**ACPred-Fuse: fusing multi-view information improves the prediction of anti-cancer peptides**

**Bing Rao1#, Chen Zhou2#, Guoying Zhang1\*, Ran Su3\***, **and Leyi Wei2\***

1 School of Mechanical Electronic & Information Engineering, China University of Mining &Technology, Beijing, China;

2 School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin, China.

3 School of Software, College of Intelligence and Computing, Tianjin University, Tianjin, China.

**#** Contribute equally to this work

\* Corresponding authors

E-mail: [zhangguoying1101@163.com](mailto:zhangguoying1101@163.com); [ran.su@tju.edu.cn](mailto:ran.su@tju.edu.cn); weileyi@tju.edu.cn

**Biographical Note:**

**Bing Rao** received his Master degree in Harbin Institute of Technology, China. He is currently a Ph.d. candidate in School of Mechanical Electronic & Information Engineering, at China University of Mining and Technology Beijing. His research interests are bioinformatics and machine learning.

**Chen Zhou** received his BSc degree in Software Engineering from Jinling Institute of Technology, China. He is currently a master student in School of Computer Science and Technology, College of Intelligence and Computing, at Tianjin University, China. His research interests are bioinformatics, machine learning and data mining.

**Guoying Zhang** received her PhD in Beijing Institute of Technology, China. She is currently a professor in School of Mechanical Electronic & Information Engineering, at China University of Mining and Technology Beijing. Her research interests are pattern recognition and artificial intelligence.

**Ran Su** is currently an Associate Professor in the School of Computer Software, College of Intelligence and Computing, at Tianjin University, China. Tianjin University, China. Her research interests include pattern recognition, machine learning and bioinformatics.

**Leyi Wei** received his PhD in Computer Science from Xiamen University, China. He is currently an Assistant Professor in School of Computer Science and Technology, College of Intelligence and Computing, at Tianjin University, China. His research interests include machine learning and their applications to bioinformatics.

**KEY POINTS**

• In this study, we present ACPred-Fuse, a powerful bioinformatics tool for the identification of anti-cancer peptides.

• In ACPred-Fuse, we introduce a feature representation learning algorithm that enables to learn informative features from multi-view feature space.

• Feature analysis demonstrates that class information, probabilistic information, and sequential information, are complementary to the performance improvement for anti-cancer peptide prediction.

• Benchmarking comparison shows that ACPred-Fuse significantly outperforms several state-of-the-art predictors.

• We establish a web server implementing ACPred-Fuse, which is publicly accessible at <http://server.malab.cn/ACPred-Fuse>.

**Abstract**

Fast and accurate identification of the peptides with anti-cancer activity potential from large-scale proteins is currently a challenging task. In this study, we propose a new machine learning predictor namely ACPred-Fuse that can automatically and accurately predict protein sequences with or without anti-cancer activity in peptide form. Specifically, we establish a feature representation learning model that can explore class and probabilistic information embedded in ACPs by integrating a total of 29 different sequence-based feature descriptors. In order to make full use of various multi-view information, we further fused the class and probabilistic features with handcrafted sequential features and then optimized the representation ability of the multi-view features, which are ultimately used as input for training our prediction model. By comparing the multi-view features and existing feature descriptors, we demonstrate that the fused multi-view features have more discriminative ability to capture the characteristics of ACPs. In addition, the information from different views is complementary for the performance improvement. Finally, our benchmarking comparison results showed that the proposed ACPred-Fuse is more precise and promising in the identification of ACPs than existing predictors . To facilitate the use of the proposed predictor, we built a web server, which is now freely available via <http://server.malab.cn/ACPred-Fuse>.

**Keywords**

Anti-cancer peptide; Feature representation; Machine learning; Random forest.

**Introduction**

Cancer remains the major cause of death killing millions of people every year and it is caused by the growth and spreading of abnormal cells without control. The International Agency for Research on Cancer (IARC) estimated that 12.7 million new cancer cases and 7.6 million cancer deaths occurred worldwide during 2008 [[1-3](#_ENREF_1)]. For cancer treatment, conventional chemotherapy is currently a common way. However, it usually has an adverse effect on normal cells, resulting in that the treatment effect is not that good [[4](#_ENREF_4)]. Moreover, the traditional chemotherapy approach is very expensive. New treatment options are therefore highly demanded for symptoms relieving and ultimately the eradication of the disease [[5](#_ENREF_5)]. Anti-cancer peptides (ACPs), with 5 to 30 amino acids long, have become potential therapeutic agents that address the shortcomings of the traditional chemotherapy, opening a promising perspective for cancer treatment [[5](#_ENREF_5), [6](#_ENREF_6)]. ACPs have some exceptional advantages. For example, ACPs are naturally occurring biologics, and hence are safer. In addition, ACPs are cationic in nature, which makes them selectively kill cancer cells by interacting with the anionic cell membrane components of cancer cells [[7](#_ENREF_7), [8](#_ENREF_8)]. Moreover, as compared to traditional anti-cancer drugs, ACPs have greater efficacy, selectivity, and specificity [[9](#_ENREF_9)]. Therefore, it is fundamentally important to identify ACPs especially from large-scale protein sequences that are generated by next-generation sequencing techniques.

