Biochemical and Biophysical Research Communications 477 (2016) 150–154

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Prediction of cell-penetrating peptides with feature selection techniques

Hua Tang [a, *], Zhen-Dong Su [c], Huan-Huan Wei [c], Wei Chen [b, c], Hao Lin [c, **]

[a] Department of Pathophysiology, Southwest Medical University, Luzhou 646000, China
[b] Department of Physics, School of Sciences, Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063009, China
[c] Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics and Center for Information in Biomedicine, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China

ARTICLE INFO

Article history:
Received 30 May 2016
Accepted 8 June 2016
Available online 10 June 2016

Keywords:
Cell-penetrating peptides
Support vector machine
g-gap dipeptide composition
Analysis of variance

ABSTRACT

Cell-penetrating peptides are a group of peptides which can transport different types of cargo molecules such as drugs across plasma membrane and have been applied in the treatment of various diseases. Thus, the accurate prediction of cell-penetrating peptides with bioinformatics methods will accelerate the development of drug delivery systems. The study aims to develop a powerful model to accurately identify cell-penetrating peptides. At first, the peptides were translated into a set of vectors with the same dimension by using dipeptide compositions. Secondly, the Analysis of Variance-based technique was used to reduce the dimension of the vector and explore the optimized features. Finally, the support vector machine was utilized to discriminate cell-penetrating peptides from non-cell-penetrating peptides. The five-fold cross-validated results showed that our proposed method could achieve an overall prediction accuracy of 83.6%. Based on the proposed model, we constructed a free webserver called C2Pred (http://lin.uestc.edu.cn/server/C2Pred).

1. Introduction

Cell-penetrating peptides (CPPs), a group of short peptides, are able to mediate the intracellular delivery of a series of molecules. They were firstly found in the measurements of the activity of the tat protein from human immunodeficiency virus 1 (HIV-1) in 1988 [1]. Subsequent studies revealed that a positively charged peptide between the 47th and 57th amino acids was in charge of the translocation [2]. Since then, more and more CPPs had been experimentally identified [3,4]. Generally, the CPPs are typically hydrophobic linear arrangements of less than 30 residues and can transport both small and large molecules in vitro and in vivo. Thus, they have great potential in the biological research and medicine development. Recently, active CPPs (ACPPs) have been employed to target cancer cells over-expressing metalloproteinase-2 [5] and treat various inflammatory diseases [6]. Therefore, it is crucial to deeply understand the CPPs function so as to reveal the mechanism involved in membrane translocation. The correct identification of CPPs is the first step for understanding their translocation mechanisms and important for discovering more CPPs, which can be used as drug delivery agent.

Although the biochemical experimental approaches can provide the exact details for CPPs, the wet experimental technique is time-consuming and expensive. With a lot of biological data, it is highly desirable to develop computational methods to identify CPPs. The machine learning-based algorithms allow us to find out CPPs from huge peptide data. In fact, several methods had been proposed to computationally identify CPPs. Dobchev et al. [7] investigated 101 peptides with artificial neural networks (ANN) and principal component analysis (PCA) and obtained the prediction accuracy of 80%–100%. Sanders et al. [8] firstly established a more objective benchmark dataset including 111 experimentally confirmed CPPs and 34 known non-CPPs and then distinguished CPPs from non-CPPs with support vector machine (SVM) and 61 features. The sensitivity (*Sn*), specificity (*Sp*) and overall accuracy (*Acc*) were 75.9%, 23.2% and 75.9%, respectively. Subsequently, Gautam et al. [9] improved the prediction accuracy by considering different features including amino acid compositions and dipeptide

* Corresponding author.
** Corresponding author.
E-mail addresses: tanghua771211@aliyun.com (H. Tang), hlin@uestc.edu.cn (H. Lin).

http://dx.doi.org/10.1016/j.bbrc.2016.06.035
0006-291X/

compositions as well as physicochemical properties. However, the prediction accuracy for an independent dataset was only 81.31%. Holton et al. [10] developed a web server called CPPpred based on N-to-1 neural network to predict CPPs and obtained the *Acc* of only 75.86% in 5-fold cross-validation for a non-redundant dataset. Recently, Chen et al. [11] used pseudo amino acid composition (PseAAC) to encode peptides and achieved the *ACC* of 83.5% in 10-fold cross-validation by using random forest (RF).

