Zhao *et al. BMC Bioinformatics* 2020, **21**(Suppl 16):559
https://doi.org/10.1186/s12859-020-03812-y

BMC Bioinformatics

RESEARCH Open Access

DRACP: a novel method for identification of anticancer peptides

Tianyi Zhao†, Yang Hu† and Tianyi Zang*

From Biological Ontologies and Knowledge bases workshop 2019 San Diego, CA, USA. 18-21 November 2019

*Correspondence:
tianyi.zang@hit.edu.cn
†Tianyi Zhao and Yang Hu have contributed equally to this work.
Department of Computer Science and Technology, School of Life Science and Technology, Harbin Institute of Technology, Harbin, China

Abstract

Background: Millions of people are suffering from cancers, but accurate early diagnosis and effective treatment are still tough for all doctors. Common ways against cancer include surgical operation, radiotherapy and chemotherapy. However, they are all very harmful for patients. Recently, the anticancer peptides (ACPs) have been discovered to be a potential way to treat cancer. Since ACPs are natural biologics, they are safer than other methods. However, the experimental technology is an expensive way to find ACPs so we purpose a new machine learning method to identify the ACPs.

Results: Firstly, we extracted the feature of ACPs in two aspects: sequence and chemical characteristics of amino acids. For sequence, average 20 amino acids composition was extracted. For chemical characteristics, we classified amino acids into six groups based on the patterns of hydrophobic and hydrophilic residues. Then, deep belief network has been used to encode the features of ACPs. Finally, we purposed Random Relevance Vector Machines to identify the true ACPs. We call this method 'DRACP' and tested the performance of it on two independent datasets. Its AUC and AUPR are higher than 0.9 in both datasets.

Conclusion: We developed a novel method named 'DRACP' and compared it with some traditional methods. The cross-validation results showed its effectiveness in identifying ACPs.

Keywords: Anticancer peptides, Deep belief network, Relevance vector machine, Random forest, Cancer

Background

In recent decades, the number of cancer patients has always been increasing. The elder people concerned more on cancers than neurodegenerative diseases. Although the rapid development of medical technology helps a lot, the death rate of patients and burden of the society are still very high. The traditional methods such as radiation therapy [1], targeted therapy [2] and chemotherapy [3] can help suppress cancers, but

apart from the expensive cost, the harm of these treatments to patients are unmeasured [4]. Apparently, finding a unharmful treatment for cancers is critical.

In 1972, antimicrobial peptides' primary structure have been found by Boman [5]. Following his research, many researchers found these peptides have antitumor activity [6, 7]. Therefore, these antimicrobial peptides were named as anticancer peptides (ACPs). ACPs not only have the advantages of high specificity and high tumor penetration, but also easy to synthesis and unharmful to normal cells [8]. This significant advantage makes ACPs become the most potential treatment for cancers [9, 10].

Most of the ACPs are combined from 12–50 amino acid residues. Many of these ACPs' structure are α-helical or β-sheet and some special ACPs have particular folds. They execute their function by interacting with the anionic cell membrane components of cancer cells and then selectively kill cancer cells [11, 12]. Most of the ACPs are obtained from Antimicrobial peptides (AMPs) [13] since cationic AMPs destroy only bacteria but not the normal cells, which shows a broad spectrum cytotoxicity against various cancer cells [14]. Although the mechanism of ACPs is not fully clear at present [15, 16], the development of natural ACPs and artificially designed peptides are still important ways to against cancer.

Due to the high cost of money and time in finding ACPs, increasing number of researchers have focused on identifying the ACPs by computing method. Tyagi et al. [17] extracted amino acid composition and binary profiles as features to build a SVM model to identify ACPs. Later, Khosravian et al. [18] also used SVM to find the ACPs. Then, Hajisharifi et al. [19] used the same method to identify the ACPs, with Chou's pseudo amino acid composition. Besides, Chen et al. [20] purposed a new method named IACP to find ACPs, which has made a great progress. Recently, Manavalan et al. [21] used both Random Forest and SVM to identify the ACPs. Felício et al. [7] reviewed the development of ACPs in 2017 and pointed ACPs decreases the probability of resistance and discussed the relationship between AMP and ACP. Grisoni et al. [22] used long short-term memory (LSTM) to identify ACPs based on sequence.

