computational methods have been proposed for identification of ACPs [25, 45-49] and design of membranolytic ACPs [50] in the past several years. The framework of these existing machine learning methods for identifying ACPs is shown in (Fig. 1), which obeys the 5-step rule [51] used to establish many practically very useful predictors [52-60].

In this review, we will summarize the representative computational approaches developed for the identification of ACPs. Current challenges facing the computational prediction of ACPs and future perspectives will be discussed as well.

2. BENCHMARK DATASET

2.1. Database

CancerPPD is the first database hosts ACPs [61], and is available at http://crdd.osdd.net/raghava/cancerppd/. At present, the CancerPPD consists of 3491 entries of ACP covering 249 types of cancer cell lines. Besides the peptide information, CancerPPD also provides the predicted tertiary structures for each deposited ACP [61]. Since it has been developed, CancerPPD has become a useful resource for scientists who concerned on ACP.

2.2. Benchmark Dataset

Constructing a high-quality benchmark dataset is the first and key step for developing computational methods [51]. By
searching the antimicrobial peptides database and academic publications, Hajisharifi and his colleagues [46] obtained 138 ACPS (positive samples). Since ACPS are naturally secretory peptides [62], the non-anticancer peptides (nACP) were obtained by selecting the non-secretory peptides from the Universal Protein Resource [63]. By doing so, 206 nACPS (negative samples) with sequence similarity less than 90% were obtained. Accordingly, a benchmark dataset $S$ including 138 ACPS and 206 nACPS was built, which has been widely used to train and test the computational models for identifying ACP.

Fig. 1). The framework for identifying ACPS using machine learning methods.

In order to objectively evaluate different computational methods, based on Tyagi et al.’s dataset [47] and the CancerPDD database [61], Chen et al. constructed an independent dataset $S_T$ that includes 150 ACPS and 150 nACPS [45]. None of the peptides in the independent dataset occurred in the benchmark dataset $S$. The sequence similarity in the independent dataset is also less than 90%.

3. SEQUENCE REPRESENTING SCHEME

The second key step for developing computational methods is how to encode the protein/peptide sequences using an effective method [51, 64]. Since the peptide sequences are with different length, they couldn’t be directly recognized by machine learning methods [65-67]. It’s necessary to use sequence encoding schemes to convert the sequences into discrete vectors. Hence, the methods that have been used to represent the ACPS were briefly introduced in this section.

For a $L$-th peptide as given by the following

$$R_1R_2\cdots R_l \cdots R_L$$

(1)

The most straightforward method to encode the peptide sequence is using the $k$-tuple amino acid composition [68-72]. By doing so the peptide will be converted to a discrete vector

$$P_1 = [f_1, f_2, \cdots, f_l, \cdots, f_{20k}]^T$$

(2)

where $T$ is the transpose operator, $f_i$ is the frequency of the $i$-th $k$-tuple amino acid in the peptide and is defined as

$$f_i = \frac{N_i}{L-k+1}$$

(3)

$N_i$ is the frequency of the $i$-th $k$-tuple amino acid occurred in the peptide.

As indicated in Eq.2, the dimension of the vector would be $20^k$. In order to include the long-range correlation information in the vector, $k$ should be a much higher value. It is obvious that the dimension of the vector will increase with the increment of $k$ as well. When $k=3$, the vector dimension will be 8000 which is dramatically greater than the number of peptide samples in the benchmark dataset. Accordingly, the high-dimension disaster problem will appear which often decreases the predictive accuracy [51].

In order to deal with such a problem, the pseudo amino acid composition [73, 74], the g-gap dipeptide composition [75, 76] and the reduced amino acid alphabet composition [74, 77] methods have been proposed.

3.1. Pseudo Amino Acid Composition (PseAAC)

The pseudo amino acid composition (PseAAC) is a widely used sequence encoding scheme in computational proteomics [78-81]. By adding the physicochemical properties into the amino acid composition, PseAAC can include the long-range correlation of two residues. The two types of PseAAC, namely type I and type II PseAAC, are defined as following [73].

(i) Type I PseAAC

If a peptide is encoded using the type I PseAAC, it will be converted to a $(20+\lambda)$ dimensional vector defined by

$$P_2 = [f_1, f_2, \cdots, f_{20}, f_{21}, \cdots, f_{20+\lambda}]^T$$

(4)

where

$$f_u' = \left\{ \begin{array}{ll} f_u & 1 \leq u \leq 20 \\ \frac{w \theta_j}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j} & 21 \leq u \leq 20 + \lambda \\ \end{array} \right.$$

(5)

where $f_u$ is the frequency of the 20 amino acid, $w$ is weight factor, $\lambda$ reflects the rank of correlation. And $\theta_j$ is the $j$-tier sequence correlation factor, which reflects the long-range correlation effect and is calculated by following equation:

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

(6)

where $R_i, R_j$ is the correlation function and is given by:

$$\theta(R_i, R_j) = \frac{1}{k} \sum_{l=1}^{k} (H_l(R_i) - H_l(R_j))^2$$

(7)

where $k$ is the number of physicochemical properties and $H_l(R_j)$ is the $l$-th normalized physicochemical properties of the residue $R_i$.

