Identification of Phage Virion Proteins by Using the g-gap Tripeptide Composition

Article *in* Letters in Organic Chemistry · September 2018

DOI: 10.2174/1570178615666180910112813

4 authors, including:

Liangwei Yang
University of Illinois at Chicago
13 PUBLICATIONS **106** CITATIONS
SEE PROFILE

Hui Gao
University of Shanghai for Science and Technology
103 PUBLICATIONS **1,679** CITATIONS
SEE PROFILE

Lixia Tang
58 PUBLICATIONS **1,522** CITATIONS
SEE PROFILE

Send Orders for Reprints to reprints@benthamscience.ae

Letters in Organic Chemistry, **Year,** *Volume*

Identification of phage virion proteins by using the g-gap tripeptide composition

Liangwei Yang [a], Hui Gao[*a], Zhen Liu [1a], and Lixia Tang [2b]

[a] *School of Computer Science and Engineering, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China.;* [b]*Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China*

ARTICLE HISTORY

Received:
Revised:
Accepted:

DOI:

Abstract: Phages are widely distributed in locations populated by bacterial hosts. Phage proteins can be divided into two main categories, that is, virion and non-virion proteins with different functions. In practice, people mainly use phage virion proteins to clarify the lysis mechanism of bacterial cells and develop new antibacterial drugs. Accurate identification of phage virion proteins is therefore essential to understanding the phage lysis mechanism. Although some computational methods have been focused on identifying virion proteins, the result is not satisfying which gives more room for improvement. In this study, a new sequence-based method was proposed to identify phage virion proteins using g-gap tripeptide composition. In this approach, the protein features were firstly extracted from the g-gap tripeptide composition. Subsequently, we obtained an optimal feature subset by performing incremental feature selection (IFS) with information gain. Finally, support victor machine (SVM) was used as the classifier to discriminate virion proteins from non-virion proteins. In 10-fold cross validation test, our proposed method achieved an accuracy of 97.40% with AUC of 0.9958, which outperforms state-of-the-art methods. The result reveals that our proposed method could be a promising method in the work of phage virion proteins identification.

Keywords: phage virion proteins, g-gap tripeptide composition, SVM, IFS, information gain, 10-fold cross validation

1. INTRODUCTION

The bacteriophage, also known as phage, is a kind of virus that infects and replicates within bacteria and archaea. They are among the most common and diverse entities in the biosphere. Bacteriophages are composed of proteins that encapsulate a DNA or RNA genome and may have relatively simple or elaborate structures. Sometimes a layer of lipid film is wrapped around the protein shell when it is outside the cell. The proteins coded by phage genes are regarded as a possible therapy against multi-drug-resistant strains of many bacteria for which it is important to get a further understanding of the phage proteins.

Phage proteins can be divided into virion proteins and non-virion proteins. The virion protein produced by phage gene compilation is an important part of the assembled phage particle, including capsid protein, envelope protein and virus granzyme [1]. These virion proteins determine the specificity of the host bacteria and play an important role in phage virus recombination, receptor recognition, bacterial attachment and infiltration [2]. Phage non-virion proteins, also compiled from the phage genome and synthesized in infected cells, are not been packaged in mature bacteriophage particles [1]. These non-viral proteins, mainly enzymes and regulatory proteins, play an important role in the biological processes of phage gene replication, transcription and expression [3]. The function of virion proteins is quite different from non-virion proteins while in practice, virion proteins are considered. Having a knowledge of phage virion proteins is important for understanding the mechanism of interaction between phage and its host bacteria as well as the development of new antibacterial drugs. For example, phage can be used in the identification of bacteria for the high specificity of phages and that it can rapidly multiply and produce phage proteins after injected. So the existence of the corresponding bacteria can be proved if the number of phages detected in experiment increases sharply. Phage can also be used in the treatment of diseases. They have the function of lysing corresponding bacteria, which can be used to deal with the drug-resistant bacteria and it is safer and more effective than antibiotics. The treatment tests using phage have already begun. Pherecydes Pharma from France is now testing the anti-infective effect of different combinations of bacteriophage in burn wards of hospitals in Western Europe with the support of the European Union. In previous animal experiments, phage therapy has shown significant anti-E.coli (Escherichia coli) and

* Correspondence: Tel: +8613550202554 E-mails: huigao@uestc.edu.cn

Pseudomonas aeruginosa infections with considerable reliability.

Identification of phage virion proteins becomes important for the value behind it. Machine learning approaches have been proven as powerful and efficient tool in dealing with various biological problems. Actually, Seguritan et al. have proposed an Artificial Neural Network (ANN)-based method to classify viral structural proteins by using amino acid composition and protein isoelectric points [4]. A Naïve Bayes-based method was proposed to predict phage virion proteins using amino acid composition and dipeptide composition by Feng et al. [5]. A sequence-based method was developed to identify phage virion proteins by the ANOVA (Analysis of variance) feature selection and analysis by Ding et al. [6]. Although the aforementioned methods could yield encouraging results, the accuracies of these methods still need improvements.

This study is devoted to improve the prediction quality of phage virion protein prediction. Firstly, we propose a new feature constructing method based on g-gap tripeptide composition that extracts features from the protein sequences. Subsequently, the information gain is used as the feature ranking criterion to judge the importance of each feature on the classification results. We then use an incremental feature selection (IFS) method after sorting the features in descending order based on the information gain to find the optimal feature subset. Finally, the support vector machine (SVM) is used as the classifier to identify phage virion proteins. A 10-fold cross validation test was used to evaluate the method, results of which revealed that our proposed method could be a high accuracy prediction tool to identify phage virion proteins. Furthermore, for the convenience of other related works, an online web-sever was established and can be freely accessed from the website (http://bigroup.uestc.edu.cn/virionpred/).

