Current Drug Targets, 2019, *20*, 481-487

REVIEW ARTICLE

Recent Advances in Computational Methods for Identifying Anticancer Peptides

Pengmian Feng[1,*] and Zhenyi Wang[2]

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

ARTICLE HISTORY

Received: May 10, 2018
Revised: May 28, 2018
Accepted: May 28, 2018

DOI:
10.2174/1389450119666180801121548

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

*Address correspondence to this author at the School of Public Health, North China University of Science and Technology, Tangshan, 063000, China; Tel: +86-315-8805582; Fax: +86-315-8805582;
E-mail: fengpengmian@gmail.com

1873-5592/19 $58.00+.00

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 $\mathbb{S}$ including 138 ACPs and 206 nACPs was built, which has been widely used to train and test the computational models for identifying ACP.

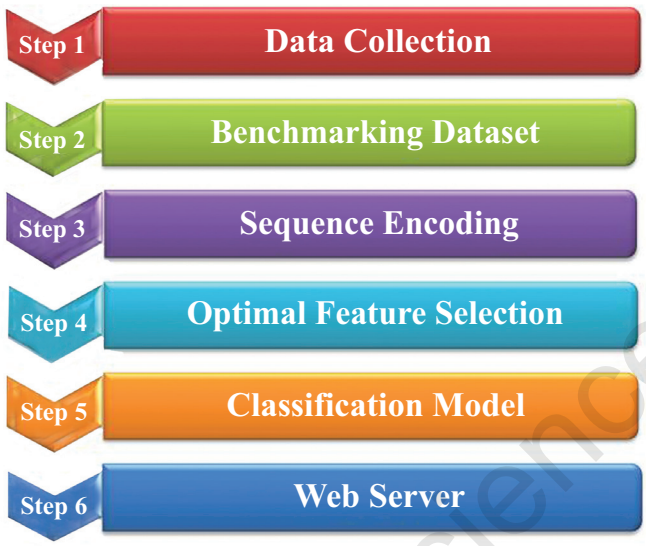

Fig. (1). The framework for identifying ACPs using machine learning methods.

In order to objectively evaluate different computational methods, based on Tyagi *et al.*'s dataset [47] and the CancerPPD database [61], Chen *et al.* constructed an independent dataset $\mathbb{S}_T$ that includes 150 ACPs and 150 nACPs [45]. None of the peptides in the independent dataset occurred in the benchmark dataset $\mathbb{S}$. 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

$$R_1R_2\cdots R_i\cdots R_L \tag{1}$$

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

$$P_1 = [f_1, f_2, \cdots, f_i, \cdots, f_{20^k}]^T \tag{2}$$

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

$$f_i = \frac{N_i}{L-k+1} \tag{3}$$

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+λ) dimensional vector defined by

$$P_2 = [f_1', f_2', \cdots, f_{20}', f_{21}', \cdots, f_{20+\lambda}']^T \tag{4}$$

where

$$f_u' = \begin{cases} \frac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\Theta_j} & 1 \le u \le 20 \\ \frac{w\Theta_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\lambda}\Theta_j} & 21 \le u \le 20+\lambda \end{cases} \tag{5}$$

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

$$\Theta_j = \frac{1}{L-j}\sum_{i=1}^{L-j}\Theta(R_i, R_{i+1}), (j < L) \tag{6}$$

where $\Theta(R_i, R_j)$ is the correlation function and is given by:

$$\Theta(R_i, R_j) = \frac{1}{k}\sum_{l=1}^{k}[H_l(R_j) - H_l(R_i)]^2 \tag{7}$$

where k is the number of physicochemical properties and $H_l(R_i)$ is the l-th normalized physiochemical 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+$n\lambda$) dimensional vector defined as

$$P_3 = [f_1', f_2', \cdots, f_{20}', f_{21}', \cdots, f_{20+n\lambda}']^T \tag{8}$$

where

$$f'_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{n\lambda}\tau_j} & 1 \le u \le 20 \\ \frac{w\Theta_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{n\lambda}\tau_j} & 21 \le u \le 20 + n\lambda \end{cases} \quad (9)$$