(ii) Type II PseAAC

If a peptide is encoded using the type II PseAAC, it will be converted to a $(20+\lambda)$ dimensional vector defined as

$$P_3 = [f_1, f_2, \cdots, f_{20}, f_{21}, \cdots, f_{20+\lambda}]^T$$

(8)

where
omics has been demonstrated in a series of recent studies defined, in which the 20 native amino acids were grouped by Lin representative residues which is called the reduced amino naive amino acids can be clustered into a smaller number of

3.3. Reduced Amino Acid Alphabet Composition

peptide length. $g=0$ indicates the correlation of two proximate residues. Its effectiveness has been
tide composition, GDC can describe the long-range correlations between two residues. Its effectiveness has been

According to different physiochemical properties, the 20 native amino acids can be clustered into a smaller number of representative residues which is called the reduced amino acid alphabet (RAAA). Based on the Protein Blocks proposed by de Brevern et al., a novel type of RAAA has been defined, in which the 20 native amino acids were grouped into five different cluster profiles (see ref. [85] for more details). The effectiveness of RAAA in computational proteomics has been demonstrated in a series of recent studies [86-89].

By using the RAAA, a peptide sequence can be encoded by the following vector

$$P_a = \left[f_{c1}, f_{h1}, f_{n1}, f_{o1}, f_{g1}\right]^T$$

where $f_i^n$ is the occurrence frequency of the $i$-th $n$-peptide RAAA defined as

$$f_i^n = \frac{n_i^n}{L-n+1}$$

$N_i^n$ is the number of the $i$-th $n$-peptide RAAA in the $L$-length sequence. $\Omega$ is the dimension of the vector and its value depends on $n$ and the cluster profiles (see ref. [88] for more details).

3.4. Structural and Physicochemical Property Composition (SPPC)

Besides the above mentioned sequence-based encoding methods, the structural and physicochemical property based method has also been used to encoding the peptides for identifying ACP [25].

The 20 natural amino acids are made up of five types of atoms, namely C, H, N, O, and S [90]. By calculating the frequency of each atom presents in a given peptide, the sequence could be converted to a 5-dimensional discrete vector

$$P_e = \left[f_C, f_H, f_N, f_O, f_S\right]^T$$

where the elements $(f_C, f_H, f_N, f_O, f_S)$ in the vector are the frequency of each atom in the peptide and equals to the number of each atom in the peptide divided by their total numbers in the peptide length. This vector reflects the atomic composition of a peptide.

In terms of physicochemical property, the 20 natural amino acids can be clustered into different groups. D, E, R, K, Q and N are acidic amino acid residues, C, V, L, I, M, F and W are hydrophobic amino acid residues, D, E, K, H and R are charged amino acid residues, I, L and V are aliphatic amino acid residues, F, H, W and Y are aromatic amino acid residues, H, K and R are positively charged amino acid residues, D and E are negatively charged amino acid residues, A, C, D, G, S and T are tiny amino acid residues, E, H, I, L, K, M, N, P and Q are small amino acid residues, F, R, W and Y are large amino acid residues.

By using the percentage composition of the above mentioned ten kinds of physicochemical properties and together with the peptide mass, a peptide sequence can be represented by an 11 dimensional vector.

$$P_7 = \left[f_1^{sppc}, f_2^{sppc}, \ldots, f_i^{sppc}, \ldots, f_{10}^{sppc}, f_{11}^{sppc}\right]^T$$ (15)

The first 10 elements is the percentage composition of physicochemical properties, and the left one is the peptide mass.

4. METHODS FOR IDENTIFYING ACP

In 2014, Hajisharifi et al. proposed the first computational method for identifying ACPs [46]. By using the following six kinds of physicochemical properties (hydrophobicity, hydrophilicity, side chain mass, pK of the $\alpha$-COOH group, pK of the $\alpha$-NH$_3$ group and pI at 25°C), the peptide sequences in the benchmark dataset $S$ were converted to a 21 dimensional vector by using the type I PseAAC (Eq.5-Eq.7).
The vector thus obtained was then fed into the support vector machine (SVM) for identifying ACPs. In the 5 fold cross validation test, the proposed method obtained an accuracy of 83.82%.