To develop a really useful sequence-based predictor for a biological system as reported in a series of recent publications [7]–[21] , we adopted Chou's 5-step rule [22], which states as follows: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. The aforementioned rules were addressed in detail in the material and methods section.

2. RESULTS AND DISCUSSION

2.1 Feature selection for improving accuracy

According to the g-gap tripeptide composition mentioned in the 3.2 section, a total of 8000-dimensional features was constructed while the training sample number is only 307, which will lead to the curse of dimensionality. For example, the final model accuracy of the 8000-dimensional 2-1-gap tripeptide composition features valued by 10-fold cross validation is only 69.74%. Although the low-dimensional features can make the model more robust, an inadequate feature will make the information provided by the features insufficient and the model can only obtain a low accuracy. When we only take the first 10 features of the 2-1-gap tripeptide feature set, the accuracy of the model after 10-fold cross validation records only 72.96%. Hence, some evaluation methods were used to select the optimal feature subset from the feature set to reduce the dimension of the feature. Obviously, we can get the optimal feature subset by testing all possible combinations of features, which is impossible to test one by one. Take the 20-dimensional features of a single amino acid as an example. The possible combinations of these 20-dimensional features are $C_{20}^{1} + C_{20}^{2} + C_{20}^{3} + \cdots C_{20}^{19} + C_{20}^{20} = 1048575$. Obviously, for an 8000-dimension feature set, the combination is much larger. To save computing time and resources, information gain was used as the criterion to measure the importance of the feature and then the incremental feature learning (IFS) method was used to construct the optimal feature set. The final performance of model was evaluated by accuracy (*Acc*), sensitivity (S_n), specificity (S_p) and Mathew's correlation coefficient (*MCC*).

Table 1: The best result of each former gap

g-gap TC	Num. of features	Sn(%)	Sp(%)	*Acc*(%)	*MCC*
0-5-gap	601	86.86	98.55	94.79	0.8801
1-5-gap	751	89.89	99.51	96.41	0.9181
2-1-gap	771	91.91	100	97.40	0.9408
3-3-gap	791	89.89	98.55	95.75	0.9025
4-4-gap	781	89.89	100	96.75	0.9261
5-8-gap	721	90.90	98.55	96.10	0.9100
6-0-gap	771	89.89	100	96.73	0.9261
7-8-gap	541	87.87	100	96.08	0.9115
8-3-gap	791	89.89	99.51	96.41	0.9181
9-5-gap	781	89.89	99.51	96.43	0.9641

Using the feature selection approach mentioned above, an SVM based classifier was built to classify the phage protein dataset and valued by 10-fold cross validation. There are 100 results for the two gaps are all vary from 0 to 9. The best result of each former gap is selected and shown in the Figure 1. The accuracy rate and other indicators on their respective feature subsets are compared. Specific results are shown in table 1.

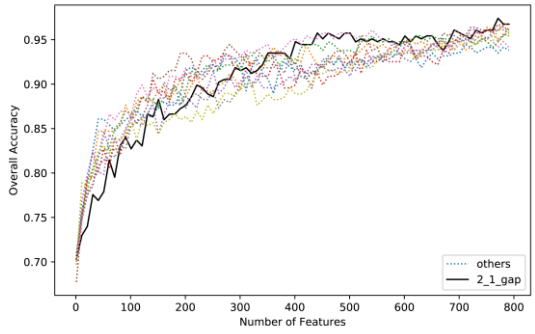

Figure 1: The IFS procedure of each former gap

Table 2: The final results of the 4 methods

method	feature	feature ranking criterion	*Sn*(%)	*Sp*(%)	*Acc*(%)	*MCC*
1(Ding's)	g-gap DC	ANOVA	75.76	89.42	85.02	0.6554
2	g-gap DC	information gain	71.71	89.90	84.09	0.6287
3	g-gap TC	ANOVA	92.92	99.51	97.39	0.9403
4(our method)	g-gap TC	information gain	91.91	100	97.40	0.9408

From Table 1, it can be seen that the classification results using the g-gap tripeptide is impressive, with a classification accuracy of up to 97.40% and some classification accuracy of negative samples up to 100%. The model built on the features consist of the top 771 highest information gain of the 2-1-gap tripeptide composition achieves the best accuracy of 97.40% with a positive recall rate of 91.91% and negative recall rate of 100%. From figure 1, we may notice that the accuracy is high when the number of features are more than 200 no matter what the gap value is, which indicates that the quality of features built by g-gap tripeptide composition is stable.

2.2 Comparison with other methods

In the study of phage proteins, Feng et al used 20 types of amino acids and 400-dimensional dipeptide to form 420-dimensional eigenvectors, which were classified by Naïve Bayes and achieved an accuracy of 79.15% [5]. Ding et al constructed 400-dimensional features using the g-gap dipeptide composition leading to an accuracy of 85.02% [6]. Zhang et al used random forest on four groups of characteristics constructed respectively from the composition, transformation distribution and pseudo-amino acid composition, which achieved an accuracy of 85% after the models were ensembled [13]. The results of this study were compared with Ding's method (method 1) since it is the most similar one to ours. For more comparison, we built another two methods based on the same process except for the method of feature extraction and feature ranking. Namely, method 2 (g-gap DC & information gain) is different in feature extraction and method 3 (g-gap TC & ANOVA) is different in feature ranking criterion. The detailed results are shown in table 2.

It can be seen from the table that the method proposed in this study based on g-gap tripeptide composition and information gain achieves the best accuracy of 97.40% with the best specificity and *MCC*. Compared with the Ding's method (method 1), the accuracy of our method is higher by 12.38% with sensitivity surpasses by 16.15%, specificity surpasses by 10.58% and a higher MCC of 0.9408, which is a significant improvement in the identification of phage virion proteins. Besides, it is obvious that the results of method 3 and method 4 are much better than method 1 and 2, which indicates that the features extracted by the g-gap tripeptide composition are of better quality than the ones built by the g-gap dipeptide composition. It can also be seen from the table that the choice of feature ranking criterion between ANOVA and information gain makes little difference when the feature extraction method is fixed. In addition, we plot the ROC curves of the 4 methods for a deeper comparison. The AUC values of each method are also shown in figure 2.