Experimental identification can help to precisely determine whether the peptides have anti-cancer activity or not, but experimental approaches are usually time-consuming and cost-ineffective. It is not appropriate for high-throughput identification of ACPs. Subsequently, more recent research efforts are focused on developing data-driven computational methods like machine learning to identify ACPs. For machine learning based methods, the regular way is to specify the features of ACPs first, and then train a prediction model based on the features to automatically predict the peptides are more likely to be ACPs or non-ACPs. Therefore, how to characterize the discriminative features between ACPs and non-ACPs is becoming extremely important for accurate identification. In this regard, many computational efforts have been done. Notably, most of them aim to extract features from protein primary sequences only. Various sequence-based features have been proposed, including amino acid composition, dipeptide composition, and binary profile of pattern, etc. [[10](#_ENREF_10), [11](#_ENREF_11)]. Based on different sequence-based features, researchers have developed a variety of predictors. For example, using Support Vector Machine (SVM) as classification algorithm and protein relatedness measure as features, Vijayakumar et al. developed a prediction tool namely ACPP (Anti-Cancer Peptide Predictor) to assess query proteins for the presence of any apoptotic domains or not and then to predict anti-cancer activity in proteins [[12](#_ENREF_12)]. Chen et al. proposed a sequence-based predictor called iACP by optimizing the g-gap dipeptide components [[9](#_ENREF_9)]. Wei et al. proposed a novel predictor named ACPred-FL that can automatically extract and learn a set of informative features from a feature pool of support vector machine-based models trained using sequence-based feature descriptors[[13](#_ENREF_13)]. More recently, they further improved the performance by introducing more sequential information into their feature representation learning model, and established a generic predictor called PEPred-Suite for therapeutic peptides including anti-cancer peptides [[14](#_ENREF_14)]. Besides, there are other commonly used methods developed specific for the ACP prediction, such as AntiCP [[10](#_ENREF_10)], MLACP [[15](#_ENREF_15)], and Hajisharifi’s method [[16](#_ENREF_16)], etc. Although the progress has been made by existing predictors, the overall predictive performance is still not satisfactory enough for real therapeutic applications.

In this study, we proposed a novel predictor called ACPred-Fuse. To improve the predictive performance, we further improved the feature representation learning scheme to generate the discriminative distribution information from two aspects: predicted class and probability. To further boost the feature ability, we sufficiently combined multi-view information together, like the class label, prediction probability, and sequential information that benefit the prediction of ACPs. The experimental results show that the multi-view features have the most discriminative ability as compared with single-view features and existing feature descriptors. Moreover, we compared our ACPred-Fuse with six state-of-art predictors, and the comparison results showed that the ACPred-Fuse method is significantly superior to existing prediction tools, demonstrating it has great potential to a useful tool to predict ACPs.

**Methods and materials**

**Datasets**

In this study, we used the same dataset collected in our previous study [[17](#_ENREF_17)]. The positive dataset contains experimentally validated ACPs and the negative dataset is composed of anti-microbial peptides (AMPs). They collected experimentally validated ACPs (in Fasta format) from three main resources including Chen’s work [[9](#_ENREF_9)], Tyagi’s work [[10](#_ENREF_10)] and the largest ACP database CancerPPD [[18](#_ENREF_18)] as positive samples. To be specific, the positive dataset includes 138 samples from Chen’s work, 225 from Tyagi’s work, and 2849 from CancerPPD, respectively. Regarding the negative dataset, the AMPs which have proven to not have anti-cancer activity were considered as non-ACPs. Notably, to avoid over-estimation in performance by the homology bias, the sequence identity in both the positive and negative datasets were reduced to 0.8 using the CD-HIT [[19](#_ENREF_19)] program. Finally, the dataset containing 332 ACPs (positives) and 1023 non-ACPs (negatives).

To sufficiently measure the specificity of the predictor, we collected more other non-ACPs in addition to the negative dataset above. Our procedure for collecting non-ACPs is done as follows. Firstly, we collected protein sequences that don’t have any anti-cancer property from Swiss-Prot [[20](#_ENREF_20)]. The sequences are then fused into a long sequence, which was randomly spliced to yield the negative samples. By doing so, a total of 2,000 sequences were randomly selected and added to the dataset constructed by Wei et al. [[17](#_ENREF_17)]. Notably, we also used the CD-HIT program [[19](#_ENREF_19)] with a threshold of 0.8 to reduce the sequence identity of the dataset, ultimately yielding 332 ACPs and 2878 non-ACPs in our dataset.

***Training dataset****.* In order to train and evaluate the performance of the predictive model, we used 250 ACPs derived from the training dataset in Wei's study [[17](#_ENREF_17)] as the positive training samples. As for negative training samples, we randomly picked 125 non-ACPs from Wei’s study and 125 non-ACPs from the negative set generated in this study. To this end, a balanced training dataset is yielded.