Although these methods play an important role in the development of this area, there still need more complex algorithm to achieve higher accuracy. Biological networks are common methods to identify biological molecule [23]. In recent years, deep learning algorithms have been widely used in bioinformatics field [24–27]. Deep belief network (DBN) has been proven to be a powerful tool to encode [28]. Therefore, we purposed a novel method named DRACP to identify ACPs. To verify the effectiveness of our method, we used the method on two different datasets. For each, we did cross-validation to do the test to verify the stability.

Results

Data description

The datasets of ACPs was downloaded from Wei Chen et al. [20]. We obtained two datasets. One of them contains 138 real ACPs samples and 206 non-ACPs samples. The other one has 150 real ACPs samples and 150 non-ACPs samples. All the negative samples are randomly generated.

In this paper, 10-cross validation was used to test our method, that is, dividing the whole dataset into 10 groups and one of the groups is used as testing dataset and the rest of groups are used as training dataset.

The performance of DRACP compared with previous method

In this study, the label of pseudo ACPs is 0, and the label of real ACPs is 1.

Firstly, we executed DRACP on the two datasets. The average accuracy of first dataset is 86.87% and the number is 85.17% for the second dataset.

Tyagi et al. [17] developed a method for identifying ACPs based on SVM. We compared our method with their method.

Compared with Tyagi et al. method, we used different features and method. Although different features are used by Tyagi et al., their best performance one is dipeptide composition-based SVM model. However, they ignored the chemical characteristics of amino acids. To test the importance of our feature, we also built a SVM model by using our features. We called this method SVM_{NF}.

The performance these three methods are shown in Table 1. As shown in Table 1, DRACP performed best among these method with the accuracy 0.96 and 0.95. SVM_{NF} ranked second, which means our features are better than Tyagi et al.'s.

The necessity of using DBN

Without using DBN, we put 56-dimension features into Random Relevance Vector Machines (RRVMs) to built the model. Same testing method was used to compare the performance of DRACP and RRVMs. This time, AUC and AUPR are used to evaluate the accuracy of classification.

Figure 1 shows the ROC curves of DRACP and RRVMs. The blue lines denote the ROC curves of DRACP and the red lines denote the ROC curves of RRVMs. The results of dataset2 are represented by dotted lines and solid lines for the results of dataset1. As we can see, DRACP performed much better than RRVMs. Then, we tested the AUPR of these two methods.

Figure 2 shows the PR curves of DRACP and RRVMs. The blue bars denote PR curves of DRACP and red bars denote PR curves of RRVMs. DRACP performed better than RRVMs too.

These experiments showed that using DBN to encode could improve the accuracy of the model.

Table 1 The accuracy of three methods

	Dataset 1	Dataset 2
DRACPs[a]	0.96	0.95
SVM_{NF}[b]	0.92	0.91
Tyagi et al[c]	0.88	0.86
Naive Bayes	0.84	0.81
Random forest	0.89	0.85