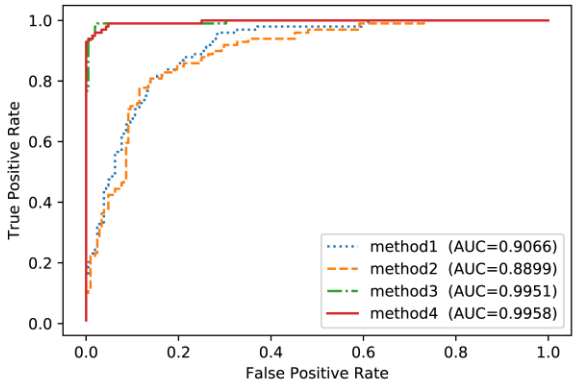

Figure 2: The ROC curves and AUC values of the 4 methods

As we can see from the figure 2, the method proposed in this paper (method 4) obtains the highest AUC of 0.9958, which surpasses the AUC of Ding's method. The ROC curve of Ding's method is all wrapped within ours, which shows the method proposed in this paper is more accurate. Moreover, compared with Ding's, our method still gets a higher accuracy of 86.96% when the feature number is exactly the same as Ding's method, which reveals that our method is more robust. Note that the main difference between method 3 and Ding's method is the feature construction method. Ding used the g-gap dipeptide composition while method 3 used the g-gap tripeptide composition. It can be revealed that the feature quality extracted from the g-gap tripeptide composition leading to a higher AUC is better than the g-gap dipeptide composition. Other conclusions are the same as what we got from the table 2, indicating that the g-gap tripeptide composition (g-gap TC) is a useful feature extraction method for related works.

3. MATERIALS AND METHODS

3.1 Benchmark dataset

The database is the starting point for all bioinformatics work. At present, there are many free and open databases of protein sequences. The raw data in this study was obtained from a database named UniProt [23] for being the least redundant and most comprehensive of functional annotations in all the related database. For the purposes of obtaining a reliable benchmark dataset, the following rules were considered. Firstly, positive samples (virion proteins) are the phage proteins whose subcellular location is a virion while negative samples (non-virion proteins) are the ones whose subcellular location is otherwise. The protein sequences which are fragments of other proteins were dislodged. Secondly, a protein will be dislodged if its sequence length is less than 50 for the reason that these fragments cannot

describe the information contained in the entire protein sequence completely, which will lead to a bad accuracy of prediction results. Thirdly, if the protein sequences contain nonstandard letters, such as B, U, X or Z, these proteins were excluded for their ambiguous meanings. After following the previous strict screening procedures, a total of 121 phage virion and 231 phage non-virion proteins were obtained. The model obtained by training on a benchmark dataset which has a high degree of homology will be overestimated for not being representative. In order to get rid of redundant data, the CD-HIT tool was used to remove the high similarity sequence by setting a cutoff threshold of 40%. After the final step, 307 sequences were remained in the benchmark dataset, including 99 phage virion protein sequences and 208 phage non-virion protein sequences. These proteins can be found from http://bigroup.uestc.edu.cn/virionpred/data/.

3.2 The *g*-gap tripeptide composition (g-gap TC)

The protein sequence is an indefinite length string consist of *20* English letters representing *20* different kinds of amino acids. Our first major concern is to translate a protein sequence into a mathematical expression for statistical prediction and to develop a sequence-based predictor for identifying phage virion protein. The most straightforward translation method is to translate the sample of protein *P* with *L* residues with its entire amino acid sequence as:

$$P = R_1 R_2 R_3 \cdots R_L \tag{1}$$

where R_1 represents the 1st residue of protein *P*, R_2 represents the 2nd residue of protein *P*, and so forth. Subsequently, we can utilize sequence-similarity-search-based tools, such as BLAST algorithm [24] to identify protein virions based on sequence similarity. However, it fails to work when the query sequence doesn't have a high similarity with any sample in the training dataset. Thus, researchers often use vectors to represent the samples, which is much easier to handle via existing operation engines. There have been many feature extraction methods proposed.

With the explosive growth of biological sequences in the post-genomic era, extracting features from proteins is not just one of the most important but also most difficult problems in computational biology. This is because all the existing machine learning algorithms can only handle vector but not sequence samples, as elucidated in a comprehensive review [25]. The simplest representation of protein sequence discretization is based on the amino acid composition (AAC) proposed by Nakashima and Nishikawa et al. [26]. This method is based on the hypothesis that the amino acid composition ratio determines protein characteristics, and the protein sequence P is represented by a 20-dimensional vector. However, this method completely ignores the information brought about by the sequence order. To avoid completely losing the sequence-pattern information for proteins, Chou et al. proposed a method that considers the composition of the pseudo-amino acids that affect the sequence residues. They added the physicochemical properties of amino acids when constructing sequence features and proposed the pseudo amino acid composition (PseAAC) [27]. Ever since the concept of Chou's PseAAC was proposed, it has been widely used in nearly all areas of computational proteomics [11], [26]–[41], for the increasingly usage of PseAAC, recently three powerful open access soft-wares, called 'PseAAC-Builder', 'propy' and 'PseAAC-General' were established to generate various modes of Chou's special PseAAC [42]. Notably, 'PseAAC-General' provides higher level feature vectors such as 'Functional Domain' mode, 'Gene Ontology' mode, and 'Sequential Evolution' and 'PSSM' mode. Based on the global description of protein sequences, Dubchak et al. proposed the composition, transformation and distribution (CTD) of amino acids feature building method [43]. In order to obtain more sequence-related information and find important relevant features, Lin et al. proposed a more general dipeptide composition called g-gap dipeptide composition (g-gap DC) [6]. While compared with dipeptide composition, the tripeptide composition contains more sequence order and composition information of protein sequences. Wang et al. used amino acid polarity and side chain group masses to classify 20 amino acids into 7 groups, and then counted three consecutive amino acids, namely the frequency of occurrence of tripeptide composition, and constructed a 343-dimensional feature into SVM for training [44]. Lai et al. used a tripeptide composition consisting of three amino acids in succession and had a total of 8000-dimensional features. They also used SVM for classification and achieved good results on cancer relating proteins [45]. Particularly, recently a very powerful web-server called 'Pse-in-One' [46] now updated to version 'Pse-in-One2.0' have been established that can be used to generate feature vectors for protein/peptide and DNA/RNA sequences according to the users' need or their own definition.