where all the terms have exactly the same meanings as those in Eq. (5) except for n and τ_j. The former is number of physiochemical properties considered and the later is defined as:

$$\begin{cases} \tau_1 = \frac{1}{L-1}\sum_{K=1}^{L-1} H_{k,k+1}^1 \\ \tau_2 = \frac{1}{L-1}\sum_{K=1}^{L-1} H_{k,k+1}^2 \\ \vdots \\ \tau_n = \frac{1}{L-1}\sum_{K=1}^{L-1} H_{k,k+1}^n \\ \tau_{n+1} = \frac{1}{L-2}\sum_{K=1}^{L-2} H_{k,k+2}^1 \ (\lambda < L) \\ \tau_{n+2} = \frac{1}{L-2}\sum_{K=1}^{L-2} H_{k,k+2}^2 \\ \vdots \\ \tau_{n+n} = \frac{1}{L-2}\sum_{K=1}^{L-2} H_{k,k+2}^n \\ \vdots \\ \tau_{n\lambda} = \frac{1}{L-\lambda}\sum_{K=1}^{L-\lambda} H_{k,k+\lambda}^n \end{cases} \quad (10)$$

where $H_{k,k+\lambda}^n$ is the correlation function and is given by:

$$H_{k,k+\lambda}^n = h^n(R_k) \cdot h^n(R_{k+\lambda}) \quad (11)$$

where $h^n(R_k)$ is the n-th kind of the physicochemical values of the residue R_k.

3.2. g-gap Dipeptide Composition (GDC)

The g-gap dipeptide composition (GDC) was proposed by Lin *et al.* in 2008 [76]. Compared with the classic dipeptide composition, GDC can describe the long-range correlations between two residues. Its effectiveness has been demonstrated in computational proteomics [82-84]. For a peptide sequence as defined in Eq.1, its GDC can be expressed as

$$P_4 = [f_1^g, f_2^g, \cdots, f_i^g, \cdots, f_{400}^g]^T \quad (12)$$

where f_i^g denotes the frequency of the i-th g-gap dipeptide in the sequence and is defined as

$$d_i^g = \frac{n_i^g}{L-g} \ (g = 0, 1, 2, \cdots) \quad (13)$$

where n_i^g is the number of the i-th g-gap dipeptide, L is the peptide length. g=0 indicates the correlation of two proximate residues; g=1 is the correlation between two residues with one residue interval; and g=2 is the correlation between two residues with the interval of two residues and so forth.

3.3. Reduced Amino Acid Alphabet Composition (RAAAC)

According to different physiochemical properties, the 20 naive 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

$$P_5 = [f_1^n, f_2^n, \cdots, f_i^n, \cdots, f_\Omega^n]^T \quad (14)$$

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

$$f_i^n = \frac{N_i^n}{L-n+1} \quad (15)$$

N_i^n is the number of the i-th n-peptide RAAA in the L-length sequence. Ω 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

$$P_6 = [f_C, f_H, f_N, f_O, f_S]^T \quad (14)$$

where the elements (f_C, f_H, f_N, f_O and 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 polar 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 are 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, Q and V 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.

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

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 α-COOH group, pK of the α-NH 3+ group and pI at 25°C), the peptide sequences in the benchmark dataset $\mathbb{S}$ 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 $\mathbb{S}$. For the convenience of scthe ientific 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 $\mathbb{S}$ 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 $\mathbb{S}$. 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 $\mathbb{S}_T$. 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 $\mathbb{S}_T$. 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. 132777133).

REFERENCES

[1] Torre LA, Bray F, Siegel RL, *et al.* Global cancer statistics, 2012. CA: A Cancer J Clinicians 2015; 65(2): 87-108.

[2] Arnold M, Karim-Kos HE, Coebergh JW, *et al.* Recent trends in incidence of five common cancers in 26 European countries since 1988: Analysis of the European Cancer Observatory. Eur J Cancer 2015; 51(9): 1164-87.

[3] 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.

[4] Al-Benna S, Shai Y, Jacobsen F, Steinstraesser L. Oncolytic activities of host defense peptides. Int J Mol Sci 2011; 12(11): 8027-51.

