OPEN ACCESS Freely available online

PLoS one

Prediction of Antimicrobial Peptides Based on Sequence Alignment and Feature Selection Methods

Ping Wang[1], Lele Hu[2,3], Guiyou Liu[1], Nan Jiang[1], Xiaoyun Chen[1], Jianyong Xu[1], Wen Zheng[1], Li Li[1], Ming Tan[1], Zugen Chen[1,4], Hui Song[1]*, Yu-Dong Cai[2,3,5]*, Kuo-Chen Chou[5]

1 Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China, **2** Institute of Systems Biology, Shanghai University, Shanghai, China, **3** Department of Chemistry, College of Sciences, Shanghai University, Shanghai, China, **4** Department of Human Genetics, University of California Los Angeles, Los Angeles, California, United States of America, **5** Gordon Life Science Institute, San Diego, California, United States of America

Abstract

Antimicrobial peptides (AMPs) represent a class of natural peptides that form a part of the innate immune system, and this kind of 'nature's antibiotics' is quite promising for solving the problem of increasing antibiotic resistance. In view of this, it is highly desired to develop an effective computational method for accurately predicting novel AMPs because it can provide us with more candidates and useful insights for drug design. In this study, a new method for predicting AMPs was implemented by integrating the sequence alignment method and the feature selection method. It was observed that, the overall jackknife success rate by the new predictor on a newly constructed benchmark dataset was over 80.23%, and the Mathews correlation coefficient is 0.73, indicating a good prediction. Moreover, it is indicated by an in-depth feature analysis that the results are quite consistent with the previously known knowledge that some amino acids are preferential in AMPs and that these amino acids do play an important role for the antimicrobial activity. For the convenience of most experimental scientists who want to use the prediction method without the interest to follow the mathematical details, a user-friendly web-server is provided at http://amp.biosino.org/.

Citation: Wang P, Hu L, Liu G, Jiang N, Chen X, et al. (2011) Prediction of Antimicrobial Peptides Based on Sequence Alignment and Feature Selection Methods. PLoS ONE 6(4): e18476. doi:10.1371/journal.pone.0018476

Editor: Vladimir N. Uversky, University of South Florida College of Medicine, United States of America

Received December 5, 2010; **Accepted** March 8, 2011; **Published** April 13, 2011

Funding: This research was supported by grants from the Tianjin Science and Technology Support Program (10ZCZDSY06400, 10ZCKFSY05500), One Hundred Person Project of the Chinese Academy of Sciences (KSCX2-YW-BR-3), and the National Basic Research Program of China, 2011CB510102, 2011CB510101. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: song_h@tib.cas.cn (HS); cai_yud@yahoo.com.cn (Y-DC)

Introduction

Natural gene-encoded antimicrobial peptides (AMPs) are a group of small, innate immune molecules, generally containing 12–100 amino acid residues [1]. AMPs have been discovered in most life forms, including bacteriocins, fungal peptide antibiotics, plant thionins and defensins, insect defensins and cecropins, amphibian magainins and temporis, as well as defensins and cathelicidins from higher vertebrates [1,2,3]. Owing to the broad spectrum antimicrobial activity [4,5], antibacteria, antifungi, antivirus, and even anticancer, are thought to be less likely to induce resistance. Thus, AMPs have attracted the attention of many investigators as a substitute for conventional antibiotics [1]. Currently, most researchers in this area are focused on screening and in silico modeling novel AMPs [6,7] as computational approaches can accelerate the process of antimicrobial drug discovery and design [8]. Many bioinformatics methods have been developed for predicting new AMPs. For example, the APD method predicted whether the new peptide had the potential to be antimicrobial based on some known principles [9]. The AMPer method [10] was developed by constructing the hidden Markov models (HMMs) to automatically discover AMPs. The BACTIBASE [11,12] and PhytAMP [13] methods were specifically designed for bacteriocin and plant respectively. The AntiBP method [14] and AntiBP2 method [15] used the Artificial Neural Network (ANN), Quantitative Matrices (QM) and Support Vector Machine (SVM) to predict antibacterial peptides. Their training sets were limited to N and/or C terminus residues of peptides. The CAMP method [16] was developed based on the Random Forests (RF), SVM, and Discriminant Analysis (DA), trained on all classes of AMPs (antibacterial, antifungal and antiviral) and full length of mature AMP sequences. However, none of the aforementioned methods has the function to identify which kinds of features are optimal for accurately predicting and meaningfully interpreting their biological implications.

The present study was initiated in an attempt to establish a new classification method for predicting AMPs by integrating the sequence alignment method and the feature selection method. In the sequence alignment method, the prediction was carried out by assigning the query peptide to the category of the peptide that has the highest sequence similarity with the query peptide. In the feature selection method, each peptide was coded with 270 features, including amino acid composition [17,18] and pseudo-amino acid composition [19] that incorporated electrostatic charge, codon diversity, molecular volume, polarity, and secondary structure [20]. Subsequently, the feature selection and analysis methods, including the Maximum Relevance Minimum Redundancy method (mRMR) [21] and the Incremental Feature Selection (IFS) [22] method, were employed to select the optimal features for the prediction of AMPs versus non-AMPs. The

prediction model was built using the well-known Nearest Neighbor Algorithm (NNA) [23,24,25]. As a result, the methods achieved a satisfactory overall success rate.

Materials and Methods

Datasets

Training set. The AMP sequences were downloaded from CAMP [16]. The 1,216 AMP sequences validated by experiments and the 1,651 AMP sequences filed with patents were used. After eliminating those sequences with non-standard residues 'B', 'J', 'O', 'U', 'X', or 'Z', the final positive dataset contained 2,752 AMP sequences, of which only 35 peptides in UniPort database [26,27] are annotated with experimentally-verified no antimicrobial activity. Because AMPs are generally secretory in nature [28], we also randomly selected 10,000 non-secretory protein sequences from UniProt database without annotated by 'antimicrobial'. Since most of the AMPs in positive dataset are with 10–80 amino acids, we randomly cut out a fragment with the same length range from each sequence and added them to the negative dataset. After eliminating those sequences with non-standard residues 'B', 'J', 'O', 'U', 'X', or 'Z', the final negative dataset thus obtained contained 10,014 non-AMP sequences.