Inspired by the above methods, we proposed a feature extraction method called g-gap tripeptide composition (*g*-gap TC) to mine the information contained in protein sequences. We still use formula (1) to represent the protein sequence. As for the *g*-gap tripeptide composition, it can be showed as

$$F = R_1 \overset{\text{gap}_1}{\therefore} R_1 \overset{\text{gap}_2}{\therefore} R_3 \tag{2}$$

where the R_1, R_2 and R_3 represent the standard amino acids. gap_1 is the gap between the first two residues and gap_2 is the gap between the last two residues. There are $20 \times 20 \times 20 = 8000$ kinds of tripeptide composition for 20 kinds of standard amino acids. In this study the gaps vary from *0* to *9* to find the optimal gaps combination. When $gap_1 = g_1$ *and* $gap_2 = g_2$, a protein sequence can be discretized as

$$\mathbf{P} = \left[f_1^{g_1,g_2}, f_2^{g_1,g_2}, \cdots, f_\xi^{g_1,g_2}, \cdots, f_{8000}^{g_1,g_2} \right]^{\mathrm{T}} \tag{3}$$

where **T** represents a transpose operation, $f_\xi^{g_1,g_2}$ represents the frequency of the ξ-th tripeptide composition. $f_\xi^{g_1,g_2}$ is calculated by

$$f_\xi^{g_1,g_2} = \frac{n_\xi^{g_1,g_2}}{\sum_{\xi=1}^{8000} n_\xi^{g_1,g_2}} = \frac{n_\xi^{g_1,g_2}}{L - g_1 - g_2 - 2} \tag{4}$$

where $n_\xi^{g_1,g_2}$ represents the occurrence number of the ξ-th $g_1\_g_2\_gap$ tripeptide composition, *L* is the length of protein *P*.

3.3 Feature selection

In order to economize the run-time and computational resources, it is a wise strategy to use a feasible algorithm to find the optimal features and eventually improve the prediction quality. To reduce the dimensions of the feature space and improve the precision of phage virion and non-virion protein classification, information gain combined with

incremental feature selection (IFS) was performed during the process of feature selection in current works. Information gain is widely used as the term importance criterion in test classification problems. For a specific feature, the change in the amount of information with and without the feature is the feature's information gain. The so-called amount of information is the entropy, described as follows:

$$H(C) = -\sum_{i=1}^{n} p(c_i) \log p(c_i) \quad (5)$$

where *C* represents a variety of n different possible values, each of which represented by c_i. $p(c_i)$ represents the probability of c_i. *H(C)* is the entropy of variety *C*. In this study, *C* only has *2* possible values *0* and *1*, which represents the negative and positive samples respectively. The possibility of each kind of category is calculated from the frequency. For protein classification, a feature contributes more if the classification based on this single feature makes the entropy of the classification result smaller. Each feature was initially processed by bi-partition method and calculated the best information gain as its information gain value just as *C4.5* decision tree deals with continuous variable [47]. For each round, the information gain brought by feature f to classification *C* can be calculated as

$$IG(f) = H(C) - H(C|f) \quad (6)$$

where $H(C|f)$ has two cases. One is the appearance of the feature *f*, labeled *f*, and the other is that the feature *f* does not appear, marked as f'. So the *H(C|f)* can be calculated by:

$$H(C|f) = P(f)H(C|f) + P(f')H(C|f') \quad (7)$$

Obviously, the larger the $IG(f)$ is, the greater the discrimination degree of the feature to the samples. Hence, we can sort all the features according to the $IG(f)$ in descending order to obtain the following feature set:

$$F = \{f_1, f_2, \cdots f_N\} \quad (8)$$

where f_1 denotes the feature with highest information gain, f_2 denotes the feature with second highest information gain and so forth. Subsequently, the incremental feature selection (IFS) method is used to select the optimal number of features. IFS method was used as follows for the reason that the number of features should not be too large because of the limited amount of samples and it would be too much time-consuming if the features were added one after the other. Firstly, we chose the feature subset started from a feature with the highest information gain in the ranked feature set. Secondly, the features with the next 10 highest information gain was added to the subset to obtain a new feature subset. The process was repeated until 800 candidates were added. The feature subset generated by the *i*-th iteration can be expressed as:

$$F_i = \{f_1, f_2, \cdots, f_{1+10*i}\} \quad (9)$$

A model was built for each feature subset on the benchmark dataset. The subset with the highest accuracy is selected as the final optimized feature subset.