[5] Kalyanaraman B, Joseph J, Kalivendi S, *et al.* Doxorubicin-induced apoptosis: implications in cardiotoxicity. Mol Cell Biochem 2002; 234-235(1-2): 119-24.

[6] Karpinski TM, Adamczak A. Anticancer activity of bacterial proteins and peptides. Pharmaceutics 2018;10(2).

[7] Vlieghe P, Lisowski V, Martinez J, Khrestchatisky M. Synthetic therapeutic peptides: science and market. Drug Discov Today 2010; 15(1-2): 40-56.

[8] Thundimadathil J. Cancer treatment using peptides: current therapies and future prospects. J Amino Acids 2012; 2012: 967347.

[9] Hoskin DW, Ramamoorthy A. Studies on anticancer activities of antimicrobial peptides. Biochim Biophys Acta 2008; 1778(2): 357-75.

[10] Riedl S, Zweytick D, Lohner K. Membrane-active host defense peptides--challenges and perspectives for the development of novel anticancer drugs. Chem Phys Lipids 2011;164(8): 766-81.

[11] Wu D, Gao Y, Qi Y, *et al.* Peptide-based cancer therapy: opportunity and challenge. Cancer Lett 2014; 351(1): 13-22.

[12] Figueiredo CR, Matsuo AL, Massaoka MH, Polonelli L, Travassos LR. Anti-tumor activities of peptides corresponding to conserved complementary determining regions from different immunoglobulins. Peptides 2014; 59: 14-9.

[13] Gaspar D, Freire JM, Pacheco TR, Barata JT, Castanho MA. Apoptotic human neutrophil peptide-1 anti-tumor activity revealed by cellular biomechanics. Biochim Biophys Acta 2015;1853(2): 308-16.

[14] Huang Y, Feng Q, Yan Q, Hao X, Chen Y. Alpha-helical cationic anticancer peptides: A promising candidate for novel anticancer drugs. Mini Rev Med Chem 2015;15(1): 73-81.

[15] Gaspar D, Veiga AS, Castanho MA. From antimicrobial to anticancer peptides. A review. Front Microbiol 2013; 4: 294.

[16] Ruiz-Torres V, Encinar JA, Herranz-Lopez M, *et al.* An updated review on marine anticancer compounds: The use of virtual screening for the discovery of small-molecule cancer drugs. Molecules 2017; 22(7).

[17] Blunden G. Biologically active compounds from marine organisms. Phytotherapy research : PTR 2001; 15(2): 89-94.

[18] Molina-Guijarro JM, Garcia C, Macias A, *et al.* Elisidepsin interacts directly with glycosylceramides in the plasma membrane of tumor cells to induce necrotic cell death. PloS one 2015; 10(10): e0140782.

[19] Hariharan S, Gustafson D, Holden S, M, *et al.* Assessment of the biological and pharmacological effects of the alpha nu beta3 and alpha nu beta5 integrin receptor antagonist, cilengitide (EMD 121974), in patients with advanced solid tumors. Ann Oncol 2007;18(8):1400-7.

[20] Gregorc V, De Braud FG, De Pas TM, *et al.* Phase I study of NGR-hTNF, a selective vascular targeting agent, in combination with cisplatin in refractory solid tumors. Clin Cancer Res 2011; 17(7): 1964-72.

[21] Boohaker RJ, Lee MW, Vishnubhotla P, Perez JM, Khaled AR. The use of therapeutic peptides to target and to kill cancer cells. Curr Med Chem 2012; 19(22): 3794-804.

[22] Manavalan B, Shin TH, Lee G. DHSpred: Support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. Oncotarget 2018; 9(2): 1944-56.

[23] Manavalan B, Shin TH, Lee G. PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine. Front Microbiol 2018; 9: 476.

[24] Manavalan B, Lee J. SVMQA: support-vector-machine-based protein single-model quality assessment. Bioinformatics 2017; 33(16): 2496-503.

[25] Manavalan B, Basith S, Shin TH, *et al.* MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget 2017; 8(44): 77121-36.