[a] The method we purposed

[b] SVM with our feature

[c] Available at https://crdd.osdd.net/raghava/anticp/multi_pep.php

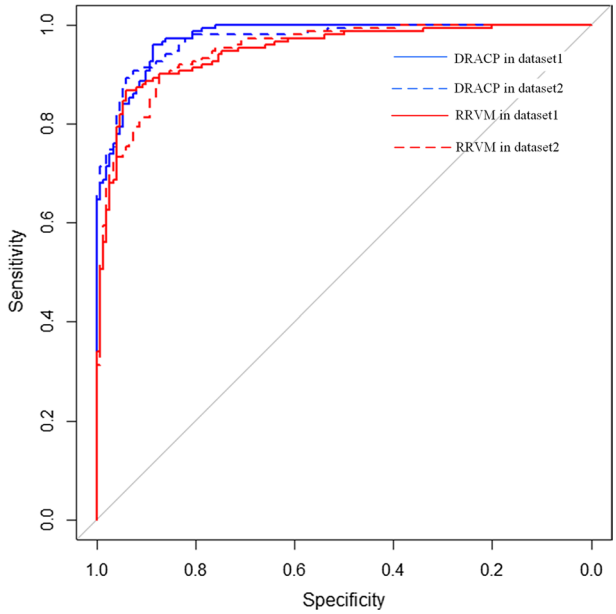

Fig. 1 The ROC curves of DRACP and RRVMs

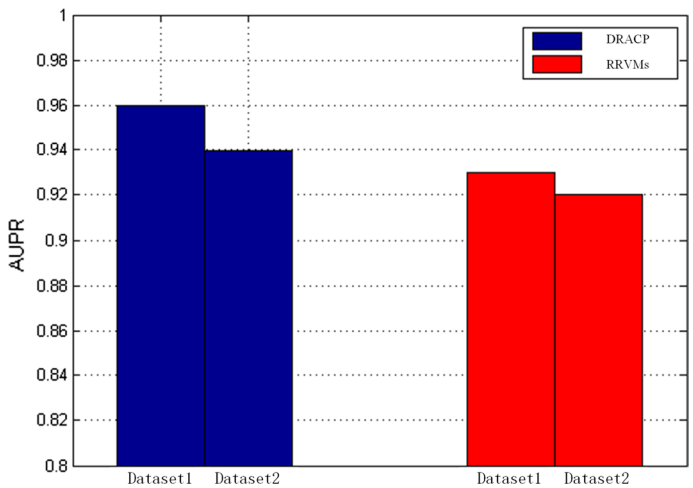

Fig. 2 The PR curves of DRACP and RRVMs

Discussion

Most of the previous methods for identifying ACPs are based on the traditional methods such as SVM. As the development of algorithms, more powerful methods should be applied into identifying ACPs.

In this paper, we used DBN to encode the feature of ACPs. DBN reduces dimension of ACP features through unsupervised learning. Then, we developed RRVMs which is a method based on RVM and RF to identify true ACPs. The experiments showed high precision of DRACP, which verified DRACP is an effective method for identifying ACPs. In addition, we also show the power of DBN by comparing the results of DRACP with RRVM's. This experiment explained the necessity of reducing dimension of features by

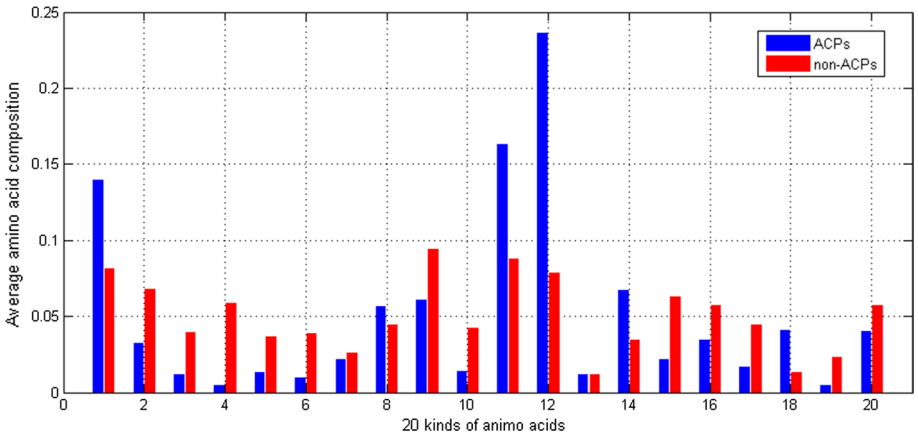

Fig. 3 Comparison of average whole amino acids composition of ACPs and non-ACPs. x-axis is the index of 20 kinds of amino acids and y-axis is the ratio of the amino acid to the total sequence length

DBN. Finally, we compared our method with previous methods and some traditional methods to show the superiority of DRACP.

DRACP can prior the potential ACPs based on their sequence. This work will help biologist reduce the cost of money and time on finding ACPs.

Conclusions

With its harmless advantages to the human body, ACPs have a great potential for treating cancers. However, due to the high cost of finding ACPs, not many ACPs have been found and there is still long way to go to use ACPs as a treatment.

To reduce the cost of money and time for finding ACPs, in this study, we proposed a method named DRACP to identify ACPs based on sequence and chemical characteristics of amino acids. Since the dimension of each ACP's feature is high, DBN was used to encode the features in a unsupervised way. It can effectively reduce the dimension and keep the information of features. After obtaining the final features, we randomly selected features and samples to build RVM models. 101 RVM models were built to generate a final classifier. This building process draw the idea of RF.

To verify the performance of DRACP, we use two independent datasets with 10-cross validation to do the test. We not only proved the performance of DRACP was better than previous method, but also showed the power of our features. In addition, we also test the performance of using RRVM without DBN and found DBN is an essential part for improving accuracy.