Later on, Tyagi and his colleagues also proposed an SVM-based method for identifying ACPs [47]. Different from Hajisharifi et al.’s work [46], they constructed new benchmark datasets: (1) the main dataset including 225 experimentally validated ACPs and 2250 nACPs; (2) the alternate dataset including 225 experimentally validated ACPs and 1372 nACPs (antimicrobial peptides without anticancer activities); (3) two balanced datasets including 225 ACPs and 225 nACPs fetched from the main dataset and the alternate dataset; (4) the independent dataset including 50 ACPs from literatures and patents and 50 random peptides from the SwissProt proteins, none of which is identical to the peptides in the other datasets. By encoding these peptides using amino acid composition, they obtained the highest accuracies of 92.65%, 75.70%, 88.89%, 87.73 and 86.00% for identifying the ACPs in the above-mentioned datasets, respectively. Since user-friendly web servers represent the future direction for developing practically more useful predictors [91, 92], based on this model, a user-friendly web-server called AntiCP has been developed, which is available at http://crdd.osdd.net/raghava/anticp/index.html. However, it should be pointed out that the peptides in their datasets share high-sequence similarities. For example, some of the peptides in the main dataset have a sequence similarity >90%.

Inspired by these two works, Chen et al. proposed a new bioinformatics model called iACP for identifying anticancer peptides [45]. In their model, the peptide sequences were encoded by using the g-gap (g=1) dipeptide composition scheme. In order to improve the predictive accuracy, the ANOVA (analysis of variance) procedure was carried out to select the optimal features that were further fed to SVM to perform the predictions. In the most objective jackknife test [93-97], iACP obtained an accuracy of 95.06% for identifying anticancer peptides in the benchmark dataset S. For the convenience of the scientific community, an online webserver was developed for iACP, which can be freely accessed at http://lin.uestc.edu.cn/server/iACP.

Since the promising performance of ensemble classification methods has been proved in bioinformatics [98-100], Akbar et al. proposed a genetic algorithm-based ensemble classification method for identifying anticancer peptides [48]. By using the g-gap dipeptide composition, type II PseAAC and RAAAC sequence encoding schemes, the peptides were represented by a hybrid feature vector with a dimension of 544. And then a genetic algorithm-based ensemble classifier called iACP-GAEnsC was proposed, in which five classification algorithms, namely random forest (RF), k-nearest neighbor (KNN), support vector machine (SVM), generalized neural network, and probabilistic neural network (PNN) were employed to build the model. As a result, iACP-GAEnsC yielded an accuracy of 96.45% for identifying anticancer peptides in the benchmark dataset S in the jackknife test.

More recently, Manavalan and his colleagues developed a machine learning based method called MLACP to identify anticancer peptides [48]. Features, such as amino acid composition, dipeptide composition, atomic composition as well as the physicochemical property composition were combined and added to SVM and RF classifier, respectively. By doing so, two prediction models, namely SVM-based and RF-based model were obtained. In the 10-fold cross validation test, the RF-based model can accurately identify 94.60% of the anticancer peptides in the benchmark dataset S. A user-friendly webserver was also established and could be freely accessible at http://thegleelab.org/MLACP.html.

5. COMPARISON OF EXISTING METHODS

Since AntiCP, iACP and MLACP were validated based on different datasets and different cross-validation method to evaluate the models (jackknife test or 10-fold cross validation test), it is difficult to directly compare their performances. In order to fairly compare these methods, we evaluated the three methods on the independent dataset $S_T$. It was found that AntiCP obtained an accuracy of 66.33%, iACP obtained an accuracy of 92.67% and MLACP obtained an accuracy of 78% for identifying the anticancer peptides in the independent dataset $S_T$. This result indicates that the performance of iACP is the best.

CONCLUSION

ACPs have been regarded as one of the therapeutic agents to treat various cancers. Accurate identification of ACPs will pave the way to understand their functions and then promote their clinical applications. Since the experimental method to identify ACPs is still cost-ineffective, development of computational methods to accurately identify ACP from natural peptides is urgent.

In the past several years, a series of computational methods have been proposed and they indeed provided novel insights for computationally identifying ACPs. As pointed out in [101] and demonstrated in a series of recent publications [102-106], user-friendly and publicly accessible web-servers and database represent the future direction for developing practically more useful prediction methods and computational tools. Therefore, it represents a big step forward that most methods introduced here have their own web-servers well established. However, the following aspects can be considered in future work. (i) The existing methods are trained on a small size dataset. Constructing a reliable database could provide more convenience to most of the scholars [103, 105, 107-110]. Therefore, it is necessary to collect more ACPs from current databases, literature and patents and to construct a new high quality benchmark dataset. (ii) Although different sequence encoding schemes have been proposed to represent APCs in the current studies, few of these studies perform feature selections. Since feature selection can alleviate the interference from noise or irrelevant features so as to improve the performance of the computational model, the feature selection techniques are suggested to select optimal features to represent the peptides.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

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

The authors would like to thank the anonymous reviewers for their constructive suggestions. This work was supported by Foundation of Science and Technology Department of Hebei Province (no. 13277713).

REFERENCES

sequence statistical moments into PseAAC. Analytical Biochem 2018; 550: 109-16.


PMID: 30068270