[26] Lin H, Liang ZY, Tang H, Chen W. Identifying sigma70 promoters with novel pseudo nucleotide composition. IEEE/ACM transactions on computational biology and bioinformatics. 2017.

[27] Dao FY, Yang H, Su ZD, *et al.* Recent advances in conotoxin classification by using machine learning methods. Molecules 2017; 22(7).

[28] Cao RZ, Adhikari B, Bhattacharya D, *et al.* QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. Bioinformatics 2017; 33(4): 586-8.

[29] Cao R, Freitas C, Chan L, *et al.* ProLanGO: Protein function prediction using neural machine translation based on a recurrent neural network. Molecules 2017; 22(10).

[30] Tang H, Su ZD, Wei HH, Chen W, Lin H. Prediction of cell-penetrating peptides with feature selection techniques. Biochem Biophysical Res Commun 2016; 477(1): 150-4.

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

[32] Cao RZ, Bhattacharya D, Hou J, Cheng JL, Deep QA: Improving the estimation of single protein model quality with deep belief networks. BMC Bioinformatics 2016; 17: 495

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

[34] Cao R, Wang Z, Wang Y, Cheng J. SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. BMC bioinformatics. 2014; 15: 120.

[35] Kang J, Fang Y, Yao P, *et al.* NeuroPP: A Tool for the Prediction of Neuropeptide Precursors Based on Optimal Sequence Composition. Interdiscip Sci 2018.

[36] Li N, Kang J, Jiang L, *et al.* PSBinder: A web service for predicting polystyrene surface-binding peptides. BioMed Res Int 2017; 2017: 5761517.

[37] He B, Kang J, Ru B, *et al.* SABinder: A web service for predicting streptavidin-binding peptides. BioMed Res Int 2016; 2016: 9175143.

[38] Jia C, Lin X, Wang Z. Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition. Int J Mol Sci 2014; 15(6): 10410-23.

[39] Zhang J, Zhao X, Sun P, Ma Z. PSNO: predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC. Int J Mol Sci 2014; 15(7): 11204-19.

[40] Xu Y, Shao XJ, Wu LY, Deng NY, Chou KC. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. Peer J 2013; 1: e171.

[41] Jia J, Liu Z, Xiao X, Liu B, Chou KC. iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget 2016; 7(23): 34558-70.

[42] Qiu WR, Xiao X, Xu ZC, Chou KC. iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. Oncotarget 2016; 7(32): 51270-83.

[43] Liu LM, Xu Y, Chou KC. iPGK-PseAAC: Identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. Med Chem 2017; 13(6): 552-9.

[44] Khan YD, Rasool N, Hussain W, Khan SA, Chou KC. iPhosT-PseAAC: Identify phosphothreonine sites by incorporating