Overall, we developed an effective method for identifying ACPs. Although our method performed well, larger datasets are still needed to further prove the power of DRACP.

Methods

Feature extraction

Compositional analysis

We conjectured the composition of real ACPs are different from other normal peptides. Therefore, the average percentage of each amino acid is shown in Fig. 3.

Table 2 The six groups of the 20 amino acids

Groups	Amino acids
Strongly hydrophilic	R, D, E, N, Q, K, H
Strongly hydrophobic	L, I, A, V, M, F
Weakly hydrophilic Weakly hydrophobic	S, T, Y, W
Proline	P
Glycine	G
Cysteine	C

As shown in Fig. 3, the blue bar denotes the composition of real ACP and the red one is the non-ACPs'. Among the 20 amino acids, only 3 amino acids almost share the same percentage. Most of the composition of amino acids have significant differences between ACPs and non-ACPs. Therefore, the composition of 20 amino acids could be the features of ACPs.

The reduced amino acid composition

Protein structure is closely related to the patterns of hydrophobic and hydrophilic residues. The amino acids are divided into 6 groups based on the ranges of the hydropathy scale. Table 2 shows the six groups of the 20 amino acids.

Therefore, we can use six characters to represent the sequence of peptides. Since the dipeptides are consisted by two peptides, there would be 6^2 features to describe a sequence. Then we could extract the feature as following:

$$F = [f_1, f_2, f_3 \ldots f_{36}] \tag{1}$$

where f_x is the absolute occurrence frequencies of the 36 hydropathy dipeptides. It can be calculated as following:

$$f_i = \frac{n_i}{L-1} \tag{2}$$

where n_i is the occurrence number of the 36 hydropathy dipeptides of the protein, L is the length of peptide.

The Fig. 4 shows the flow chart of feature extraction. In total, we extracted 56 D features to identify the ACPs.

Methods and framework

Firstly, DBN was used to encode the features we obtained above. Then RRVM was used to classify ACPs. The workflow of our method is shown in Fig. 5.

DBN

DBN is an efficient semi-supervised algorithm. A layer-by-layer greedy algorithm is used to train the parameters of the deep belief network, breaking the deadlock that has been difficult for deep networks for a long time.

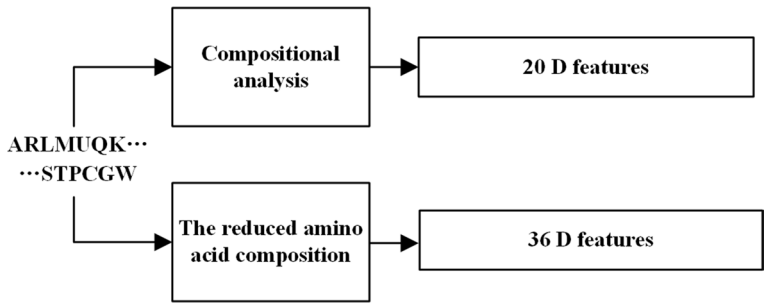

Fig. 4 Flow chart of feature extraction. We extracted 56 D features to identify the ACPs and it includes 20-dimensional composition and 36-dimensional reduced amino acid composition