3.4 Support vector machine (SVM)

Support Vector Machine (SVM) algorithm is widely used in bioinformatics and have had a good performance on protein classification problems. The classification idea of SVM is to find a hyperplane in the feature space to divide data into two categories, which makes the interval between classes maximal. An important method in SVM is called the kernel function. This implicitly maps low-dimension linearly indivisible data into high-dimensional space. By using SVM, an optimal separation hyperplane will be constructed in the high-dimensional feature space to make a better separation of data than in the low-dimension space. Another advantage of SVM is that it still can effectively classify cases where the feature dimension is larger than the number of samples. For protein sequences, the number of samples after screening is usually small, but the dimensions of the features constructed are generally much larger than the total number of samples. For the reason mentioned above, SVM was adopted as the classification algorithm in this work. A grid search method was used to optimize the regularization parameter *C* and kernel parameter γ through 5-fold cross-validation. The search spaces for *C* and γ are $[2^{15}, 2^{-5}]$ and $[2^{-5}, 2^{-15}]$, respectively.

Note that before using SVM to train the data, the experimental data must be normalized. Zero-mean normalization is the most common standardized method, also known as standard deviation standardization. The method is based on the mean μ and standard deviation σ of the original data to standardize the data. The distribution of processed data meets the standard normal distribution, which means the mean is 0 and the standard deviation is 1. Its conversion method is:

$$x^* = \frac{x - \mu}{\sigma} \quad (10)$$

3.5. Performance evaluation

In statistical forecasting, there are several methods to evaluate the model, such as using independent dataset test, *K*-fold cross-validation, and jackknife test. The independent test set refers to training the model using the training set while using the mutually independent test set to evaluate the quality of the model. Although this method seems simple, it also has certain defects. Since the training set and the test set are randomly divided, different divisions will produce different results owing to an unstable test result. Generally speaking, the larger the amount of data used to train the model, the better the trained model will generally be. The resulting model will be affected by using the test set for the un-fully use of data. Therefore, cross-validation method was proposed. *K*-fold cross-validation divides the entire data into *k* disjoint subsets, each of which is not repeated as a test set, and the other *k-1* copies are used as a training set to train the model each time. The results of *k* tests are combined to measure the performance of the model. In the jackknife test, all the samples in the benchmark dataset will be singled out one-by-one and tested by the predictor trained by the remaining

samples. Of the three test methods, specifically the jackknife test is deemed as the least arbitrary that can always yield a unique result for a given benchmark dataset as elaborated in [22]. Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of predictors [26], [42]–[46], [48]–[56]. However, to reduce the computational time, we adopted the 10-fold cross-validation in this study as done by many investigators with SVM as the prediction engine [57]–[65]. There are several kinds of evaluation metrics used to estimate the performance of the model. In the function prediction of protein sequences, researchers usually use accuracy (*Acc*), sensitivity (S_n), specificity (S_p), and Mathew's correlation coefficient (*MCC*), which is calculated by

$$\begin{cases} S_n = 1 - \dfrac{N_-^+}{N^+} \\ S_p = 1 - \dfrac{N_+^-}{N^-} \\ Acc = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} \\ MCC = \dfrac{1 - (\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-})}{\sqrt{(1 + \frac{N_+^- - N_-^+}{N^+})(1 + \frac{N_-^+ - N_+^-}{N^-})}} \end{cases} \quad (11)$$

Where N^+, N^- represents the number of positive and negative samples respectively. N_-^+, N_+^- represents the number of samples in which the positive sample is mistakenly classified into negative samples and the negative samples are mistakenly divided into positive samples. Compared with traditional metrics copied from math books, these metrics derived in [19], [23], [24] based on the Chou's symbols [66]–[68] are more intuitive and have been widely used among investigators [7], [11], [12], [15], [19], [20], [23], [28]–[39]. Note that, no matter the kind of metrics used, it is only valid only for the single-label systems, where each sample only belongs to one class. For the multi-label systems, where a sample may simultaneously belong to several classes, whose existence has become more frequent in system biology [13], [24], [40], [69], [70], system medicine [71], [72] and biomedicine [73], a completely different set of metrics as defined in [22] is needed. Of the aforementioned indicators, the most important are the *Acc* and *MCC*. The *Acc* reflects the overall accuracy of the model, and the *MCC* represents the reliability of the results of the algorithm. S_n, S_p can be seen as the recall rates of positive and negative categories.

To fully evaluate the predictive power of the model and display the results more vividly, the receiver operating characteristic curve (ROC curve) was used in this study. The ROC curve takes the true positive rate as the vertical axis and the false positive rate as the horizontal axis, depicting the trend of the model's prediction ability under all thresholds. AUC (Area Under Curve) is the area under the ROC curve, and the greater the value of AUC, the stronger the predictability of the model.

4. WEB-SERVER

As pointed out in [74] and demonstrated in a series of recent publications [9], [13]–[20], [23], [39], [41], [69], [70], [75], user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful prediction methods and computational tools. Actually, many practically useful web-servers have increasing impacts on medical science [25], driving medicinal chemistry into an unprecedented revolution [76]. We have also established a web-server called VirionPred for the identification method as mentioned in this paper and is shown in Figure 3. Users may access the web server at http://bigroup.uestc.edu.cn/virionpred/. The input of the web-server is a set of protein sequences in FASTA format, which can be either uploaded as a single file or copied/pasted into the input box. After submitting the protein sequences and click the submit button, results will be shown in a new interface.