sequence statistical moments into PseAAC. Analytical Biochem 2018; 550: 109-16.
[45] Chen W, Ding H, Feng P, Lin H, Chou KC. iACP: A sequence-based tool for identifying anticancer peptides. Oncotarget 2016; 7(13): 16895-909.
[46] Hajisharifi Z, Piryaiee M, Mohammad Beigi M, Behbahani M, Mohabatkar H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity *via* Ames test. J Theoretical Biol 2014; 341: 34-40.
[47] Tyagi A, Kapoor P, Kumar R, *et al. In silico* models for designing and discovering novel anticancer peptides. Scientific Reports 2013; 3: 2984.
[48] Akbar S, Hayat M, Iqbal M, Jan MA. iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. Artificial Intelligence Med 2017; 79: 62-70.
[49] Zhang J, Ju Y, Lu H, Xuan P, Zou Q. Accurate identification of cancerlectins through hybrid machine learning technology. Int. J. Genomics 2016, 2016: 7604641.
[50] Grisoni F, Neuhaus C, Gabernet G, *et al.* Designing anticancer peptides by constructive machine learning. Chem Med Chem 2018, 13(13): 1300-02
[51] Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition. J Theoretical Biol 2011; 273(1): 236-47.
[52] Qiu WR, Sun BQ, Xiao X, *et al.* iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics. 2017, 110(5): 239-46.
[53] Qiu WR, Jiang SY, Xu ZC, Xiao X, Chou KC. iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget 2017; 8(25): 41178-88.
[54] Chen W, Feng PM, Deng EZ, Lin H, Chou KC. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Analytical biochemistry. 2014;462:76-83.
[55] Yang H, Qiu WR, Liu GQ, *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.
[56] Chen W, Feng PM, Lin H, Chou KC. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res 2013; 41(6): e68.
[57] Chen W, Lin H, Feng PM, Ding C, Zuo YC, Chou KC. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes *via* physicochemical properties. PloS one. 2012;7(10):e47843.
[58] Cheng X, Zhao SG, Lin WZ, Xiao X, Chou KC. pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. Bioinformatics 2017; 33(22): 3524-31.
[59] Liu B, Yang F, Chou KC. 2L-piRNA: A Two-Layer Ensemble Classifier for Identifying Piwi-Interacting RNAs and Their Function. Mol Ther Nucleic Acids 2017;7: 267-77.
[60] Cheng X, Xiao X, Chou KC. pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. Genomics 2018; 110(1): 50-8.
[61] Tyagi A, Tuknait A, Anand P, *et al.* CancerPPD: A database of anticancer peptides and proteins. Nucleic Acids Res 2015; 43: D837-43.
[62] Mader JS, Hoskin DW. Cationic antimicrobial peptides as novel cytotoxic agents for cancer treatment. Expert Opin Vestigational Drugs 2006; 15(8): 933-46.
[63] UniProt C. Activities at the universal protein resource (UniProt). Nucleic Acids Res 2014; 42: D191-8.
[64] Cao R, Cheng J. Protein single-model quality assessment by feature-based probability density functions. Scientific Reports 2016; 6: 23990.
[65] Feng PM, Ding H, Chen W, Lin H. Naive Bayes classifier with feature selection to identify phage virion proteins. Comput Math Methods Med 2013; 2013: 530696.
[66] Feng PM, Lin H, Chen W. Identification of antioxidants from sequence information using naive Bayes. Comput Math Methods Med. 2013; 2013: 567529.
[67] Zou Q, He W. Special protein molecules computational identification. Int J Mole Sci 2018; 19(2): 536.
[68] Chen W, Lin H. Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine. Comput Biol Med 2012; 42(4): 504-7.
[69] Feng P, Chen W, Lin H. Identifying antioxidant proteins by using optimal dipeptide compositions. Interdiscip Scie 2016; 8(2): 86-91.
[70] Ding H, Deng EZ, Yuan LF, *et al.* iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed Res Int 2014; 2014: 286419.
[71] Wei L, Tang J, Zou Q. SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. Bmc Genomics 2017; 18: 742.
[72] Lai HY, Chen XX, Chen W, Tang H, Lin H. Sequence-based predictive modeling to identify cancerlectins. Oncotarget 2017; 8(17): 28169-75.
[73] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 2001; 43(3): 246-55.
[74] Du P, Gu S, Jiao Y. PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. Int J Mol Sci 2014; 15(3): 3495-506.
[75] Lin H, Chen W, Ding H. AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes. PloS one 2013; 8(10): e75726.
[76] Lin H. The modified mahalanobis discriminant for predicting outer membrane proteins by using chou's pseudo amino acid composition. J Theoretical Biol 2008; 252(2): 350-6.
[77] Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 1999; 291(1): 177-96.
[78] Yang H, Tang H, Chen XX, *et al.* Identification of secretory proteins in mycobacterium tuberculosis using pseudo amino acid composition. BioMed Res Int 2016; 2016: 5413903.
[79] Chen XX, Tang H, Li WC, *et al.* Identification of bacterial cell wall lyases *via* pseudo amino acid composition. BioMed Res Int 2016; 2016: 1654623.
[80] Zhu PP, Li WC, Zhong ZJ, *et al.* Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition. Mol Bio Sys 2015; 11(2): 558-63.
[81] Zhao YW, Su ZD, Yang W, *et al.* IonchanPred 2.0: A tool to predict ion channels and their types. Int J Mol Sci 2017; 18(9): pii: E1838
[82] Lin H, Liu WX, He J, *et al.* Predicting cancerlectins by the optimal g-gap dipeptides. Scientific Reports 2015; 5: 16964.
[83] Ding H, Feng PM, Chen W, Lin H. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Mol Bio Sys 2014;10(8): 2229-35.
[84] Tang H, Zou P, Zhang C, *et al.* Identification of apolipoprotein using feature selection technique. Scientific Reports 2016; 6: 30441.
[85] Etchebest C, Benros C, Bornot A, Camproux AC, de Brevern AG. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. Eur Biophys J 2007; 36(8): 1059-69.
[86] Feng P, Lin H, Chen W, Zuo Y. Predicting the types of J-proteins using clustered amino acids. Bio Med Res Int 2014; 2014: 935719.
[87] Chen W, Feng P, Lin H. Prediction of ketoacyl synthase family using reduced amino acid alphabets. J Ind Microbiol Niotechnol 2012; 39(4): 579-84.
[88] Feng PM, Chen W, Lin H, Chou KC. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. Analytical Biochem 2013; 442(1): 118-25.
[89] Zuo YC, Li QZ. Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet. Peptides 2009; 30(10): 1788-93.
[90] Kumar R, Chaudhary K, Singh Chauhan J, *et al.* An *in silico* platform for predicting, screening and designing of antihypertensive peptides. Scientific Reports 2015; 5: 12512.
[91] Chen W, Feng P, Ding H, Lin H, Chou KC. iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. Analytical Biochem 2015; 490: 26-33.
[92] Chen W, Tang H, Ye J, Lin H, Chou KC. iRNA-PseU: Identifying RNA pseudouridine sites. Mol Ther Nucleic Acids 2016; 5: e332.
[93] Chen W, Yang H, Feng P, Ding H, Lin H. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. Bioinformatics 2017; 33(22): 3518-23.