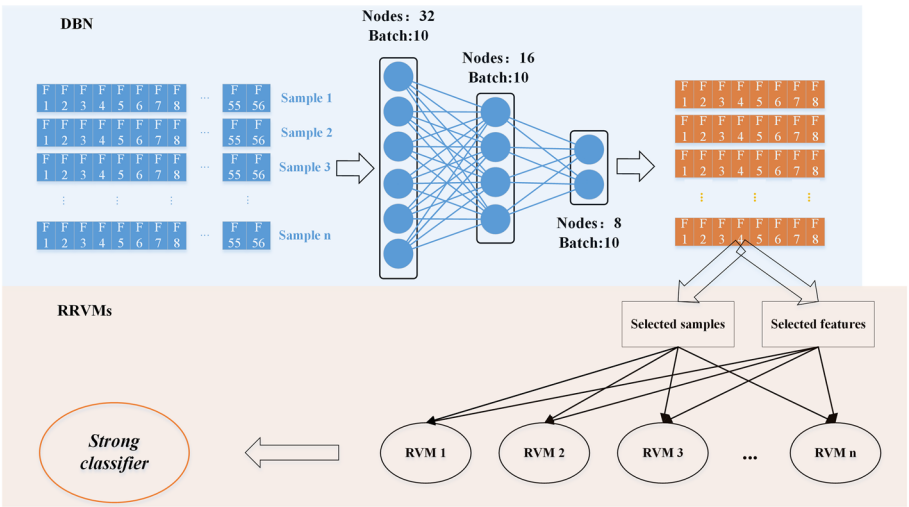

Fig. 5 Frame of DRACP. The first step is to use DBN to reduce the dimension of features. Then RRVMs is used to do classification

Restricted Boltzmann Machine (RBM) is the basic unit of DBN. The variables in RBM are divided into hidden variables and observable variables. These two sets of variables are represented by observable and hidden layers, respectively. There is no connection between nodes in the same layer, and nodes in one layer are connected to all nodes in another layer, which is as same as the fully connected neural network structure.

An RBM is composed of m observable variables and n hidden variables, and its energy function is defined as:

$$E(v,h) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i w_{ij} h_j = -a^T v - b^T v - v^T W h \tag{3}$$

Here, v is an observable variable $v = [v_1, v_2, ...v_m]^T$ and h is a hidden random vector $h = [h_1, h_2, ...h_n]^T$. W is a weight matrix, its dimension is m*n, and each element is the weight of the edge between the observable variable and the hidden variable. Both a and b are biases, a is the bias of the observable variable v, and b is the bias of the hidden variables.

The joint probability distribution of RBM is $p(v, h)$ which could be calculated by:

$$p(v,h) = \frac{1}{Z}\exp(-E(v,h)) = \frac{1}{Z}\exp(a^T v)\exp(b^T h)\exp(v^T Wh) \tag{4}$$

where $Z = \sum_{v,h}\exp(-E(v,h))$ is the partition function.

The essence of DBN is the stacking of RBMs. For a DBN containing L-level hidden variables, the lowest level is $v = h^{(0)}$ which is the observable variable. The top two layers are an undirected graph used to generate the prior distribution of $p(h^{(L-1)})$. Except for the top two layers, each layer can be calculated by the layer above it:

$$p(h^{(l)}|h^{(l+1)}, ..., h^{(L)}) = p(h^{(l)}|h^{(l+1)}) \tag{5}$$

The joint probability of variables in DBN can be denoted by:

$$p(v, h^{(1)}, ..., h^{(L)}) = p(v|h^{(1)})\left(\prod_{l=1}^{L-2} p(h^{(l)}|h^{(l+1)})\right)p(h^{(L-1)}, h^{(L)}) \tag{6}$$

where $p(h^{(l)}|h^{(l+1)})$ is sigmoid conditional probability distribution.

RRVMs

We learnt the basic idea from random forest (RF) to propose a new method RRVMs. By randomly selecting features and samples, RVM was built as a weak classifier. We repeated this process 101 times to construct a strong classifier.

First, we randomly select 5 features and 100 samples to build up a RVM model. Then, we put these features and samples back and started another round of building model. This process could be repeated 101 times, so 101 RVM models would be obtained. In the end, the strong classifier could be obtained by getting votes from these 101 RVM models.

The construction of RVM classifier

Compared with Support vector machine (SVM), the kernel function of RVM is not limited by Mercer conditions. It could be more sparse and has less super-parameters, so it reduces the computational burden of kernel functions.

For a given dataset $\{x_i, t_i\}_{i=1}^N$, $x_i \in \mathrm{R}^d$, non-linear model is :

$$t = y(x) + \varepsilon \tag{7}$$

where N is the sample number, y() is the non-linear function, ε is the noise, $\varepsilon \sim N(0, \sigma^2)$.

The final function of RVM is:

$$t = \mathbf{\Phi}\omega + \varepsilon \tag{8}$$

where $\omega = (\omega_0, \ldots, \omega_N)^{\mathrm{T}}$ is the weight, $\mathbf{\Phi}$ is the matrix of the kernel function. K() is the kernel function.$\phi_i(x_i) = [1, K(x_i, x_1), \ldots, K(x_i, x_N)], i = 1, 2, \ldots, N$.