***Independent test dataset.*** The use of an independent test dataset is to verify the generalization capability of the predictive model. Therefore, we constructed a test dataset that includes the remaining 82 ACPs from the positive dataset and the remaining 2628 non-ACPs in the negative set.

**Feature representation learning**

In our previous work, we first introduced the concept of feature representation learning[[13](#_ENREF_13), [14](#_ENREF_14), [21](#_ENREF_21)]. Inspired by the work, we further improved the feature representation learning scheme as shown in Figure 1. The procedure of the feature representation learning scheme is described in detail below.

In the first step, we constructed a feature pool. We used 29 different sequence-based feature encoding algorithms, as listed in Table 1. The details of all the feature algorithms can be found in Supporting Information. Through tuning the feature parameters, we yielded a total of 114 different types of feature descriptors. Notably, most of feature descriptors can capture only global sequential information based on full-length sequences. To explore local sequential information, we extracted features based on local sequences, instead such as the N-terminus (denoted as NT), C-terminus (denoted as CT), and the combination of NT and CT (denoted as NTCT) of the sequences.

In the second step, we used the Random Forest (RF) [[22](#_ENREF_22)] classifier to train the 114 feature descriptors respectively and build respective basic models. For each data sample, each well-trained model can provide us the probabilistic information and label information, respectively. The probabilistic information is the predicted confidence of the given sample, while the class information is the predicted label. Afterwards, by separately integrating the probabilistic information and the class information from all the 114 models, we obtained two 114-dimensional (114-D) feature vectors. Subsequently, the two newly formed 114-D features are optimized by MRMD (Maximal Relevance and Maximal Distance) approach and the SFS (sequential feature search) method. To be specific, the features are sorted according to their classification importance using the MRMD (Maximal Relevance and Maximal Distance) approach, and the sequential feature search (SFS) method is used to select the best feature subset for the generated probabilistic and class features, respectively.

In the final step, we combined the best individual feature descriptor from the basic feature pool with the optimal probabilistic features and class features to form multi-view features. Likewise, we further did the similar feature optimization procedure as the second step did and yielded the optimal features as our prediction model’s input.

**Feature selection based on Maximal Relevance and Maximal Distance**

Here we used the MRMD (Maximal Relevance and Maximal Distance) method for feature selection [[23](#_ENREF_23), [24](#_ENREF_24)]. The MRMD method is mainly for calculating the feature importance and ranking the features based on the feature importance. It can measure not only the relevance between features and their target classes but also the redundancy (also denoted as distance) between features themselves. In MRMD, the more important features have the better trade-off between maximum relevance and minimum redundancy. How to quantitatively measure the relevance and the redundancy is described as follows.

The relevance between features (*f*) and their target classes (*c*) is measured by Pearson Correlation Coefficient (PCC). To maximize the relevance, it is computed as:

where and ; N represents the number of features in the feature descriptor; represents the *k*-th value of the feature vector ; represents the *k*-th value of the target class vector c.

To measure the distance (or redundancy) of two feature vectors, we used the Euclidean distance function, which is calculated as:

where M is the number of samples in the feature descriptor; represents the Euclidean distance function, which is represented by:

Then use the correlation function and the distance function to calculate the maximum correlation and distance, which is represented by:

By applying the MRMD method to our feature set, we can obtain a ranked list of features, where the higher ranked features have a better trade-off between maximum correlation and minimum redundancy. On other words, higher the feature ranks, more important it is.

**Random Forest (RF)**

Random forest is one of the most powerful machine learning algorithms. It has been extensively used in many fields of computational biology [[25-28](#_ENREF_25)]. In this work, we also employed Random Forest (RF) algorithm to build our prediction models [[22](#_ENREF_22)]. The RF algorithm is kind of ensemble learning. It relies on the voting strategy of multiple decision trees to determine the final classification result. The details of RF can be referred to [[22](#_ENREF_22)]. In this study, the implementation of the RF algorithm was done in Python 2.7. It’s worth noting that the tree number is the main parameter in RF. To optimize the performance of RF, we tuned the tree number with a range of 100~500 followed by the step-size of 20. Especially for each feature descriptor, we trained and tuned to yield an optimized RF model. The optimal parameters for the RF models trained with different feature descriptors can be found in Supporting Information. Moreover, for feature selection, we also did the parameter optimization as well. The difference is that the parameter range is 50~500, while the step-size is 1. Detailed results of the parameter optimization can be also found in Supporting Information.

**Performance measurement**

To measure the performance, we used four metrics widely used in machine learning for two-class prediction problems, including Sensitivity (SE), Specificity (SP), Accuracy (ACC), and Matthew's correlation coefficient (MCC). The formulas of the four metrics are as follows:

where TP (True Positive) is the number of real ACPs predicted as real ACPs; TN (True Negative) is the number of non-ACPs predicted as non-ACPs; FP (False Positive) is the number of non-ACPs predicted as real ACPs. FN (False Negative) is the number of real ACPs predicted as non-ACPs. The SE and SP metrics measure the predictive power of the predictor for positive and negatives, respectively. The other two metrics, ACC and MCC, measures the overall performance of the predictor. In addition, we also used AUC (Area Under Curve) as the additional overall performance metric. AUC is defined as the area enclosed by the ROC (Receiver Operating Characteristic) curve and the coordinate axis. The higher score AUC achieves, the better performance the predictor is. To further compute the statistical significance in performance between two different predictive methods, we used the DeLong test [[29](#_ENREF_29)] which was implemented in MedCalc [[30](#_ENREF_30)].