Test set. CAMP [16] predicted dataset contained 1,153 sequences identified as antimicrobial based on the evidences of similarity or annotations in NCBI as 'antimicrobial regions' without exprerimental evidences. After eliminating those sequences containing non-standard residues 'B', 'J', 'O', 'U', 'X', or 'Z', 1,136 sequences were left that will serve as independent positive test dataset. As mentioned above, only 35 peptides are experimentally-verified no antimicrobial activity, and we had used these peptides as negative samples in the training dataset. Therefore, there were no more peptides left that could be used as independent negative samples for the test dataset in this study.

Cutoff threshold for sequence identity. Generally, homologous sequences in the datasets often influence the performance of the predictors. In order to remove the homologous peptides inside the training dataset and between the training and test datasets, a cutoff threshold of 70% was imposed to exclude those peptides from the training set that have equal to or greater than 70% sequence identity to any other in the training/test set by the CD-HIT program [29]. As a result, the training set thus obtained contained 9731 sequences, including 870 AMPs and 8661 non-AMPs.

It is known to us that the peptide's function is strongly related to its sequence order. Therefore we first apply the sequence alignment algorithm to predict AMPs. Secondly, we use amino acid composition and pseudo amino acid composition which can approximately reflect the sequence order [30], to deal with those peptides which can't be performed by the sequence alignment method.

Sequence alignment method

Sequence alignment is a very important problem in Bioinformatics [31]. The sequences segments with high identify are inclined to share the structure and function. In the past decades, various sophisticated method such as FASTA, BLAST, HMMER and Smith-Waterman algorithm [32,33,34,35] were developed for local and global alignments for DNA and protein sequences. Here, BLASTP [36] was used to predict AMPs, which can be described as follows. First, let us suppose a query peptide P and the training set $\{P_1,P_2,...,P_n\}$, then the high-scoring segment pairs (HSPs) score between the query peptide and each peptide in the training set are calculated by BLASTP with default parameters. Then the peptide is predicted to share the same category as the peptide P_k if the HSP score between P and P_k is higher than other scores. Expressed in a formula, P_k subjects to

$$\text{HSPs Score (P, P}_\text{k}) = \max\{\text{HSPs Score(P, P}_\text{i})|\text{i} = 1, 2, ..., n\} \quad (1)$$

If more than one P_k fulfils the Eq. (1), one of them is chosen at random and its category was assigned to the query peptide P.

Feature selection method

In this research, amino acid composition and pseudo-amino acid composition were used to code the AMP sequences.

Amino acid composition. Amino acid composition is a basic feature of protein sequence [25], which is closely correlated with its attributes, such as subcellular location [37,38,39,40,41], folding type [17,42], secondary structure content [43], and domain [44]. Amino acid composition consists of 20 discrete numbers, each of which represents the occurrence frequency of the native amino acid in a protein sequence. Therefore, the protein can be coded into a 20-D (dimensional) numerical vector by the amino acid composition.

Pseudo-amino acid composition. The concept of pseudo-amino acid composition (PseAAC) was originally introduced by Chou for predicting the protein subcellular locations and membrane protein types [19]. Based on the conventional amino acid composition, Chou proposed a set of discrete numbers to take into account some sequence order effects. PseAAC has been proved to be an extremely effective feature in treating many protein and protein-related systems (see, e.g., [45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72] as well as the Wikipedia web page at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition). For the detailed description about PseAAC, refer to [19,73] and a recent comprehensive review [74]. Here, for reader's convenience, the concept of PseAAC is briefly described as follows.

Suppose a protein sequence of L amino acid residues:

$$\mathrm{R_1R_2R_3\cdots R_{L-2}R_{L-1}R_L} \quad (2)$$

The sequence order effect of the protein can be reflected by a set of discrete correlation factors, which are calculated as follows:

$$\begin{cases} \theta_1 = \dfrac{1}{L-1}\displaystyle\sum_{i=1}^{L-1}\Theta(R_i,R_{i+1}) \\ \theta_2 = \dfrac{1}{L-2}\displaystyle\sum_{i=1}^{L-2}\Theta(R_i,R_{i+2}) \\ \theta_3 = \dfrac{1}{L-3}\displaystyle\sum_{i=1}^{L-3}\Theta(R_i,R_{i+3}) \\ \cdots \\ \theta_\lambda = \dfrac{1}{L-\lambda}\displaystyle\sum_{i=1}^{L-\lambda}\Theta(R_i,R_{i+\lambda}) \quad (\lambda < L) \end{cases} \quad (3)$$

where θ_1, θ_2, θ_3, θ_λ are the first-tier, second-tier, third-tier, λ-th tier correlation factors. And the correlation function is

$$\Theta(R_i,R_j) = [F(R_j) - F(R_i)]^2 \quad (4)$$

where $F(R_i)$ is the feature (e.g. hydrophilicity) value of the amino acid R_i. The value is converted from the original feature value of the amino acid according to the following equation:

$$F(R_i)=\frac{F_o(R_i)-\sum_{i=1}^{20}\frac{F_o(R_i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20}\left[F_o(R_i)-\sum_{i=1}^{20}\frac{F_o(R_i)}{20}\right]^2}{20}}} \quad (5)$$

where $F_o(R_i)$ is the original feature value of the amino acid R_i. Thus, the PseAAC of a protein can be represented by a (20+λ)-D vector as follows:

$$V=[v_1,v_2,\cdots,v_{20},v_{21},\cdots,v_{20+\lambda}]^{\mathrm{T}} \quad (6)$$

where superscript **T** is the transpose operator and

$$v_x=\begin{cases}\dfrac{f_x}{\sum_{i=1}^{20}f_i+\omega\sum_{j=1}^{\lambda}\theta_j}, & (1\le x\le 20)\\ \dfrac{\omega\theta_{x-20}}{\sum_{i=1}^{20}f_i+\omega\sum_{j=1}^{\lambda}\theta_j}, & (21\le x\le 20+\lambda)\end{cases} \quad (7)$$

where $f_x(x=1,2,...,20)$ represent the occurrence frequencies of the 20 amino acids in the protein sequence, θ_j represents the *j-th* tier sequence correlation factor calculated according to Eq. (3), and ω represents the weight for the sequence order effect. Based on the above description, we know that the first 20 components in Eq. (6) reflect the effect of the conventional amino acid composition, while the remaining λ components are the correlation factors reflecting the effect of sequence order. A set of such 20+λ numbers is named PseAAC. In this study, we chose $\omega=0.15$ and $\lambda=50$ for getting the optimal results.