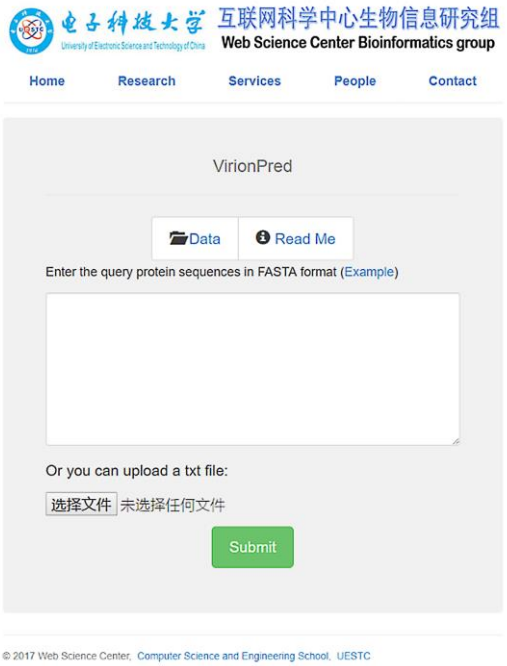

Figure 3: A semi-screenshot of the VirionPred webserver

5. CONCLUSION

Phage may provide a useful tool to find novel antibacterial drugs and understand the relationship between phage and host bacteria. Accurate identification of phage virion proteins from phage protein sequences is significant to understanding the complex virulence mechanism in host bacteria and the influence of bacteriophages on the development of antibacterial drugs. Although the performance of existing methods has been attested to have encouraging results, the accuracy of these methods is still far from satisfactory. The method proposed in this paper introduced a new feature construction method called the g-gap tripeptide composition. Combined with information gain & IFS feature selecting method and SVM algorithm, the method reached an accuracy up to 97.40% with MCC 0.9408 evaluated by 10-fold cross validation. The result outperforms other state-of-the-art methods upon. Our proposed method can be adopted as an improved method of identifying phage virion proteins and provide an effective feature constructing method for reference on other related works.

As future work, we will combine the g-gap TC and g-gap DC features to perform protein identification tasks. The combination of features of different gaps also meets our interests. Furthermore, we will work on other similar problems by using the g-gap TC to see the scope of application.

CONFLICT OF INTEREST

This research was funded by National Natural Science Foundation of China (No. 265 31771471, No. 21673034 and No. 71661167005.). The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

Conceptualization, Hui Gao, Zhen Liu and Lixia Tang; Methodology, Liangwei Yang, Hui Gao and Zhen Liu; Software, Songtao Li and Zhen Liu; Investigation, Liangwei Yang; Writing-Original Draft Preparation, Liangwei Yang and Hui Gao; Writing-Review & Editing, Liangwei Yang, Hui Gao; Supervision, Hui Gao, Zhen Liu and Lixia Tang; Funding Acquisition, Hui Gao, Zhen Liu and Lixia Tang.

REFERENCES

Journal Reference:

[1] A. Martelet, G. L'hostis, P. Tavares, S. Brasiles, F. Fenaille, C. Rozand, A. Theretz, G. Gervasi, J. Tablet, E. Ezan, "Bacterial detection using unlabeled phage amplification and mass spectrometry through structural and nonstructural phage markers," *J. Proteome Res.*, vol. 13, no. 3, pp. 1450–1465, 2014.

[2] P. V. Aguilar, A. P. Adams, E. Wang, W. Kang, A. S. Carrara, M. Anishchenko, L. Frolov, S. C. Weaver, "Structural and nonstructural protein genome regions of eastern equine encephalitis virus are determinants of interferon sensitivity and murine virulence," *J. Virol.*, vol. 82, no. 10, pp. 4920–4930, 2008.

[3] N. J. Moreland, M. YF. Tay, E. Lim, P. N. Paradkar, D. NP. Doan, Y. H. Yau, S. G. Shochat, S. G. Vasudevan, "High affinity human antibody fragments to dengue virus non-structural protein 3," *PLoS Negl. Trop. Dis.*, vol. 4, no. 11, p. e881, 2010.

[4] V. Seguritan, N. Alves, M. Arnoult, A. Raymond, D. Lorimer, A. B. Burgin Jr, P. Salamon, A. M. Segall, "Artificial neural networks trained to detect viral and phage structural proteins," *PLoS Comput. Biol.*, vol. 8, no. 8, p. e1002657, 2012.

[5] P.-M. Feng, H. Ding, W. Chen, and H. Lin, "Naive Bayes classifier with feature selection to identify phage virion proteins," *Comput. Math. Methods Med.*, vol. 2013, 2013.

[6] H. Ding, P.-M. Feng, W. Chen, and H. Lin, "Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis," *Mol. Biosyst.*, vol. 10, no. 8, pp. 2229–2235, 2014.

[7] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC," *J. Theor. Biol.*, vol. 377, pp. 47–56, Jul. 2015.

[8] F. Li, C. Li, T. Marquez-lago, A. Leier, T. Akutsu, A. W. Purcell, A. Smith, T. Lithgow, R. J. Daly, J. Song, K. C. Chou, "Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome.," *Bioinforma. Oxf. Engl.*, Jun. 2018.

[9] X. Cheng, X. Xiao, and K. C. Chou, "pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information," *Bioinformatics*, 2017.

[10] J. Song, F. Li, K. Takemoto, G. Haffari, T. Akutsu, K. C. Chou, G. Webb, "PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework," *J. Theor. Biol.*, vol. 443, pp. 125–137, Apr. 2018.

[11] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition," *J. Biomol. Struct. Dyn.*, vol. 34, no. 9, pp. 1946–1961, Sep. 2016.

[12] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets," *Molecules*, vol. 21, no. 1, p. 95, Jan. 2016.

[13] X. Cheng, X. Xiao, and K.-C. Chou, "pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC," *Genomics*, vol. 110, no. 1, pp. 50–58, Jan. 2018.

[14] B. Liu, F. Weng, D. S. Huang, and K. C. Chou, "iRO-3wPseKNC: Identify DNA replication origins by three-window-based PseKNC.," *Bioinformatics*, 2018.

[15] W. Chen, P. Feng, H. Yang, H. Ding, H. Lin, and K.-C. Chou, "iRNA-3typeA: Identifying Three Types of Modification at RNA's Adenosine Sites," *Mol. Ther. - Nucleic Acids*, vol. 11, pp. 468–474, Jun. 2018.

[16] B. Liu, F. Yang, D. S. Huang, and K. C. Chou, "iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC," *Bioinformatics*, vol. 34, no. 1, 2017.