**Results and discussion**

**Feature selection results**

In this study, we did the feature selection for the class and probabilistic features using the MRMD and SFS method, respectively. The MRMD is used to rank the features based on their classification importance and the SFS is used to determine the best feature subset. The feature selection results for the probabilistic and class features were illustrated in Supplementary Figure 1A and 1B, respectively. The detailed feature selection results can be also found in Supporting Information.

For the class features, as the feature number increases, the AUC of the prediction model reaches to a maximum of 0.881 while the ACC achieved 79.6% when the feature number 105. As the feature number reaches to 107, the ACC reaches a maximum of 80.6% while the AUC decreases to 0.875. Therefore, we determined the feature subset as the optimal when the feature number is 107, as it gives more balance between the ACC and AUC. Similarly, as for the probabilistic features, we found that when the feature number increases to 36, the ACC reaches a maximum of 81.6% and the corresponding AUC value is 0.878. When the feature number reaches to 96, the AUC achieves a maximum of 0.887, but the value of ACC drops to 80.6%. Accordingly, the 36 features are chosen as the best feature subset for the probabilistic features.

**Comparison of the class features, probabilistic features, and individual feature descriptors.**

To verify the effectiveness of our feature representation learning algorithm, we compared the features generated based on the class information and probabilistic information with existing sequence-based feature descriptors. In our feature representation learning framework, we used 29 sequence-based feature encoding approaches. By varying the feature parameters, we yielded a total of 114 different feature descriptors. The cross-validation results of all the feature descriptors are presented in Supporting Information. Among the 114 feature descriptors, we chose the top three feature descriptors [ZSC(NT=10), CTDC, and AAC(NT=9)] to compare with the class features (F107\_c) and the probabilistic features (F36\_p). Table 2 presents the cross-validation results for all the compared features. As seen in Table 2, the SEs of the three individual feature descriptors are 74.0%, 71.6%, and 68.4%, while the SEs of the class features (F107\_c) and the probabilistic features (F36\_p) descriptors are 76.0% and 76.4%, respectively. The improvement in SE is significant. In terms of SP, the class and probabilistic features are also slightly better than the other three individual feature descriptors. Therefore, the overall performance by the two model-generated features got improved. Specifically, using the class information, we boosted the ACC to 80.6% and AUC to 0.875. leading the three compared feature descriptors by 2% - 7.6% in ACC and 0.026-0.039 in AUC. Likewise, the probabilistic information is also helpful for the performance improvement. The generated probabilistic features improve the ACC by 3% - 8.6% and the AUC by 0.029 - 0.042 as compared with the best three exiting feature descriptors. In all, it demonstrates that the class and probabilistic information generated in the feature representation learning scheme can effectively improve the prediction performance.

To better interpret the feature space distribution, we adopted t-distributed Stochastic Neighborhood Embedding (t-SNE) [[31](#_ENREF_31)] to visualize and compare the feature space distribution of the two model-generated features and existing feature descriptors. The distribution is shown in Figure 2. It can be seen that either the probabilistic features or the class features can more clearly separate the positives from the negatives in space distribution than existing features. This demonstrates that the model-generated information is more effective to capture the difference between the positives and negatives. In addition, we further studied the classification contribution of the two feature descriptors, respectively. Table 3 presents the top ten important features that affect the predictive performance of the class and probabilistic feature descriptors. As shown in Table 3, the rankings for the top ten features between the two feature descriptors are different. This indicates that using different information, the features showed different importance for the classification.

**Fusing multi-view features improves the predictive performance**

To train the predictive model, we fused three features from different perspectives: (1) the class feature vector - F107\_c, (2) the probabilistic feature vector – F36\_p, and (3) the best individual sequence-based feature descriptor - ZSC (NT = 10). After combing the three features, we obtained a new feature vector of 193 dimensions, which we further did the feature optimization on by using MRMD and the SFS method. Figure 3A depicts SFS curve regarding the feature selection results of the multi-view features. It can be seen that as the number of features increases, the performance is steadily improved. When the feature number is increased to 174, the ACC and AUC are 82.4% and 0.882, respectively, achieving the best overall performance. Detailed results of feature selection using the SFS method can be found in Supporting Information.

Next, we investigated whether different views of features can be complementary to the performance improvement. We compared the multi-view features (denoted as F\_fuse) with three single-view features [ZSC(NT=10), F107\_c, and F36\_p]. The comparison results are illustrated in Figure 3B. We can observe that the F\_fuse improves the performances of single-view feature descriptors in terms of all the four metrics (ACC, AUC, MCC, SE, and SP), demonstrating that the different-view information are complementary to each other to improve the predictive performance.