In this study, the codon diversity, electrostatic charge, molecular volume, polarity, and secondary structure are used to describe the physicochemical and biochemical properties of amino acids. And the values of the 5 features of the amino acids are retrieved from [20,75,76], as shown in **Table 1**. For each of the five features, a set of discrete correlation factors can be calculated according to Eq. (3) and Eq. (4) so as to contribute $\lambda=50$ additional components for defining the protein sequence according to Eq. (6). Likewise, the similar approach can also be used to code the AMPs.

Since each of the aforementioned five features (cf. **Table 1**) can generate $\lambda=50$ discrete numbers, the AMPs will be defined in a $(20+50\times5=270)$-D vector space.

In the feature space, we firstly prioritized the 270 features by the Maximum Relevance, Minimum Redundancy (mRMR) method. Based on the feature order, Incremental Feature Selection (IFS) method was employed to select the optimal feature subset. The prediction model was constructed according to Nearest Neighbor Algorithm (NNA) and evaluated by the jackknife test.

mRMR method. In pattern recognition, feature selection is an important procedure for constructing the classifier. Generally, a "good" feature for classification is considered to be not only highly correlated to the class, but also lowly redundant to the already selected features. Here, the Maximum Relevance, Minimum Redundancy [21] (mRMR) method was employed to sort the 270 features according to the descending order. The key ideas of the method are the Maximum Relevance criterion and Minimum Redundancy criterion as meant by its name. According to the Maximum Relevance criterion, the feature to be selected should have the maximal correlation with the class variable; while according to the Minimum Redundancy criterion, the feature to be selected should have minimal redundancy to the already selected features. Features are selected from the 270-D feature space one by one, being put into the MaxRel feature list by applying the Maximum Relevance criterion, and being put into the mRMR feature list by applying both the criteria. Both the relevance and redundancy are quantified by the mutual information (MI) defined as follows

$$I(x,y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \quad (8)$$

where $p(x,y)$ is the joint probabilistic density for feature x and feature y, $p(x)$ and $p(y)$ are the marginal probabilistic densities for feature x and feature y, respectively.

Suppose the whole feature set is denoted by Ω, the already selected feature set with m features by Ω_s and the feature set with n features by Ω_t. The relevance D between the feature f in set Ω_t and the class c is calculated by

$$D=I(f,c) \quad (9)$$

The redundancy R of f with all the features in Ω_s is calculated by

$$R=\frac{1}{m}\sum_{f_i\in\Omega_s}I(f,f_i) \quad (10)$$

To select the feature f_i in set Ω_t with the maximum relevance and minimum redundancy to already selected features in set Ω_s, Eq. (9) and Eq. (10) are combined to generate the function:

Table 1. The physicochemical and biochemical properties of the 20 amino acids.

Amino Acid	Polarity	Secondary structure	Molecular volume	Codon diversity	Electrostatic charge
A	−0.591	−1.302	−0.733	1.57	−0.146
C	−1.343	0.465	−0.862	−1.02	−0.255
D	1.05	0.302	−3.656	−0.259	−3.242
E	1.357	−1.453	1.477	0.113	−0.837
F	−1.006	−0.59	1.891	−0.397	0.412
G	−0.384	1.652	1.33	1.045	2.064
H	0.336	−0.417	−1.673	−1.474	−0.078
I	−1.239	−0.547	2.131	0.393	0.816
K	1.831	−0.561	0.533	−0.277	1.648
L	−1.019	−0.987	−1.505	1.266	−0.912
M	−0.663	−1.524	2.219	−1.005	1.212
N	0.945	0.828	1.299	−0.169	0.933
P	0.189	2.081	−1.628	0.421	−1.392
Q	0.931	−0.179	−3.005	−0.503	−1.853
R	1.538	−0.055	1.502	0.44	2.897
S	−0.228	1.399	−4.76	0.67	−2.647
T	−0.032	0.326	2.213	0.908	1.313
V	−1.337	−0.279	−0.544	1.242	−1.262
W	−0.595	0.009	0.672	−2.128	−0.184
Y	0.26	0.83	3.097	−0.838	1.512

Listed below are the scores of the physicochemical and biochemical properties of the 20 amino acids, each of which can be coded by a 5-dimensional vector.
doi:10.1371/journal.pone.0018476.t001

$$\max_{f_j \in \Omega_t} \left[I(f_j,c) - \frac{1}{m} \sum_{f_i \in \Omega_s} I(f_j, f_i) \right] (j = 1,2,...,n) \quad (11)$$

Subsequently, the selected feature f_i will be taken away from the set Ω_t and added into the set Ω_s. Such a process will be repeated until all the features are taken away from the set Ω_t and added into the set Ω_s. The better the feature is, the earlier it will be selected.

Nearest Neighbor Algorithm. Nearest Neighbor Algorithm (NNA) [23] is a simple and effective instance-based learning method. It assigns the unknown sample to the class of the nearest neighbor. The distance function, the core of the algorithm, can be defined as follows [68]:

$$D(v_i, v_j) = 1 - \frac{v_i \cdot v_j}{||v_i|| \cdot ||v_j||} \quad (12)$$

where the symbol $||v||$ stands for the vector module of the sample, and $v_i \cdot v_j$ stands for the dot product of the two coding vectors.