The aforementioned methods did yield quite encouraging results. However, the prediction accuracies are still far from satisfactory. Moreover, few online resources of CPPs prediction are available. Therefore, we developed a powerful tool to correctly identify CPPs. In the study, we enhanced the prediction power and quality in identifying CPPs.

According to previous studies [12–14], the machine learning method for predicting biological molecules included four steps: benchmark dataset construction, representation of peptide samples, machine learning method selection and webserver construction. In the following sections, we will describe the four steps in detail.

2. Material and methods

2.1. Benchmark datasets

A reliable benchmark dataset is the foundation of an accurate model. Thus, in this study, all the CPPs and non-CPPs were derived from CPPsite2.0 [15]. To avoid any similarity bias which would result in an overestimate of predicted results, we used the CD-HIT program to remove highly similar sequences. Finally, a total of 411 experimentally confirmed CPPs and 411 know non-CPPs were gained. All the data can be freely downloaded from the website (http://lin.uestc.edu.cn/server/CPPIden/data).

2.2. Representation of peptide samples

Formulating a given peptide sample **P** with a mathematics descriptor is the second step to develop a sequence-based predictor for CPPs prediction. The most straightforward method in the current benchmark dataset is to use its entire amino acid sequence to formulate a peptide **P** with L residues as follows

$$\mathbf{P} = \mathrm{R_1R_2R_3R_4}\ldots\mathrm{R_L} \tag{1}$$

where $\mathrm{R_1}$, $\mathrm{R_2}$ and $\mathrm{R_L}$ respectively denote the 1st, 2nd, and L-th residues of the peptide **P**. This sequential model can be statistically analyzed by using BLAST program. Unfortunately, the straightforward and intuitive approach fails when a query peptide sequence has no significant similarity to any known CPP sequence.

To solve this problem, a peptide sequence may be firstly translated into a vector with the same dimension and then machine learning method is used to perform prediction. The peptide samples in the vector format can be more easily handled than those in the sequence format with many existing operation engines. The most simple vector model for a peptide sequence is its amino acid composition (AAC), which has been widely applied in protein classification [16–19]. However, if AAC was used to represent a peptide sample, all of its sequence order information would be lost. Many studies demonstrated that the residue-order information was very important for peptide structure and function annotation [12–14,20]. Thus, in this work, CPPs and non-CPPs were described with dipeptide composition as follows:

$$\mathbf{p} = [f_1, f_2, \cdots, f_u, \cdots, f_{400}]^{\mathrm{T}} \tag{2}$$

where the f_u is the frequency of the u-th ($u = 1, 2, \ldots, 400$) dipeptide defined as

$$f_u = \frac{x_u}{\sum_u x_u} \tag{3}$$

where x_u denotes the number of the u-th dipeptide in a peptide.

According to Eqs. (2) and (3), each sample can be transformed into a 400-dimension vector. However, it is well known that the length of CPPs usually ranges from 12 to 26 residues, suggesting that noise or redundant information is included in this vector model. Generally, garbage information will prevent the proposed model from correctly identifying CPPs. Moreover, the computational time will increase. Thus, it is necessary to pick out the useful features via a feature selection technique. Currently, the technique based on analysis of variance (ANOVA) has been proposed to rank the features and improve the predictive accuracies in protein classification field [13,19,21]. Thus, we also used the feature selection technique to optimize the feature set for improving the predictive performance.

On the basis of the ANOVA theory, the importance of each dipeptide for CPP prediction can be defined as:

$$F(u) = \frac{\sum_{i=1}^{2} m_i \left(\frac{\sum_{j=1}^{m_i} f_u(i,j)}{m_i} - \frac{\sum_{i=1}^{2}\sum_{j=1}^{m_i} f_u(i,j)}{\sum_{i=1}^{2} m_i} \right)^2}{\sum_{i=1}^{2}\sum_{j=1}^{m_i} \left(f_u(i,j) - \frac{\sum_{j=1}^{m_i} f_u(i,j)}{m_i} \right)^2 \Big/ (m_1 + m_2 - 2)} \tag{4}$$