The distribution of $p(t|x)$ meets $N(t|y(x), \sigma^2)$. Likelihood estimation of data is:

Table 3 Parameters and functions of RVM

Setting items	The value set
Max iterations	100
Kernel function	Gaussian
Kernel function width	6
Sample number	50
Feature number	10

$$p(t|\omega,\sigma^2) = (2\pi\sigma^2)^{-N/2}\exp\{-\|t-\Phi\omega\|^2/(2\sigma^2)\} \tag{9}$$

Tipping defines a zero mean Gauss type prior distribution on ω:

$$p(\omega/\alpha) = \prod_0^N N(\omega_i|0,\alpha_i^{-1}) = \prod_0^N \frac{\alpha_i}{\sqrt{2\pi}}\exp\left(\frac{\omega_i^2\alpha_i}{2}\right) \tag{10}$$

where α is the super-parameter, it is one-to-one correspondence to the weight.

α and the variance of noise σ^2 meet the Gamma distribution.

$$\begin{aligned} p(\alpha) &= \prod_{i=0}^N \text{Gamma}(\alpha_i|a,b) \\ p(\sigma^2) &= \prod_{i=0}^N \text{Gamma}(\beta|c,d) \end{aligned} \tag{11}$$

When there is a new set of observations, the prediction based on the sparse Bayesian learning framework can be expressed as:

$$p(t_{N+1}|t) = \int p(t_{N+1}|\omega,\alpha,\sigma^2)p(\omega,\alpha,\sigma^2|t)d\omega d\alpha d\sigma^2 \tag{12}$$

where t_{N+1} is the target value of the new observation x_{N+1}.

For a new set of inputs x_*, the output t_* should meet the distribution $p(t_*|t) \sim N(\mu^{\mathrm{T}}\Phi(x_*), \sigma_*^2)$.

$$t_* = \mu^{\mathrm{T}}\Phi(x_*) \tag{13}$$

$$\sigma_*^2 = \sigma_{MP}^2 + \Phi(x_*)^{\mathrm{T}}\sum\Phi(x_*) \tag{14}$$

where σ_{MP}^2 is the final variance of noise.

To accomplish the construction of classifier, we also need to set the various parameters as Table 3 shows.

Abbreviations

ACP: Anticancer peptides; DBN: Deep belief network; AMP: Antimicrobial peptide; RBM: Restricted Boltzmann machine; RF: Random forest; RRVM: Random Relevance Vector Machines; SVM: Support vector machine; LSTM: Long short-term memory.

Acknowledgements

The authors thank the anonymous referees for their many useful suggestions.

About this supplement
This article has been published as part of BMC Bioinformatics Volume 21 Supplement 16, 2020: Selected articles from the Biological Ontologies and Knowledge bases workshop 2019. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-21-supplement-16.

Authors' contributions
TYZ2 helped revise this paper. YH and TYZ1 wrote this paper and did the experiments. All authors have read and approved the final manuscript.

Funding
Publication costs are funded by the National Key Research and Development Program of China (No.: 2016YFC0901605) and the National Science and Technology Major Project (No. 2016YFC1202302). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
All the datasets used in this paper could be downloaded from https://github.com/zty2009/ACP.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 1 October 2020 Accepted: 13 October 2020
Published: 16 December 2020