**Comparison of the proposed predictor and existing predictors on cross-validation test**

To evaluate the predictive performance of the proposed ACPred-Fuse, we compared its performance with five state-of-the-art predictors, including AntiCP, Hajisharifi’s method, iACP, ACPred-FL, and PEPred-Suite. It is worth noting that we employed two prediction models of AntiCP for comparison. One is called AntiCP\_ACC trained with amino acid composition while the other is namely AntiCP\_DC trained with dipeptide composition. For fair comparison, they were trained with the same training dataset used in this study, and the performances were evaluated by 10-fold cross-validation. The cross-validation results are presented in Table 4.

As shown in Table 4 and Figure 4A, we observed that the performance of our proposed ACPred-Fuse is significantly better than other predictors in four out of five metrics (SE, MCC, AUC, and ACC). Specially, the SE and ACC of our predictor are 77.2% and 82.4%, respectively, which are 4.6% - 20% and 2% - 11.8% higher than other predictors. Although our SP is not the highest, it is slightly worse than the best SP of PEPred-Suite, only by 0.4% lower. Additionally, we used the ROC curve to compare the performance of ACPred-Fuse with other state-of-the-art predictors. As seen in Figure 4B, ACPred-Fuse is superior to other predictors in AUC, which is 0.022 - 0.073 higher than existing predictors. Furthermore, to analyze how reliable the compassion is, we computed the statistical significance between the proposed ACPred-Fuse and other predictors in terms of AUC. It can be seen also from Table 4 that our predictor is significantly better as compared with other predictors. These results indicate that ACPred-Fuse is capable of identifying ACPs more precisely than existing predictors.

**Comparison of the proposed predictor and existing predictors on independent test.**

To validate the robustness of ACPred-Fuse, we further tested and compared their performances using an independent test set (containing 82 ACPs and 2628 non-ACPs). To make the comparison fair, we used the same training dataset for model training. Figure 4C illustrates the performances of ACPred-Fuse and the six prediction models, while Figure 4D shows their corresponding ROC curves. Table 5 lists their detailed independent test results. As seen from Table 5, Figures 4C and 4D, we observed similar results with that under cross-validation test. That is, our predictor significantly outperforms the compared existing predictors for the independent test, with only one exception, which is that the proposed ACPred-Fuse is competitive with PEPred-Suite. Overall, the independent test results confirm that our predictor can better discriminate real ACPs from non-ACPs than existing predictors. However, we have to say that the above independent results might not 100% reflect their predictive performance in real applications mainly due to the limit of the positive number, which is relatively small for a reliable independent test. It can be anticipated that with the increment of experimentally validated ACPs, the evaluation and comparison would be more precise.

**Conclusion**

In this work, we have developed a new predictor called ACPred-Fuse that establishes an effective predictive model to identify ACPs. In this predictor, we have proposed a more informative feature representation scheme by integrating and learning from multi-view features. We have compared the proposed multi-view features with existing feature descriptors and single-view features. Our comparison results showed that the multi-view features can more effectively distinguish ACPs from non-ACPs than existing feature descriptors and single-view features, demonstrating the information from multiple views is complementary to each other, contributing to the performance improvement. To validate the performance of our ACPred-Fuse, we have evaluated and compared it with the state-of-art predictors using both cross validation test and independent test. Benchmarking comparison results demonstrate that the proposed predictor is more effective and promising for identifying ACPs. Moreover, we have also established a user-friendly web server for the convenient use of our predictor by a wider research community. We anticipate that ACPred-Fuse will be a useful tool for discovering new potential ACPs in a high-throughput and cost-effective manner. In future work, we will try other methods, such as deep learning [[32-34](#_ENREF_32)] and parallel computation techniques [[35-38](#_ENREF_35)], for more precise and efficient prediction.

**Acknowledgement**

The work was supported by the National Natural Science Foundation of China (Nos. 61701340 and 61702361), the Natural Science Foundation of Tianjin city (Nos. 18JCQNJC00500 and 18JCQNJC00800), and the National Key R&D Program of China (2018YFC0910405).

**References**

1. Ferlay J, Shin HR, Bray F et al. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008, International journal of cancer 2010;127:2893-2917.

2. Liu H, Luo LB, Cheng ZZ et al. Group-sparse Modeling Drug-kinase Networks for Predicting Combinatorial Drug Sensitivity in Cancer Cells, Current Bioinformatics 2018;13:437-443.

3. Liao Z, Li D, Wang X et al. Cancer diagnosis from isomiR expression with machine learning method, Current Bioinformatics 2018;13:57-63.

4. Holohan C, Van Schaeybroeck S, Longley DB et al. Cancer drug resistance: an evolving paradigm, Nature Reviews Cancer 2013;13:714.

5. Gaspar D, Veiga AS, Castanho MA. From antimicrobial to anticancer peptides. A review, Frontiers in microbiology 2013;4:294.