where $f_u(i,j)$ denotes the frequency of the u-th dipeptide of the j-th sample in the i-th group; m_i denotes the number of samples in the i-th group (here $m_1 = 411$, $m_2 = 411$). It is obviously that the larger $F(u)$ value means the better discriminative capability of the u-th feature. Thus, we may rank all features according to their F values. Then we investigated the predictive performance of the first feature subset including the feature with the largest F value on CPP prediction. Subsequently, we measured the predictive accuracy of a new feature subset produced by adding a new feature with the second highest F value into the first feature set. This process was repeated from the higher F value to the lower F value until all candidate features were added. The optimal feature subset including u_0 ranked dipeptides which could achieve the highest predictive accuracy was expressed as:

$$\mathbf{p}_{u_0} = [f_1, f_2, \cdots, f_{u_0}]^T \tag{5}$$

Based on the feature selection, the high-dimensional data were projected into a low-dimensional space. The final model was built based on the optimal feature subset.

2.3. Machine learning method

After the representation of peptide samples, the third step in CPPs prediction is to perform the classification with a machine learning method. With the progress of in mathematical theory, several machine learning methods, such as, fisher discrimination (FD) [22], RF [23], neural network (NN) [7], and k-nearest neighbors (KNN) [24] have been developed and widely applied in

bioinformatics. SVM is one of the most powerful and popular methods in protein and DNA motif classification [25–27]. Its basic idea is to transform the input vector into a high-dimension Hilbert space and seek a separating hyperplane in this space. Due to its excellent learning ability, especially for small sample size, we also used SVM to perform classification.

In this work, each sample in the benchmark dataset expressed as a vector has a corresponding label $y \in \{+1, -1\}$, where $+1$ and -1 indicate CPPs and non-CPPS, respectively. The SVM projects the input vectors into a high-dimensional feature space for constructing an optimal separating hyperplane with the largest distance between two classes, measured along a line perpendicular to this hyperplane. The decision function of SVM is expressed as:

$$f\left(\vec{p}\right) = \mathrm{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \cdot K\left(\vec{p_i}, \vec{p_j}\right)\right) + b \tag{6}$$

where N is the number of samples (here $N = 145$); $K(\vec{p_i}, \vec{p_j})$ is called kernel function, which is an inner product in a high-dimensional feature space. In this work, we used radial basis function (RBF) defined as $K(\vec{p_i}, \vec{p_j}) = \exp(-\left\|\gamma\vec{p_i} - \vec{p_j}\right\|^2)$. The coefficients α_i can be solved by the convex Quadratic Programming (QP) problem,

$$\text{Maximize} \quad \left\{\sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j \cdot y_i y_j \cdot K\left(\vec{p_i}, \vec{p_j}\right)\right\} \tag{7}$$
$$\text{Subjected to} \quad 0 < \alpha_i < C$$

where the regularization parameter C can control the tradeoff between margin and misclassification error. These $\vec{p_i}$ are called support vectors only if corresponding $\alpha_i > 0$.

For the convenience of study, a software package called LibSVM (https://www.csie.ntu.edu.tw/~cjlin/libsvm/) was designed to implement SVM. The regularization parameter C and kernel parameter γ were optimized by using a grid search method with cross-validation test. The search spaces for C and γ are $[2^{15}, 2^{-5}]$ and $[2^{-5}, 2^{-15}]$ with the steps of 2^{-1} and 2, respectively.

2.4. Performance evaluation

Three cross-validation methods are widely used in statistical prediction: independent dataset test, sub-sampling (2-, 5- or 10-fold cross-validation) test, and jackknife test [7,8–14,18–27]. The jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors because it can always yield a unique result for a given benchmark data set. [12–14,21,22]. However, it is time-consuming to run the test. Thus, in previous studies on CPPs prediction, the n-fold cross-validation was adopted [8,11]. To provide an objective comparison, we also used n-fold cross-validation in this study.

A set of simple methods to measure the prediction quality are introduced as the follows:

$$Sn = \frac{TP}{TP + FN} 0 \quad \leq Sn \leq 1 \tag{8}$$

$$Sp = \frac{TN}{TN + FP} \quad 0 \leq Sp \leq 1 \tag{9}$$

$$Acc = \frac{TP + TN}{TP + TN + PF + FN} \quad 0 \leq Acc \leq 1 \tag{10}$$

where Sn, Sp and Acc are called sensitivity, specificity and overall accuracy, respectively. TP and TN denote the numbers of correctly recognized CPPs and non-CPPs, respectively. FP and FN are the number of the non-CPPs incorrectly predicted as the CPPs and the number of CPPs incorrectly predicted as the non-CPPs, respectively.