[17] B. Liu, K. Li, D. S. Huang, and K. C. Chou, "iEnhancer-EL: Identifying enhancers and their strength with ensemble learning approach," *Bioinformatics*, 2018.

[18] P. Feng, H. Yang, H. Ding, H. Lin, W. Chen, and K.-C. Chou, "iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC," *Genomics*, Jan. 2018.

[19] W. Chen, P.-M. Feng, H. Lin, and K.-C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Res.*, vol. 41, no. 6, pp. e68–e68, 2013.

[20] J. Song, Y. Wang, F. Li, T. Akutsu, N. D. Rawling, G. I. Webb, K. C. Chou, "iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites.," *Brief. Bioinform.*, Apr. 2018.

[21] Z.-D. Su, Y. Huang, Z. Y. Zhang, Y. W. Zhao, D. Wang, W. Chen, K. C. Chou, H. Lin, "iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC.," *Bioinforma. Oxf. Engl.*, Jun. 2018.

[22] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, 2011.

[23] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, "UniProt: the Universal Protein knowledgebase.," *Nucl Acids Res*, vol. 32, no. 22, pp. D115–D119, 2004.

[24] I. Lobo, "Basic Local Alignment Search Tool (BLAST)," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, 2008.

[25] K. C. Chou, "Impacts of bioinformatics to medicinal chemistry.," *Med. Chem.*, vol. 11, no. 3, p. , 2015.

[26] H. Nakashima and K. Nishikawa, "Discrimination of Intracellular and Extracellular Proteins Using Amino Acid Composition and Residue-pair Frequencies," *J. Mol. Biol.*, vol. 238, no. 1, pp. 54–61, Apr. 1994.

[27] K.-C. Chou, "Prediction of Protein Cellular Attributes Using Pseudo- Amino Acid Composition," p. 11.

[28] M. Mandal, A. Mukhopadhyay, and U. Maulik, "Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC," *Med. Biol. Eng. Comput.*, vol. 53, no. 4, pp. 331–344, 2015.

[29] M. Arif, M. Hayat, and Z. Jan, "iMem-2LSAAC: A two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into chou's pseudo amino acid composition," *J. Theor. Biol.*, vol. 442, pp. 11–21, Apr. 2018.

[30] J. Mei and J. Zhao, "Analysis and prediction of presynaptic and postsynaptic neurotoxins by Chou's general pseudo amino acid composition and motif features.," *J. Theor. Biol.*, vol. 447, p. 147, 2018.

[31] S. M. Krishnan, "Using Chou's general PseAAC to analyze the evolutionary relationship of receptor associated proteins (RAP) with various folding patterns of protein domains," *J. Theor. Biol.*, vol. 445, pp. 62–74, 2018.

[32] M. S. Rahman, S. Shatabda, S. Saha, M. Kaykobad, and M. S. Rahman, "DPP-PseAAC: A DNA-binding protein prediction model using Chou's general PseAAC," *J. Theor. Biol.*, vol. 452, pp. 22–34, 2018.

[33] M. F. Sabooh, N. Iqbal, M. Khan, M. Khan, and H. F. Maqbool, "Identifying 5-methylcytosine sites in RNA sequence using composite encoding feature into Chou's PseKNC," *J. Theor. Biol.*, vol. 452, pp. 1–9, 2018.

[34] J. Mei and J. Zhao, "Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers," *Sci. Rep.*, vol. 8, no. 1, p. 2359, 2018.

[35] X. B. Zhou, C. Chen, Z. C. Li, and X. Y. Zou, "Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes," *J. Theor. Biol.*, vol. 248, no. 3, pp. 546–551, 2007.

[36] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *J. Theor. Biol.*, vol. 263, no. 2, pp. 203–209, 2010.

[37] L. Nanni, A. Lumini, D. Gupta, and A. Garg, "Identifying Bacterial Virulent Proteins by Fusing a Set of Classifiers Based on Variants of Chou's Pseudo Amino Acid Composition and on Evolutionary Information," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 9, no. 2, pp. 467–475, 2012.

[38] B. M. Mohammad, M. Behjati, and H. Mohabatkar, "Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach.," *J. Struct. Funct. Genomics*, vol. 12, no. 4, pp. 191–197, 2011.

[39] ZiaUrRehman and A. Khan, "Identifying GPCRs and their types with Chou's pseudo amino acid composition: an approach from multi-scale energy representation and position specific scoring matrix.," *Protein Pept. Lett.*, vol. 19, no. 8, p. , 2012.

[40] M.K. Gupta, R. Niyogi, and M. Misra, "An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition," *Sar Qsar Environ. Res.*, vol. 24, no. 7, p. 597, 2013.

[41] M. Khosravian, F. K. Faramarzi, M. M. Beigi, M. Behbahani, and H. Mohabatkar, "Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods," *Protein Pept. Lett.*, vol. 20, no. 2, p. , 2013.

[42] K.-C. Chou, "Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology," *Curr. Proteomics*, vol. 6, no. 4, pp. 262–274, 2009.

[43] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim, "Prediction of protein folding class using global description of amino acid sequence.," *Proc. Natl. Acad. Sci.*, vol. 92, no. 19, pp. 8700–8704, Sep. 1995.

[44] H. Wang and X. Hu, "Accurate prediction of nuclear receptors with conjoint triad feature," *BMC Bioinformatics*, vol. 16, no. 1, Dec. 2015.

[45] H.-Y. Lai, X.-X. Chen, W. Chen, H. Tang, and H. Lin, "Sequence-based predictive modeling to identify cancerlectins," *Oncotarget*, vol. 8, no. 17, Apr. 2017.

[46] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, "Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W65–W71, Jul. 2015.

[47] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[48] H. Lin, E.-Z. Deng, H. Ding, W. Chen, and K.-C. Chou, "iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition," *Nucleic Acids Res.*, vol. 42, no. 21, pp. 12961–12972, 2014.