6. Huang Y, Feng Q, Yan Q et al. Alpha-helical cationic anticancer peptides: a promising candidate for novel anticancer drugs, Mini reviews in medicinal chemistry 2015;15:73-81.

7. Mader JS, Hoskin DW. Cationic antimicrobial peptides as novel cytotoxic agents for cancer treatment, Expert opinion on investigational drugs 2006;15:933-946.

8. Hoskin DW, Ramamoorthy A. Studies on anticancer activities of antimicrobial peptides, Biochimica et Biophysica Acta (BBA)-Biomembranes 2008;1778:357-375.

9. Chen W, Ding H, Feng P et al. iACP: a sequence-based tool for identifying anticancer peptides, Oncotarget 2016;7:16895.

10. Tyagi A, Kapoor P, Kumar R et al. In silico models for designing and discovering novel anticancer peptides, Scientific reports 2013;3:2984.

11. Zhang J, Ju Y, Lu H et al. Accurate Identification of Cancerlectins through Hybrid Machine Learning Technology, International Journal Of Genomics 2016;2016:7604641.

12. Vijayakumar S, Lakshmi P. ACPP: a web server for prediction and design of anti-cancer peptides, International Journal of Peptide Research and Therapeutics 2015;21:99-106.

13. Wei L, Zhou C, Chen H et al. ACPred-FL: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides, Bioinformatics 2018.

14. Wei L, Zhou C, Su R et al. PEPred-Suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning, Bioinformatics 2019.

15. Manavalan B, Basith S, Shin TH et al. MLACP: machine-learning-based prediction of anticancer peptides, Oncotarget 2017;8:77121.

16. Hajisharifi Z, Piryaiee M, Beigi MM et al. Predicting anticancer peptides with Chou′ s pseudo amino acid composition and investigating their mutagenicity via Ames test, Journal of Theoretical Biology 2014;341:34-40.

17. Wei L, Zhou C, Chen H et al. ACPred-FL: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides, Bioinformatics 2018;34:4007-4016.

18. Tyagi A, Tuknait A, Anand P et al. CancerPPD: a database of anticancer peptides and proteins, Nucleic acids research 2014;43:D837-D843.

19. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 2006;22:1658-1659.

20. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, Nucleic acids research 2000;28:45-48.

21. Qiang X, Zhou C, Ye X et al. CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learningQiang et al. A predictor for CPP identification, Briefings in bioinformatics.

22. Breiman L. Random forests, Machine learning 2001;45:5-32.

23. Zou Q, Zeng J, Cao L et al. A novel features ranking metric with application to scalable visual and bioinformatics data classification, Neurocomputing 2016;173:346-354.

24. Zou Q, Wan S, Ju Y et al. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy, Bmc Systems Biology 2016;10:114.

25. Wei L, Xing P, Zeng J et al. Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier, Artificial Intelligence in Medicine 2017;83:67-74.

26. Wei L, Wan S, Guo J et al. A novel hierarchical selective ensemble classifier with bioinformatics application, Artificial Intelligence in Medicine 2017;83:82-90.

27. Su R, Wu H, Xu B et al. Developing a Multi-Dose Computational Model for Drug-induced Hepatotoxicity Prediction based on Toxicogenomics Data, IEEE/ACM Transactions on Computational Biology and Bioinformatics 2018.

28. Zhao X, Zou Q, Liu B et al. Exploratory Predicting Protein Folding Model with Random Forest and Hybrid Features, Current Proteomics 2014;11:289-299.

29. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 1988:837-845.

30. Schoonjans F, Zalata A, Depuydt C et al. MedCalc: a new computer program for medical statistics, Computer methods and programs in biomedicine 1995;48:257-262.

31. Maaten Lvd, Hinton G. Visualizing data using t-SNE, Journal of machine learning research 2008;9:2579-2605.

32. Zhang Z, Zhao Y, Liao X et al. Deep learning in omics: a survey and guideline, Brief Funct Genomics 2018:doi: 10.1093/bfgp/ely1030.

33. Zou Q, Xing P, Wei L et al. Gene2vec: Gene Subsequence Embedding for Prediction of Mammalian N6‐Methyladenosine Sites from mRNA, rna 2019;25:205-218.

34. Li Y, Niu M, Zou Q. ELM-MHC: An improved MHC Identification method with Extreme Learning Machine Algorithm, Journal Of Proteome Research 2019;18:1392-1401.

35. Zou Q, Li X-B, Jiang W-R et al. Survey of MapReduce frame operation in bioinformatics, Briefings in Bioinformatics 2014;15:637-647.

36. Guo R, Zhao Y, Zou Q et al. Bioinformatics applications on Apache Spark, GigaScience 2018;7:giy098.

37. Song T, Rodríguez-Patón A, Zheng P et al. Spiking neural P systems with colored spikes, IEEE Transactions on Cognitive and Developmental Systems 2018;10:1106-1115.

38. Cabarle FGC, de la Cruz RTA, Zhang X et al. On string languages generated by spiking neural P systems with structural plasticity, IEEE Transactions on NanoBioscience 2018;17:560-566.