3. Results and discussion

3.1. Feature selection for improving accuracy

Based on the dipeptide composition definition in Eqs. (2) and (3), each peptide in benchmark dataset may be represented by a 400-dimension vector. To improve the predictive performance, it is necessary to find out the best feature subset which can produce the highest *Acc*. It is obvious that the optimal feature set can be obtained by investigating the *Acc* of all combinations of features. However, it is impossible to examine the performance of all feature subsets due to the long computation time. For the amino acid composition including 20 features, there are over 10^6 possible combinations. If the dimension increases to 400, the number of all possible combinations will be greater than 10^{120}, which is beyond the computational capability for most computers.

Thus, in order to reduce computation time, we used the feature selection technique defined in Eq. (4) to optimize features. At first, we used $F(u)$ value to evaluate the importance of each dipeptide for CPP prediction. Secondly, 400 dipeptides were ranked according to their $F(u)$ values. Thirdly, we estimated the *Acc* of the first feature with the largest $F(u)$ by using SVM. Furthermore, a new feature subset was achieved when the feature with the second highest F value was added. Then the *Acc* of this feature subset was investigated. We repeated the process from the large $F(u)$ value to the small $F(u)$ value until the *Acc*s of all candidate features were examined. All examinations were performed by using 5-fold cross-validation to avoid over-fitting.

A large feature set bears more information. However, it will also bring about information redundancy or noise, which will result in the low capability in the generalization of a predictor or reduce the cluster-tolerant capacity so as to lower the cross-validation accuracy. For example, by investigating the accuracy of 400 features for CPPs prediction, we found that 80.7% samples could be correctly predicted in 5-fold cross-validation. However, the low-dimension feature could improve the robustness of a predictor. However, if few features are available, the obtained features are still not the optimal features for prediction because they cannot afford enough information or reflect real characteristics of the CPPs, thus leading to the low predictive accuracy. For instance, 10 dipeptides can only produce the *Acc* of 75.3% in 5-fold cross-validation.

To achieve the optimal feature subset and produce the highest *Acc*, a curve (Fig. 1) was plotted in a 2D Cartesian coordinate system with the number of features as its abscissa and the *Acc* as its ordinate. The peak of the curve corresponds to the maximum *Acc*. As shown in Fig. 1, when 164 best dipeptides are used, the maximum *Acc* of 83.6% is obtained, and the *Sn* and *Sp* are 81.5% and 85.6%, respectively.

3.2. Comparison with other methods

To demonstrate the advantages of the proposed model, we made a comparison between the proposed method and other published methods. Sanders et al. constructed a benchmark dataset including 111 experimentally confirmed CPPs and 34 know non-CPPs [8]. We examined our method on this dataset using 10-fold cross-validation and recorded results in Table 1. It shows that all index values of published methods are much lower than the corresponding ones achieved by our method. It should be noted that the features used in our model are much less than that in other published methods, suggesting that our model is more robust and

ingenious.

Sander et al. [8] designed a series of experiments to investigate the effect of unbalanced datasets on CPP prediction. Here, we compared the accuracies of our method with Sander's method through two strategies [8]. The first was to yield a set of 111 negative examples randomly repeated produced from the 34 known negatives. Combining with the 111 positive samples, we found that the overall accuracy is 89.92% in 10-fold cross-validation, which is higher than that of Sander's model (88.74%). The second strategy was to yield a set of 34 positive examples randomly selected from the 111 know positive examples. Combining with the 34 known negative examples, our model obtained the overall accuracy of 83.35% in 10-fold cross-validation, which is still higher than that of Sander's model (78.82%).

Gautam et al. [28] constructed a curated database called CPPsite including 843 cell penetrating peptides. To investigate the performance of the proposed method, we randomly selected 5 sets of 34 positive examples from CPPsite. Combining with the 34 known negative examples, we noticed that the averaged overall accuracy is 93.53% in 10-fold cross-validation, suggesting that our method could identify cell-penetrating peptides. Gautam et al. [28] also provided several search pages for users to run a search of a peptide against the CPPsite. However, these servers were based on homologous sequence in the searching dataset. Our model is more flexible because it is just dependent on the residue sequence.