Suppose a queried peptide with the 270-D coding vector p and the training set comprised of n classified peptides with the coding vector set $\{p_1, p_2, ..., p_i, ..., p_n\}$ respectively. Then the queried peptides will be assigned to the class of vector p_m, which satisfies

$$D(p, p_m) = \min\{D(p, p_i) | (i = 1,2,...,n)\} \quad (13)$$

If more than one p_m satisfies to Eq. (9), the class of one of these peptides will be randomly selected as the predicted result for the queried peptide.

Incremental Feature Selection. In essence, feature selection is a combinatorial optimization problem. Its goal is to seek the feature subset that maximizes the performance of the predictor. To find the optimal feature subset from the feature space with N features, all the combinations of N features should be tried from the point of view of the exhaustion principle, which is of computational intractability. Therefore Incremental Feature Selection [76,77] (IFS) method was utilized to get the approximate solutions for this problem.

Based on features prioritized in the mRMR feature list, 270 feature subsets were obtained according to

$$S_i = \{f_1, f_2, ..., f_i\} (1 \le i \le 270) \quad (14)$$

where f_i is the *i-th* feature in the mRMR feature list.

Then a NNA predictor was constructed for each feature subset and evaluated by the jackknife test. With the number of features of subset S_i as its x-axis and accuracy as its y-axis, IFS curve was plotted to reveal the relation between the performance of the NNA predictor and the feature subset. The optimal feature subset is considered with the highest prediction accuracy, and the predictor thus obtained was used to classify the peptides.

Overall prediction

For a query peptide, BLAST method was first applied to estimate whether it has antimicrobial activity. If it did not have any hits against the training sequences, then the Feature selection method was applied.

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its anticipated accuracy: independent dataset test, subsampling (K-fold cross-validation) test, and jackknife test [78]. In this study the jackknife test was adopted to examine the quality of the current predictor. During the jackknifing process, each of the peptide samples was in turn singled out from the benchmark dataset as a test sample, and identified by the prediction engine trained by the rest of the peptide samples in the dataset.

The following equations were often used in literatures to reflect the prediction quality:

$$\begin{aligned} S_n &= \frac{TP}{TP + FN} \\ S_p &= \frac{TN}{TN + FP} \\ AC &= \frac{TP + TN}{TP + FP + TN + FN} \\ MCC &= \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \end{aligned} \quad (15)$$

where S_n reflects the sensitivity, S_p the specificity, AC the accuracy, and MCC the Mathews correlation coefficient; while TP represents the true positive, TN, the true negative; FP, the false positive, and FN, the false negative (**Figure 1**). S_n, S_p and AC stand for the success rates of prediction on positive, negative and overall datasets, respectively. MCC is used to evaluate the performance of the predictor when the positive and negative samples in the dataset are out-of-balance. Its value ranges from -1 to 1, and a larger MCC means a better prediction.

Results and Discussion

Results of sequence alignment method

In the jackknife cross-validation, each peptide was singled out from the benchmark data set as the query peptide, and the remaining peptides would serve as the training data set to train the predictor. Then the BLASTP method was applied to classify the peptide according to Eq. (1). However, some query peptides could not be processed by the method because no hits at all were found between them and the peptides in the training dataset. Among the

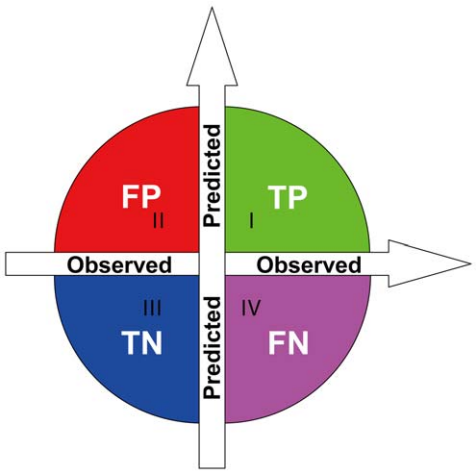

Figure 1. An illustration to show (I) TP (true positive) quadrant (green) for correct prediction of positive dataset, (II) FP (false positive) quadrant (red) for incorrect prediction of negative dataset; (III) TN (true negative) quadrant (blue) for correct prediction of negative dataset; and (IV) FN (false negative) quadrant (pink) for incorrect prediction of positive dataset.
doi:10.1371/journal.pone.0018476.g001

9731 peptides in the benchmark data set, 5855 peptides were predicted by the BLAST. The predicted results were shown in **Table 2**. The S_n, S_p, AC, and MCC were 91.22%, 95.55%, 95.12%, and 0.7723, respectively.

Results of feature selection method

As the sequence alignment method could not deal with all the peptides, we designed the feature selection method to classify the remaining 3876 ($3876 = 9731 - 5855$) peptides.

Here, the prediction model was constructed as follows. All the peptides in the benchmark data set were firstly represented by the 270 features retrieved from the amino acid composition and pseudo-amino acid composition. The mRMR program (http://penglab.janelia.org/proj/mRMR/index.htm) was then applied to prioritize the features according to the Maximum Relevance criterion and Minimum Redundancy criterion. The MaxRel feature list and mRMR feature list thus obtained can be found in Table S1 and Table S2, respectively. Based on the sorted feature in mRMR feature list, the 270 feature subsets were constructed according to Eq. (14). Each of the feature subsets was used to recode the peptides in the dataset and construct the prediction model according to NNA. The prediction accuracies of the NNA predictor evaluated by jackknife test are shown in the IFS curve (**Figure 2**). It was observed that the peak of the accuracy was corresponding to the number of features at 25. Hence, the optimal feature subset was obtained with the first 25 features in the mRMR feature list. Therefore the predictor with these 25 features was used to cope with the 3876 peptides. The predicted results were also shown in **Table 2**. The S_n, S_p, AC, and MCC were 56.83%, 93.19%, 90.58%, and 0.6426, respectively.

The overall predicted results

By combining the results of prediction from sequence alignment method and sequence based method, the overall success rates for the benchmark data set were obtained, as shown in **Table 2**. Evaluated by jackknife test, the S_n, S_p, AC, and MCC were 80.23%, 94.59%, 93.31%, and 0.7312, respectively, indicating a good prediction from the integration of the two methods. From the table, we can see that although BLASTP method obtained good predicted results, it could not deal with all the peptides. As a fall-back, the feature selection method was used to process the remaining peptides. By integrating the two methods, the hybrid one leads to satisfactory results.