39. Bhasin M, Raghava GP. Classification of nuclear receptors based on amino acid composition and dipeptide composition, Journal of Biological Chemistry 2004;279:23262-23266.

40. Chen Z, Zhao P, Li F et al. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 2018;1:4.

41. Lee T-Y, Chen S-A, Hung H-Y et al. Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites, PloS one 2011;6:e17331.

42. Saravanan V, Gautham N. Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid composition-based feature descriptor, Omics: a journal of integrative biology 2015;19:648-658.

43. Wei L, Xing P, Su R et al. CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency, Journal of proteome research 2017;16:2044-2053.

44. Wei L, Tang J, Zou Q. SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides, BMC genomics 2017;18:1.

45. Li Z-R, Lin HH, Han L et al. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, Nucleic acids research 2006;34:W32-W37.

46. Dubchak I, Muchnik I, Holbrook SR et al. Prediction of protein folding class using global description of amino acid sequence, Proceedings of the National Academy of Sciences 1995;92:8700-8704.

47. Tomii K, Kanehisa M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, Protein Engineering, Design and Selection 1996;9:27-36.

48. Song L, Li D, Zeng X et al. nDNA-prot: identification of DNA-binding proteins based on unbalanced classification, BMC bioinformatics 2014;15:298.

49. Shen H-B, Chou K-C. Using ensemble classifier to identify membrane protein types, Amino acids 2007;32:483-488.

50. Wei L, Xing P, Shi G et al. Fast prediction of protein methylation sites using a sequence-based feature selection technique, IEEE/ACM Transactions on Computational Biology and Bioinformatics 2017:1-1.

51. Antos JM, Chew G-L, Guimaraes CP et al. Site-specific N-and C-terminal labeling of a single polypeptide using sortases of different specificity, Journal of the American Chemical Society 2009;131:10800-10801.

52. Govindan G, Nair AS. Composition, Transition and Distribution (CTD)—a dynamic feature for predictions based on hierarchical structure of cellular sorting. In: India Conference (INDICON), 2011 Annual IEEE. 2011, p. 1-6. IEEE.

53. Dou Y, Yao B, Zhang C. PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine, Amino acids 2014;46:1459-1469.

54. Wei L, Xing P, Tang J et al. PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only, IEEE transactions on nanobioscience 2017;16:240-247.

55. Kawashima S, Pokarowski P, Pokarowska M et al. AAindex: amino acid index database, progress report 2008, Nucleic acids research 2007;36:D202-D205.

56. Tung C-W, Ho S-Y. Computational identification of ubiquitylation sites from protein sequences, BMC bioinformatics 2008;9:310.

57. Sandberg M, Eriksson L, Jonsson J et al. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids, Journal of medicinal chemistry 1998;41:2481-2491.

**Figure legends**

**Figure 1.** **Flowchart of the proposed feature representation learning scheme.** There are three main steps. Firstly, we built a feature pool by varying the feature parameters based on 29 different feature encoding algorithms. Each feature descriptor is subsequently trained with RF classifier, generating class and probabilsitc information, respectively. Secondly, the resulting class probabilistic information are separately combined to form two feature vectors. Afterwards, we did the feature selection by using MRMD and SFS methods, obtaining the corresponding optimal feature subset. Finally, the two optimal feature subsets are merged with sequential features to form the multi-view features, the representation ability of which are further optimized. The resulting features are used to train the prediction model.

**Figure 2.** **T-SNE distribution of the three best individual feature descriptors,** probabilistic **features, and class features.** (A-C) denote the distribution of the three individual feature descriptors: ZSC(NT=10), CTDC, and AAC (NT=9), respectively; (D) denotes the distribution of the class features (F107\_c); (E) denotes the distribution of the probabilistic features (F36\_p).

**Figure 3. Predictive performance of our multi-view feature descriptor.** (A) SFS curves of our multi-view features in terms of ACC and MCC for the feature selection. (B) Performance comparison of our feature descriptor with its single-view feature descriptor evaluated with ten-fold cross validation.

**Figure 4.** **Performance comparison of our proposed ACPred-Fuse and state-of-the-art predictors.** (A) 10-fold cross-validation results of the proposed ACPred-Fuse and existing prediction models on the training dataset; (B) The ROC curves of the proposed ACPred-Fuse and existing prediction models on the training set; (C) Independent test results of the proposed ACPred-Fuse and existing six prediction models on independent test sets; and (D) The ROC curves of the proposed ACPred-Fuse and existing prediction models on the independent test set.