Feature selection by using ANOVA can not only provide deeper insights into the intrinsic properties of peptide sequences, but also economize runtime and computational resource. Furthermore, it is robust to most violations of its assumptions and more intuitive for users to analyze the interaction of the two features. The ANOVA-based method did improve the cross-validated accuracies and robustness of the model. Therefore, our model has the better performance for CPP prediction.

Protein structure and function are correlated with the physiochemical properties of amino acids. Consequently, in the future, we will make our efforts to improve the accuracy by combining the optimal dipeptide composition with physiochemical properties of amino acids.

The positive and negative samples can be found from a number of different experimental techniques (different detection methodologies, and different cell types) [8]. Thus, in the future, we will collect more data from different cell types for the creation of CPPs dataset and the prediction of CPPs.

Table 1
Comparison with published methods.

Methods	*Sn*	*Sp*	*Acc*
Sander et al.'s method [8]	0.759	0.232	0.759
Chen et al.'s method [11]	0.955	0.441	0.835
Our method	0.973	0.765	0.924

4. Web-server guide

We established a user-friendly web-server called **C2Pred** based on the proposed model to improve the efficiency and avoid a repeated a complicated mathematic process for identifying CPPs.

To facilitate the studies of other scholars, we provided the guideline as follows. One may browse the web server at http://lin.uestc.edu.cn/server/C2Pred (top page shown in Fig. 2**)**. The Read me button provides a brief introduction about the predictor and the caveat for use. The Data button lists a link for downloading the benchmark datasets. The Citation button gives the relevant paper of **C2Pred**. The Example button provides example sequences in FASTA format. Users may type or copy/paste the query peptide sequences with FASTA format into the input box at the center of Fig. 2. After submitting their peptide sequences, results will be shown in a new interface.

We will provide the maintenance of **C2Pred**. The server will be available to the research community for at least five years. With the collection of new CPPs and the development of new algorithms, we will upgrade the server with the higher accuracy and specificity. At this stage, **C2Pred** provides users with a highly practical tool and straightforward web interface for identifying CPPs. We also plan to develop more useful applications on CPP analysis.

5. Conclusions

In this paper, we developed a novel approach to discriminate CPPs from non-CPPs. In order to improve the prediction capability of the model, we designed a feature selection technique based on the ANOVA. An overall accuracy of 83.6% was achieved in 5-fold cross-validation underlining the predictive power of **C2Pred**.

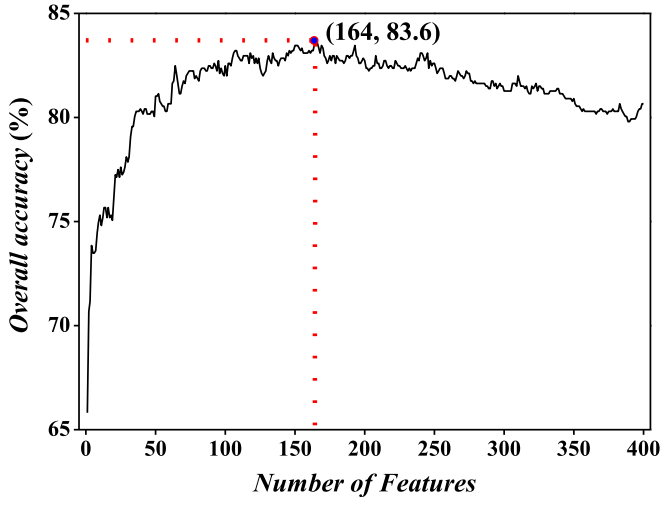

Fig. 1. The feature selection results. When the top 164 dipeptides were used to perform prediction, the overall accuracy reached its peak of 83.6%.