References

1. Timmerman RD, Paulus R, Pass HI, Gore EM, Edelman MJ, Galvin J, Straube WL, Nedzi LA, McGarry RC, Robinson CG. Stereotactic body radiation therapy for operable early-stage lung cancer: findings from the NRG Oncology RTOG 0618 Trial. JAMA Oncol. 2018;4(9):1263–6.
2. Mereiter S, Balmaña M, Campos D, Gomes J, Reis CA. Glycosylation in the era of cancer-targeted therapy: where are we heading? Cancer Cell. 2019;36(1):6–16.
3. Gandhi L, Rodríguez-Abreu D, Gadgeel S, Esteban E, Felip E, De Angelis F, Domine M, Clingan P, Hochmair MJ, Powell SF. Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. N Engl J Med. 2018;378(22):2078–92.
4. Abnet CC, Arnold M, Wei W-Q. Epidemiology of esophageal squamous cell carcinoma. Gastroenterology. 2018;154(2):360–73.
5. Boman HG. Peptide antibiotics and their role in innate immunity. Annu Rev Immunol. 1995;13(1):61–92.
6. Falcao CB, Pérez-Peinado C, de la Torre BG, Mayol X, Zamora-Carreras H, Jiménez MA, Rádis-Baptista G, Andreu D. Structural dissection of crotalicidin, a rattlesnake venom cathelicidin, retrieves a fragment with antimicrobial and antitumor activity. J Med Chem. 2015;58(21):8553–63.
7. Felício MR, Silva ON, Gonçalves S, Santos NC, Franco OL. Peptides with dual antimicrobial and anticancer activities. Front Chem. 2017;5:5.
8. Gabernet G, Müller AT, Hiss JA, Schneider G. Membranolytic anticancer peptides. MedChemComm. 2016;7(12):2232–45.
9. Freire JM, Gaspar D, Veiga AS, Castanho MA. Shifting gear in antimicrobial and anticancer peptides biophysical studies: from vesicles to cells. J Pept Sci. 2015;21(3):178–85.
10. Zhao T, Hu Y, Zang T, Wang Y. Integrate GWAS, eQTL, and mQTL data to identify Alzheimer's disease-related genes. Front Genet. 2019;10:1021.
11. Oelkrug C, Hartke M, Schubert A. Mode of action of anticancer peptides (ACPs) from amphibian origin. Anticancer Res. 2015;35(2):635–43.
12. Arias M, Hilchie AL, Haney EF, Bolscher JG, Hyndman ME, Hancock RE, Vogel HJ. Anticancer activities of bovine and human lactoferricin-derived peptides. Biochem Cell Biol. 2016;95(1):91–8.
13. Mahlapuu M, Håkansson J, Ringstad L, Björn C. Antimicrobial peptides: an emerging category of therapeutic agents. Front Cell Infect Microbiol. 2016;6:194.
14. Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res. 2015;44(D1):D1087–93.
15. Bechinger B, Gorr S-U. Antimicrobial peptides: mechanisms of action and resistance. J Dent Res. 2017;96(3):254–60.
16. Cullen T, Schofield W, Barry N, Putnam E, Rundell E, Trent M, Degnan P, Booth C, Yu H, Goodman A. Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. Science. 2015;347(6218):170–5.
17. Tyagi A, Kapoor P, Kumar R, Chaudhary K, Gautam A, Raghava G. In silico models for designing and discovering novel anticancer peptides. Sci Rep. 2013;3:2984.
18. Khosravian M, Kazemi Faramarzi F, Mohammad Beigi M, Behbahani M, Mohabatkar H. Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. Protein Pept Lett. 2013;20(2):180–6.

19. Hajisharifi Z, Piryaiee M, Beigi MM, Behbahani M, Mohabatkar H. Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. J Theor Biol. 2014;341:34–40.
20. Chen W, Ding H, Feng P, Lin H, Chou K-C. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016;7(13):16895.
21. Manavalan B, Basith S, Shin TH, Choi S, Kim MO, Lee G. MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget. 2017;8(44):77121.
22. Grisoni F, Neuhaus CS, Gabernet G, Müller AT, Hiss JA, Schneider G. Designing anticancer peptides by constructive machine learning. ChemMedChem. 2018;13(13):1300–2.
23. Peng J, Zhu L, Wang Y, Chen J. Mining relationships among multiple entities in biological networks. In: IEEE/ACM transactions on computational biology and bioinformatics. 2019.
24. Zhao T, Cheng L, Zang T, Hu Y. Peptide-major histocompatibility complex class I binding prediction based on deep learning with novel feature. Front Genet. 2019;10:1191.
25. Peng J, Hui W, Li Q, Chen B, Hao J, Jiang Q, Shang X, Wei Z. A learning-based framework for miRNA-disease association identification using neural networks. Bioinformatics. 2019;35(21):4364–71.
26. Peng J, Xue H, Wei Z, Tuncali I, Hao J, Shang X. Integrating multi-network topology for gene function prediction using deep neural networks. Brief Bioinform. 2020.
27. Peng J, Wang X, Shang X. Combining gene ontology with deep neural networks to enhance the clustering of single cell RNA-Seq data. BMC Bioinform. 2019;20(8):284.
28. Plahl C, Sainath TN, Ramabhadran B, Nahamoo D. Improved pre-training of deep belief networks using sparse encoding symmetric machines. In: 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP): 2012. IEEE. p. 4165–4168.

Publisher's Note