[94] Chen W, Feng P, Yang H, *et al.* iRNA-3typeA: identifying 3-types of modification at RNA's adenosine sites. Mol Ther Nucleic Acids 2018; 11: 468-74.

[95] Feng P, Yang H, Ding H, *et al.* iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics 2019, 11(1): 96-102.

[96] Chen W, Feng PM, Lin H, Chou KC. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. BioMed Res Int 2014; 2014: 623149.

[97] Feng P, Ding H, Yang H, *et al*. iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. Mol Ther Nucleic Acids 2017; 7: 155-63.

[98] Chen W, Xing P, Zou Q. Detecting N(6)-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines. Scientific Reports 2017; 7: 40242.

[99] Jia C, Zuo Y, Zou Q, Hancock J. 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

[100] Wan S, Duan Y, Zou Q. HPSLPred: An Ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. Proteomics 2017; 17: 1700262.

[101] Chou KC, Shen HB. Recent advances in developing web-servers for predicting protein attributes. Natural Sci 2009; 1: 63-92.

[102] Liu B, Yang F, Huang DS, Chou KC. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics 2018; 34(1): 33-40.

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

[104] Chen W, Zhang X, Brooker J, Lin H, Zhang L, Chou KC. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics 2015; 31(1): 119-20.

[105] Feng P, Ding H, Lin H, Chen W. AOD: the antioxidant protein database. Scientific Reports 2017; 7(1): 7449.

[106] Chen W, Lei TY, Jin DC, Lin H, Chou KC. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Analytical Biochem 2014; 456: 53-60.

[107] He B, Jiang L, Duan Y, *et al.* Biopanning data bank 2018: hugging next generation phage display. Database 2018; 2018.

[108] Dong C, Hao GF, Hua HL, *et al.* Anti-CRISPRdb: a comprehensive online resource for anti-CRISPR proteins. Nucleic Acids Res 2018; 46(D1): D393-D8.

[109] He B, Chai G, Duan Y, *et al.* BDB: biopanning data bank. Nucleic Acids Res 2016; 44(D1): D1127-32.

[110] Huang J, Ru B, Zhu P, *et al.* MimoDB 2.0: a mimotope database and beyond. Nucleic Acids Res 2012; 40: D271-7.

PMID: 30068270