Independent test and comparison with the existing predictors

Generally speaking, the independent dataset is used for demonstrating how to use the predictor for practical applications

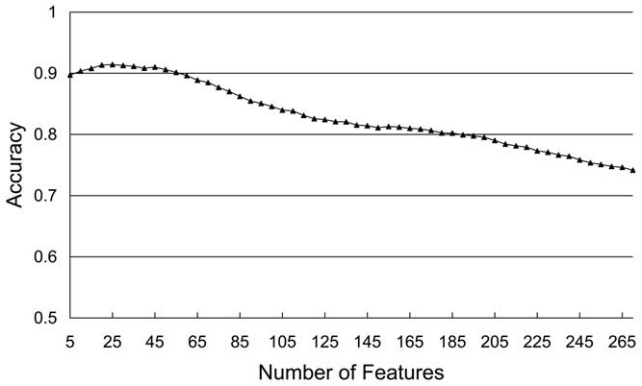

Figure 2. IFS curve. It reveals the relation between the performance of the NNA predictor and the feature subset. The IFS curve arrives at the apogee when the feature set is comprised of the first 25 features in the mRMR feature list.
doi:10.1371/journal.pone.0018476.g002

[37]. This is because each of the peptides singled-out from the benchmark data set during the jackknifing process can actually be deemed as a sample of an independent data set. Now, just as a demonstration, let us use the benchmark dataset as a training dataset to identify the 1,136 AMP sequences collected in the independent dataset. The prediction sensitivity thus obtained with the integrated method was 72.27%, somewhat lower than the rate of jackknife test S_n, this may because some AMPs in the test set were derived according to the annotations in NCBI based on the similarity principle and hence cannot avoid some sort of arbitrariness or false positive.

Up to now, several computational methods [10,11,12,13, 14,15,16] have been proposed for the predicting AMPs. However, AMPer method [10] is not available at http://www.cnbi2.com/cgi-bin/amp.pl as described in [10]. BACTIBASE [11,12] and PhytAMP [13] methods were specifically designed for bacteriocin and plant respectively. As for AntiBP [14] and AntiBP2 methods [15], they were designed for identifying the AMPs in a protein sequence, and hence could not be used to compare with our method. To make the comparison meaningful, our method was compared with CAMP method [16], which was developed based on the Random Forests (RF), SVM, and Discriminant Analysis (DA). In the comparison, the original 2,752 AMPs and 10,014 non-AMPs were treated as the training set. This is because to make the predictor better, nornally all the training samples need to be used. The comparison results are shown in the **Table 3**. The prediction S_n by our method was 84.95%, higher than the predicted results of CAMP, indicating that our method outperformed CAMP.

Table 2. The predicted results of the three methods.

Method	Number of Predicted Peptides	S_n (%)	S_p (%)	AC (%)	MCC
Sequence Alignment Method	5855	91.22	95.55	95.12	0.7723
Feature selection Method	3876	56.83	93.19	90.58	0.6426
Integrated Method	9731	80.23	94.59	93.31	0.7312

doi:10.1371/journal.pone.0018476.t002

Table 3. Comparison between CAMP and our method on the test set.

Method	Algorithm	Predicted AMPs	S_n (%)
CAMP	Support Vector Machine	866	76.23
CAMP	Random Forest	852	75.00
CAMP	Discriminant Analysis	881	77.55
Our Method	BLASTP+Nearest Neighbor Algorithm	965	84.95

doi:10.1371/journal.pone.0018476.t003

Table 4. Comparison between sequence alignment method and feature selection method.

Dataset	Method	Number of Predicted Peptides	Number of Correctly Predicted Peptides	S_n (%)
Original Dataset with high sequence similarity	Sequence Alignment	986	896	90.87
	Feature Selection	1136	791	69.63
Dataset with <0.7 sequence similarity	Sequence Alignment	869	679	78.14
	Feature Selection	1136	692	60.92

doi:10.1371/journal.pone.0018476.t004

Comparison between sequence alignment method and feature selection method

In this study, sequence alignment method and feature selection method were developed to identify the AMPs from peptides. To compare the performance between them, each method was used alone to predict the peptides in the test set. To investigate the effect of sequence homology on the performance of the methods, original dataset (2,752 AMPs and 10,014 non-AMPs) and the dataset <0.7 sequence similarity were used. The predicted results are shown in **Table 4**.

From the table, we can see that the prediction S_n by sequence alignment method is much higher than the S_n by feature selection method. However, the sequence alignment could not deal with all the 1136 peptides in the test set. The sequence alignment method has the high predicted accuracies, while the feature selection method can predict all the peptides. To utilize the two advantages, the two methods were integrated to predict AMPs as above mentioned. The accuracies dropped by about 10% from the original dataset to dataset with <0.7 sequence similarity, which indicates sequence homology influenced the predictive quality.

Analysis of optimal features

Among the 25 optimal features obtained from the feature selection method, the one for the amino acid composition took up 64% (**Figure 3**). In the previous works, except for the simple and linear AMPs, larger AMPs are prone to contain certain amino acid types, such as cysteine, proline, arginine, tryptonphan, and histidine [79]. These five amino acids are all in our optimal features. Actually, according to our results, cysteine, arginine, tryptonphan and histidine are rich in antimicrobial peptides (**Figure 4**), fully consistent with the findings in [79], while proline is not obviously different between antimicrobial and non-antimicrobial peptides. Our results further confirm that amino acid composition is important for identify whether a peptide is an effector molecules of immunity. According to the ranks of these features, cysteine is the second one. Cysteine-rich peptides are particularly typical in plants [80,81] and animals [82]. Pairs of cysteines forming intramolecular disulfide bridged are common in AMPs, thus allowing a complex three-dimensional structure, such as β-sheet [83] and β-turn [84]. Arginine, lysine and histidine are also important amino acid component features in our result. Arginine, lysine, and histidine in acidic environments are with positive net charged [85]. Meanwhile, the negative charged amino acids, glutamic acid and aspartic acid, are lack in AMPs (**Figure 4**). This may help AMPs to flip into biological membranes owing to the anionic phospholipid membranes [86]. Another AMP-rich amino acid is tryptophan. It is important for lipid binding [87,88]and preferential in the protein-membrane interface [89]. The secondary structures, codon diversity as well as polarity of AMPs would ensure their abilities to defend microorganisms. All these effects may help AMPs disrupt the microbial membranes integrity.