Table 1. Summary of 29 sequence-based feature descriptors.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Order** | **Feature descriptors** | **Description** | **Dimension** | **References** |
| 1 | **Amino acid composition (AAC(N=1))** | Appearance number of 20 different amino acids | 20-dimensional | [[39](#_ENREF_39)] |
| 2 | **Amino acid composition (AAC(N=2))** | Appearance frequency of 20 different amino acids | 20-dimensional | [[39](#_ENREF_39)] |
| 3 | **Grouped amino acid composition (GAAC)** | Physicochemical properties of amino acids | 5-dimensional | [[39-41](#_ENREF_39)] |
| 4 | **Di-Peptide Composition (DPC)** | Dipeptide composition | 400-dimensional | [[42](#_ENREF_42)] |
| 5 | **Grouped Di-Peptide Composition (GDPC)** | Physicochemical properties of dipeptides | 25-dimensional | [[40](#_ENREF_40)] |
| 6 | **Gap dipeptide composition (GGAP)** | Gap dipeptide composition (the range of parameter *g* is 1~4) | 400-dimensional | [[43](#_ENREF_43)] |
| 7 | **Adaptive skip dipeptide composition (ASDC)** | Adaptive skip dipeptide composition | 400-dimensional | [[43](#_ENREF_43), [44](#_ENREF_44)] |
| 8 | **Tri-Peptide Composition (TPC)** | Tripeptide composition | 8000-dimensional | [[39](#_ENREF_39)] |
| 9 | **Grouped Tri-Peptide Composition (GTPC)** | Physicochemical properties of tripeptides | 125-dimensional | [[39](#_ENREF_39), [40](#_ENREF_40)] |
| 10 | **Composition (CTDC)** | Percentage of particular amino acid property groups | 39-dimensional | [[45-47](#_ENREF_45)] |
| 11 | **Transition**  **(CTDT)** | Percentage of mutual conversion in amino acid properties | 39-dimensional | [[45-47](#_ENREF_45)] |
| 12 | **Distribution**  **(CTDD)** | Distribution of amino acid properties in sequences | 195-dimensional | [[45-47](#_ENREF_45)] |
| 13 | **Composition, Transition and Distribution (CTD)** | Distribution of amino acid properties in sequences | 63-dimensional | [[45-47](#_ENREF_45)] |
| 14 | **188-dimensional feature (188D)** | Distribution of amino acids and physicochemical properties | 188-dimensional | [[48](#_ENREF_48)] |
| 15 | **Composition of k-spaced Amino Acid Pairs** | Composition of k-spaced amino acid pairs | 2400-dimensional | [[40](#_ENREF_40)] |
| 16 | **Composition of k-Spaced Amino Acid Group Pairs (CKSAAGP)** | Composition of k-spaced amino acid group pairs | 150-dimensional | [[40](#_ENREF_40)] |
| 17 | **Conjoint Triad (CTriad)** | Conjoint Triad | 343-dimensional | [[49](#_ENREF_49)] |
| 18 | **K-Spaced Conjoint Triad (KSCTriad)** | K-Spaced Conjoint Triad | 343-dimensional | [[40](#_ENREF_40)] |
| 19 | **Dipeptide Deviation from Expected Mean (DDE)** | Dipeptide Deviation from Expected Mean | 400-dimensional | [[42](#_ENREF_42)] |
| 20 | **Twenty-Bit Features (BIT20)** | Binary profiles for amino acids  with N-terminal approach (NT) (the value of NT is from 1 to 10) | 20×NT-dimensional | [[43](#_ENREF_43), [50](#_ENREF_50), [51](#_ENREF_51)] |
| 21 | **Twenty-One-Bit Features (BIT21)** | Binary code for the physicochemical properties of amino acids  (NT) | 21×NT-dimensional | [[51](#_ENREF_51), [52](#_ENREF_52)] |
| 22 | **Overlapping Property Features (OLP)** | Binary code for classification of amino acid physicochemical properties  (NT) | 10×NT-dimensional | [[51](#_ENREF_51), [53](#_ENREF_53)] |
| 23 | **Information Theory Features (IT)** | Sequence information theory  (NT) | 3-dimensional | [[51](#_ENREF_51), [54](#_ENREF_54)] |
| 24 | **AAindex (AAIN)** | Public index of the physicochemical properties  (NT) | 531×NT-dimensional | [[51](#_ENREF_51), [55](#_ENREF_55), [56](#_ENREF_56)] |
| 25 | **BLOSUM62 (BLO)** | Encode according to the BLOSUM62 matrix  (NT) | 20×NT-dimensional | [[40](#_ENREF_40), [41](#_ENREF_41), [51](#_ENREF_51)] |
| 26 | **Z-Scale (ZSC)** | Five physicochemical properties to characterize each amino acid  (NT) | 5×NT-dimensional | [[51](#_ENREF_51), [57](#_ENREF_57)] |
| 27 | **Amino acid composition (AAC(NT))** | The amino acid composition (Probability) (NT) | 20-dimensional | [[39](#_ENREF_39), [51](#_ENREF_51)] |
| 28 | **Amino acid composition (AAC(NTCT))** | The amino acid composition (Probability)  N+C-terminal approach (NTCT) (the value of NTCT is 5~10) | 20-dimensional | [[39](#_ENREF_39), [51](#_ENREF_51)] |
| 29 | **Grouped amino acid composition (GAAC(NTCT))** | Physicochemical properties of amino acids  (NTCT) | 5-dimensional | [[39-41](#_ENREF_39), [51](#_ENREF_51)] |