[49] Y. Xu, X.-J. Shao, L.-Y. Wu, N.-Y. Deng, and K.-C. Chou, "iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins," *PeerJ*, vol. 1, p. e171, 2013.

[50] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, and A. Sattar, "Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou′ s general PseAAC," *J. Theor. Biol.*, vol. 364, pp. 284–294, 2015.

[51] W. Chen, P.-M. Feng, E.-Z. Deng, H. Lin, and K.-C. Chou, "iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition," *Anal. Biochem.*, vol. 462, pp. 76–83, 2014.

[52] Z. U. Khan, M. Hayat, and M. A. Khan, "Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model," *J. Theor. Biol.*, vol. 365, pp. 197–203, 2015.

[53] H. Ding, E. Deng, L. Yuan, H. Lin, W. Chen, K. C. Chou, "iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels," *BioMed Res. Int.*, vol. 2014, 2014.

[54] A. Anand and P. N. Suganthan, "Multiclass cancer classification by support vector machines with class-wise optimized genes and probability estimates," *J. Theor. Biol.*, vol. 259, no. 3, pp. 533–540, 2009.

[55] C. Chen, L. X. Chen, X. Y. Zou, and P. X. Cai, "Predicting protein structural class based on multi-features fusion," *J. Theor. Biol.*, vol. 253, no. 2, pp. 388–392, 2008.

[56] K. Chen, L. A. Kurgan, and J. Ruan, "Prediction of protein structural class using novel evolutionary collocation-based sequence representation," *J. Comput. Chem.*, vol. 29, no. 10, pp. 1596–1604, 2010.

[57] B. Park, J. Im, N. Tuvshinjargal, W. Lee, and K. Han, "Sequence-based prediction of protein-binding sites in DNA: comparative study of two SVM models.," *Comput. Methods Programs Biomed.*, vol. 117, no. 2, pp. 158–167, 2014.

[58] A. Rajput, A. K. Gupta, and M. Kumar, "Prediction and analysis of quorum sensing peptides based on sequence features," *Plos One*, vol. 10, no. 3, p. e0120066, 2015.

[59] Y. Xu, X. Wang, Y. Wang, Y. Tian, X. Shao, L. Wu, N. Deng, "Prediction of posttranslational modification sites from amino acid sequences with kernel methods," *J. Theor. Biol.*, vol. 344, no. 6, pp. 78–87, 2014.

[60] F. M. Pouzols, A. Lendasse, and A. B. Barros, "Autoregressive time series prediction by means of fuzzy inference systems using nonparametric residual variance estimation," *Fuzzy Sets Syst.*, vol. 161, no. 4, pp. 471–497, 2010.

[61] C. W. Tung, "POPISK: T-cell reactivity prediction using support vector machines and string kernels," *BMC Bioinformatics*, vol. 12, no. 1, pp. 446–446, 2011.

[62] B. A. Mckinney, D. M. Reif, M. T. Rock, K. M. Edwards, S. F. Kingsmore, J. H. Moore, C. J. Jr, "Cytokine Expression Patterns Associated with Systemic Adverse Events following Smallpox Immunization," *J. Infect. Dis.*, vol. 194, no. 4, pp. 444–453, 2006.

[63] W. Chen, L. Luo, and L. Zhang, "The organization of nucleosomes around splice sites," *Nucleic Acids Res.-01-05*, vol. 38, no. 9, pp. 2788–2798, 2010.

[64] Z. Huang, H. Chen, C. J. Hsu, W. H. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks: a market comparative study," *Decis. Support Syst.*, vol. 37, no. 4, pp. 543–558, 2004.

[65] F. Ali and M. Hayat, "Classification of membrane protein types using Voting Feature Interval in combination with Chou's Pseudo Amino Acid Composition," *J. Theor. Biol.*, vol. 384, pp. 78–83, 2015.

[66] K.-C. Chou, "Prediction of protein signal sequences and their cleavage sites," *Proteins Struct. Funct. Bioinforma.*, vol. 42, no. 1, pp. 136–139, 2001.

[67] K.-C. Chou, "Using subsite coupling to predict signal peptides," *Protein Eng.*, vol. 14, no. 2, pp. 75–79, 2001.

[68] K.-C. Chou, "Prediction of signal peptides using scaled window," *peptides*, vol. 22, no. 12, pp. 1973–1979, 2001.

[69] H. Mohabatkar, M. M. Beigi, K. Abdolahi, and S. Mohsenzadeh, "Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach," *Med. Chem.*, vol. 9, no. 1, p. , 2013.

[70] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, Jan. 2005.

[71] X. Cheng, S.-G. Zhao, X. Xiao, and K.-C. Chou, "iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals," *Bioinformatics*, vol. 33, no. 3, pp. 341–346, 2016.

[72] X. Cheng, S.-G. Zhao, X. Xiao, and K.-C. Chou, "iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals," *Oncotarget*, vol. 8, no. 35, p. 58494, 2017.

[73] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, and K.-C. Chou, "iPTM-mLys: identifying multiple lysine PTM sites and their different types," *Bioinformatics*, vol. 32, no. 20, pp. 3116–3123, 2016.

[74] K.-C. Chou and H.-B. Shen, "Recent advances in developing web-servers for predicting protein attributes," *Nat. Sci.*, vol. 1, no. 02, p. 63, 2009.

[75] P. Du, S. Cao, and Y. Li, "SubChlo: Predicting protein subchloroplast locations with pseudo-amino acid composition and the evidence-theoretic K -nearest neighbor (ET-KNN) algorithm," *J. Theor. Biol.*, vol. 261, no. 2, pp. 330–335, 2009.

[76] K.-C. Chou, "An unprecedented revolution in medicinal chemistry driven by the progress of biological science," *Curr. Top. Med. Chem.*, vol. 17, no. 21, pp. 2337–2358, 2017.