Conclusion

In this study, two methods are implemented: the sequence alignment method based on the BLASTP and the feature selection

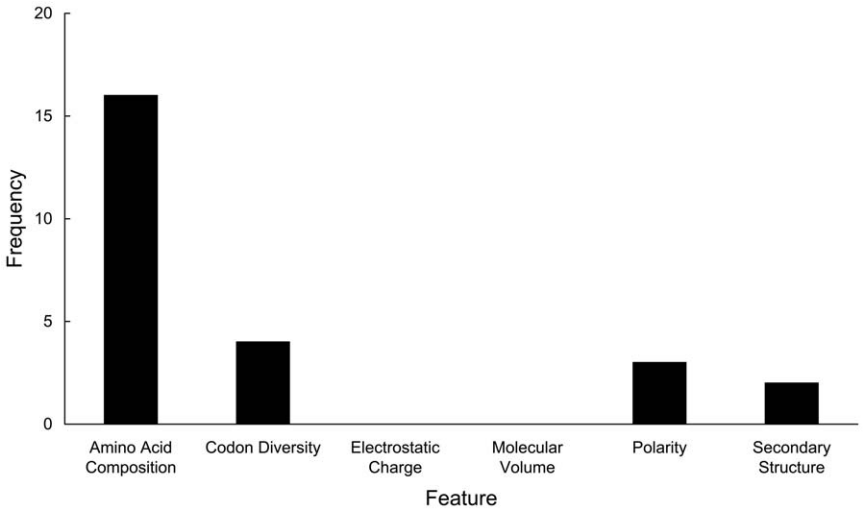

Figure 3. The numbers of each kind of features in optimal features. In the feature space, all the features can be classified into six kinds: amino acid composition, codon diversity, electrostatic charge, molecular volume, polarity and secondary structure.
doi:10.1371/journal.pone.0018476.g003

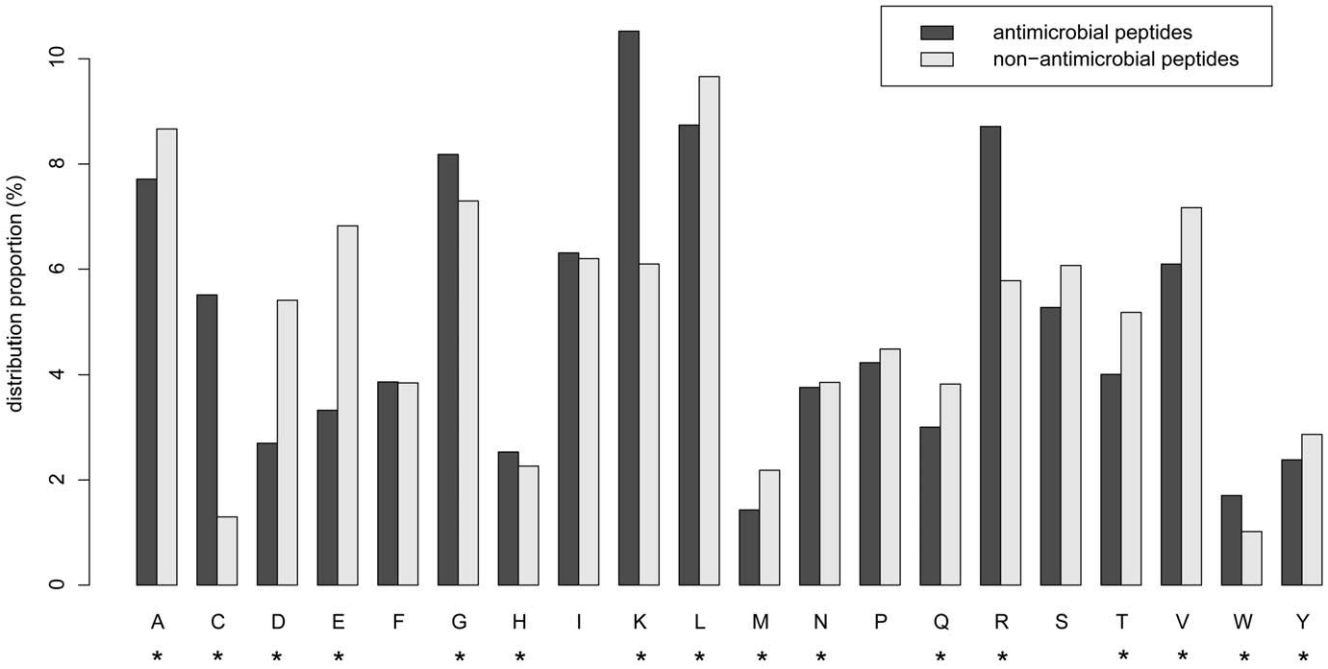

Figure 4. Amino acid distribution in AMPs and non-AMPs. * indicates amino acid in the optimal feature set.
doi:10.1371/journal.pone.0018476.g004

method with amino acid composition and pseudo amino acid composition features [90]. The prediction accuracy of the integrated method on the benchmark dataset is 80.23%. It is anticipated that the new method may be of use for helping to understand the role of peptide in antimicrobial activity, identify the natural AMPs, and design the synthetic AMPs against the resistance of microorganisms to antibiotics. For the convenience of readers, a user-friendly web-server is freely accessible at http://amp.biosino.org/.

Supporting Information

Table S1 The MaxRel feature list.
(DOC)

Table S2 The mRMR feature list.
(DOC)

Acknowledgments

We thank CAMP for supplying data to support not-for-profit research efforts.

Author Contributions

Conceived and designed the experiments: YDC. Performed the experiments: YDC LH PW. Analyzed the data: YDC LH PW. Contributed reagents/materials/analysis tools: PW LH. Wrote the paper: LH PW YDC KCC.

References

1. Sang Y, Blecha F (2008) Antimicrobial peptides and bacteriocins: alternatives to traditional antibiotics. Anim Health Res Rev 9: 227–235.
2. McPhee JB, Hancock RE (2005) Function and therapeutic potential of host defence peptides. J Pept Sci 11: 677–687.
3. Yeaman MR, Yount NY (2007) Unifying themes in host defence effector polypeptides. Nat Rev Microbiol 5: 727–740.
4. Epand RM, Vogel HJ (1999) Diversity of antimicrobial peptides and their mechanisms of action. Biochim Biophys Acta 1462: 11–28.
5. Kamysz W, Okroj M, Lukasiak J (2003) Novel properties of antimicrobial peptides. Acta Biochim Pol 50: 461–469.
6. Hadley EB, Hancock RE (2010) Strategies for the Discovery and Advancement of Novel Cationic Antimicrobial Peptides. Curr Top Med Chem.
7. Pestana-Calsa MC, Ribeiro IL, Calsa T, Jr. (2010) Bioinformatics-coupled molecular approaches for unravelling potential antimicrobial peptides coding genes in Brazilian native and crop plant species. Curr Protein Pept Sci 11: 199–209.
8. Hammami R, Fliss I (2010) Current trends in antimicrobial agent research: chemo- and bioinformatics approaches. Drug Discov Today 15: 540–546.
9. Wang Z, Wang G (2004) APD: the Antimicrobial Peptide Database. Nucleic Acids Res 32: D590–592.
10. Fjell CD, Hancock RE, Cherkasov A (2007) AMPer: a database and an automated discovery tool for antimicrobial peptides. Bioinformatics 23: 1148–1155.
11. Hammami R, Zouhir A, Ben Hamida J, Fliss I (2007) BACTIBASE: a new web-accessible database for bacteriocin characterization. BMC Microbiol 7: 89.
12. Hammami R, Zouhir A, Le Lay C, Ben Hamida J, Fliss I (2010) BACTIBASE second release: a database and tool platform for bacteriocin characterization. BMC Microbiol 10: 22.
13. Hammami R, Ben Hamida J, Vergoten G, Fliss I (2009) PhytAMP: a database dedicated to antimicrobial plant peptides. Nucleic Acids Res 37: D963–968.
14. Lata S, Sharma BK, Raghava GP (2007) Analysis and prediction of antibacterial peptides. BMC Bioinformatics 8: 263.
15. Lata S, Mishra NK, Raghava GP (2010) AntiBP2: improved version of antibacterial peptide prediction. BMC Bioinformatics 11 Suppl 1: S19.
16. Thomas S, Karnik S, Barai RS, Jayaraman VK, Idicula-Thomas S (2010) CAMP: a useful resource for research on antimicrobial peptides. Nucleic Acids Res 38: D774–780.
17. Nakashima H, Nishikawa K, Ooi T (1986) The folding type of a protein is relevant to the amino acid composition. Journal of biochemistry 99: 153–162.
18. Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 21: 319–344.
19. Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43: 246–255.
20. Atchley WR, Zhao J, Fernandes AD, Druke T (2005) Solving the protein sequence metric problem. Proc Natl Acad Sci U S A 102: 6395–6400.
21. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27: 1226–1238.
22. Kohavi R (1997) Artificial Intelligence.
23. Friedman JH, Baskett F, Shustek LJ (1975) An algorithm for finding nearest neighbors. IEEE Trans Comput 24: 1000–1006.
24. Chou KC, Cai YD (2004) Prediction of protein subcellular locations by GO-FunD-PseAA predictor. Biochem Biophys Res Commun 320: 1236–1239.

25. Cai YD, Chou KC (2006) Predicting membrane protein type by functional domain composition and pseudo-amino acid composition. J Theor Biol 238: 395–400.
26. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, et al. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 10: 136.
27. Ng KL, Ciou JS, Huang CH (2010) Prediction of protein functions based on function-function correlation relations. Computers in Biology and Medicine 40: 300–305.
28. Bals R (2000) Epithelial antimicrobial peptides in host defense against infection. Respir Res 1: 141–150.
29. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659.
30. Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins-Structure Function and Genetics 43: 246–255.
31. Agrawal A, Huang X (2011) Pairwise statistical significance of local sequence alignment using sequence-specific and position-specific substitution matrices. IEEE/ACM transactions on computational biology and bioinformatics/IEEE, ACM 8: 194–205.
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of molecular biology 215: 403–410.
33. Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227: 1435–1441.
34. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755–763.
35. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. Journal of molecular biology 147: 195–197.
36. Altschul SF (1997) Evaluating the statistical significance of multiple distinct local alignments. In: Suhai S, ed. Theoretical and Computational Methods in Genome Research. New York: Plenum. pp 1–14.
37. Chou KC, Elrod DW (1999) Protein subcellular location prediction. Protein Eng 12: 107–118.
38. Chou KC, Elrod DW (1999) Prediction of membrane protein types and subcellular locations. Proteins 34: 137–153.
39. Garg A, Raghava GP (2008) A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. In silico biology 8: 129–140.
40. Tamura T, Akutsu T (2007) Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition. BMC bioinformatics 8: 466.
41. Hoglund A, Donnes P, Blum T, Adolph HW, Kohlbacher O (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics 22: 1158–1165.
42. Chou KC (1995) Does the folding type of a protein depend on its amino acid composition? FEBS Lett 363: 127–131.
43. Lee S, Lee BC, Kim D (2006) Prediction of protein secondary structure content using amino acid composition and evolutionary information. Proteins 62: 1107–1114.
44. Dumontier M, Yao R, Feldman HJ, Hogue CW (2005) Armadillo: domain boundary prediction by amino acid composition. Journal of molecular biology 350: 1061–1073.
45. Lin H, Wang H, Ding H, Chen YL, Li QZ (2009) Prediction of subcellular localization of apoptosis protein using Chou's pseudo amino acid composition. Acta Biotheor 57: 321–330.
46. Fang Y, Guo Y, Feng Y, Li M (2008) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34: 103–109.
47. Jiang X, Wei R, Zhao Y, Zhang T (2008) Using Chou's pseudo amino acid composition based on approximate entropy and an ensemble of AdaBoost classifiers to predict protein subnuclear location. Amino Acids 34: 669–675.
48. Li ZC, Zhou XB, Dai Z, Zou XY (2009) Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis. Amino Acids 37: 415–425.
49. Nanni L, Lumini A (2008) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34: 653–660.
50. Zhang SW, Chen W, Yang F, Pan Q (2008) Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach. Amino Acids 35: 591–598.
51. Zhang SW, Zhang YL, Yang HF, Zhao CH, Pan Q (2008) Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. Amino Acids 34: 565–572.
52. Qiu JD, Huang JH, Liang RP, Lu XQ (2009) Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. Anal Biochem 390: 68–73.
53. Zou D, He Z, He J, Xia Y (2010) Supersecondary structure prediction using Chou's pseudo amino acid composition. J Comput Chem.
54. Esmaeili M, Mohabatkar H, Mohsenzadeh S (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J Theor Biol 263: 203–209.
55. Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2009) Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. J Theor Biol 257: 17–26.
56. Lin H, Ding H, Guo FB, Zhang AY, Huang J (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein and Peptide Letters 15: 739–744.
57. Zeng YH, Guo YZ, Xiao RQ, Yang L, Yu LZ, et al. (2009) Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J Theor Biol 259: 366–372.
58. Ding Y-S, Zhang T-L (2008) Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: An approach with immune genetic algorithm-based ensemble classifier. Pattern Recogn Lett 29: 1887–1892.
59. Qiu JD, Huang JH, Shi SP, Liang RP (2010) Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform. Protein Pept Lett 17: 715–722.
60. Chen C, Chen L, Zou X, Cai P (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Pept Lett 16: 27–31.
61. Jiang X, Wei R, Zhang T, Gu Q (2008) Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. Protein Pept Lett 15: 392–396.
62. Li FM, Li QZ (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein Pept Lett 15: 612–616.
63. Lin H, Ding H, Guo FB, Zhang AY, Huang J (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein Pept Lett 15: 739–744.
64. Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept Lett 17: 1207–1214.
65. Gu Q, Ding YS, Zhang TL (2010) Prediction of G-protein-coupled receptor classes in low homology using Chou's pseudo amino acid composition with approximate entropy and hydrophobicity patterns. Protein Pept Lett 17: 559–567.
66. Chou KC, Cai YD (2004) Predicting enzyme family class in a hybridization space. Protein science : a publication of the Protein Society 13: 2857–2863.
67. Chou KC, Shen HB (2008) Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nature protocols 3: 153–162.
68. Chou KC, Shen HB (2007) Review: Recent progresses in protein subcellular location prediction. Analytical Biochemistry 370: 1–16.
69. Zhang TL, Ding YS, Chou KC (2008) Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. Journal of theoretical biology 250: 186–193.
70. Liu T, Zheng X, Wang C, Wang J (2010) Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: an approach from auto covariance transformation. Protein Pept Lett 17: 1263–1269.
71. Wang YC, Wang XB, Yang ZX, Deng NY (2010) Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature. Protein Pept Lett 17: 1441–1449.
72. Kandaswamy KK, Pugalenthi G, Moller S, Hartmann E, Kalies KU, et al. (2010) Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino Acid composition. Protein Pept Lett 17: 1473–1479.
73. Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19.
74. Chou KC (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Current Proteomics 6: 262–274.
75. Rubinstein ND, Mayrose I, Pupko T (2009) A machine-learning approach for predicting B-cell epitopes. Molecular immunology 46: 840–847.
76. Huang T, Shi XH, Wang P, He Z, Feng KY, et al. (2010) Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS One 5: e10972.
77. Huang T, Cui W, Hu L, Feng K, Li YX, et al. (2009) Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. PLoS One 4: e8126.
78. Chou KC, Zhang CT (1995) Prediction of protein structural classes. Critical reviews in biochemistry and molecular biology 30: 275–349.
79. Boman HG (2003) Antibacterial peptides: basic facts and emerging concepts. J Intern Med 254: 197–215.
80. Silverstein KA, Moskal WA, Jr., Wu HC, Underwood BA, Graham MA, et al. (2007) Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants. Plant J 51: 262–280.
81. Manners JM (2007) Hidden weapons of microbial destruction in plant genomes. Genome Biol 8: 225.
82. Lehrer RI (2004) Primate defensins. Nat Rev Microbiol 2: 727–738.
83. Chou KC, Scheraga HA (1982) Origin of the right-handed twist of beta-sheets of poly(LVal) chains. Proc Natl Acad Sci U S A 79: 7047–7051.
84. Chou KC (2000) Prediction of tight turns and their types in proteins. Anal Biochem 286: 1–16.
85. Kacprzyk L, Rydengard V, Morgelin M, Davoudi M, Pasupuleti M, et al. (2007) Antimicrobial activity of histidine-rich peptides is dependent on acidic conditions. Biochim Biophys Acta 1768: 2667–2680.
86. Mozsolits H, Wirth HJ, Werkmeister J, Aguilar MI (2001) Analysis of antimicrobial peptide interactions with hybrid bilayer membrane systems using surface plasmon resonance. Biochim Biophys Acta 1512: 64–76.

87. Wang G (2002) How the lipid-free structure of the N-terminal truncated human apoA-I converts to the lipid-bound form: new insights from NMR and X-ray structural comparison. FEBS Lett 529: 157–161.
88. Wang G, Pierens GK, Treleaven WD, Sparrow JT, Cushley RJ (1996) Conformations of human apolipoprotein E(263–286) and E(267–289) in aqueous solutions of sodium dodecyl sulfate by CD and 1H NMR. Biochemistry 35: 10358–10366.
89. Wimley WC, White SH (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol 3: 842–848.
90. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of theoretical biology.