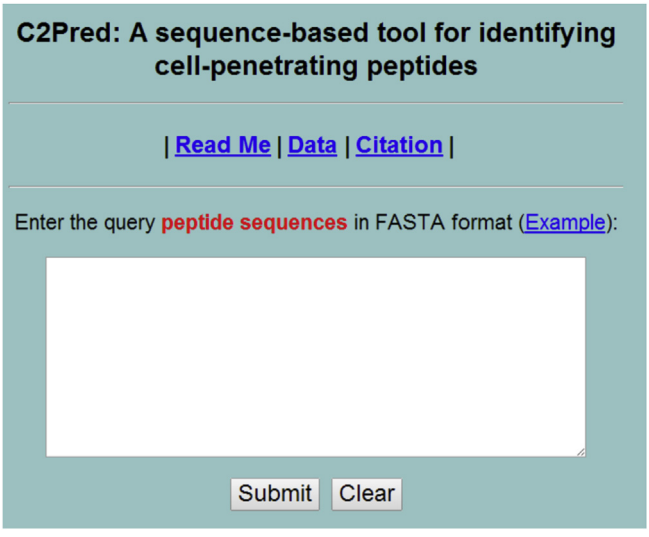

Fig. 2. A semi-screenshot to show the top page of the **C2Pred** webserver. Its website address is http://lin.uestc.edu.cn/server/C2Pred.

Conflict of interest

The authors declare that there is no conflict of interests.

Acknowledgements

This work was supported by the Applied Basic Research Program of Sichuan Province (No. 2015JY0100 and LZ-LY-45), the Scientific Research Foundation of the Education Department of Sichuan Province (11ZB122), the Nature Scientific Foundation of Hebei Province (No. C2013209105), the Fundamental Research Funds for the Central Universities of China (No. ZYGX2015J144) and Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (No. BJ2014028).

References

[1] A.D. Frankel, C.O. Pabo, Cellular uptake of the tat protein from human immunodeficiency virus, Cell 55 (1988) 1189–1193.
[2] E. Vives, P. Brodin, B. Lebleu, A truncated HIV-1 Tat protein basic domain rapidly translocates through the plasma membrane and accumulates in the cell nucleus, J. Biol. Chem. 272 (1997) 16010–16017.
[3] M. Lindgren, M. Hallbrink, A. Prochiantz, U. Langel, Cell-penetrating peptides, Trends Pharmacol. Sci. 21 (2000) 99–103.
[4] H. Matsui, K. Tomizawa, M. Matsushita, Protein transduction by poly-arginine, Nihon Yakurigaku Zasshi 121 (2003) 435–439.
[5] E.S. Olson, T.A. Aguilera, T. Jiang, L.G. Ellies, Q.T. Nguyen, E.H. Wong, L.A. Gross, R.Y. Tsien, In vivo characterization of activatable cell penetrating peptides for targeting protease activity in cancer, Integr. Biol. (Camb) 1 (2009) 382–393.
[6] J.S. Orange, M.J. May, Cell penetrating peptide inhibitors of nuclear factor-kappa B, Cell Mol. Life Sci. 65 (2008) 3564–3591.
[7] D.A. Dobchev, I. Mager, I. Tulp, G. Karelson, T. Tamm, K. Tamm, J. Janes, U. Langel, M. Karelson, Prediction of Cell-Penetrating Peptides Using Artificial Neural Networks, Curr. Comput. Aided Drug Des. 6 (2010) 79–89.
[8] W.S. Sanders, C.I. Johnston, S.M. Bridges, S.C. Burgess, K.O. Willeford, Prediction of cell penetrating peptides by support vector machines, PLoS Comput. Biol. 7 (2011) e1002101.
[9] A. Gautam, K. Chaudhary, R. Kumar, A. Sharma, P. Kapoor, A. Tyagi, c. Open source drug discovery, G.P. Raghava, In silico approaches for designing highly effective cell penetrating peptides, J. Transl. Med. 11 (2013) 74.
[10] T.A. Holton, G. Pollastri, D.C. Shields, C. Mooney, CPPpred: prediction of cell penetrating peptides, Bioinformatics 29 (2013) 3094–3096.
[11] L. Chen, C. Chu, T. Huang, X. Kong, Y.D. Cai, Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models, Amino Acids 47 (2015) 1485–1493.
[12] C. Ding, L.F. Yuan, S.H. Guo, H. Lin, W. Chen, Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions, J. Proteom. 77 (2012) 321–328.
[13] H. Ding, P.M. Feng, W. Chen, H. Lin, Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis, Mol. Biosyst. 10 (2014) 2229–2235.
[14] P.P. Zhu, W.C. Li, Z.J. Zhong, E.Z. Deng, H. Ding, W. Chen, H. Lin, Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition, Mol. Biosyst. 11 (2015) 558–563.
[15] P. Agrawal, S. Bhalla, S.S. Usmani, S. Singh, K. Chaudhary, G.P. Raghava, A. Gautam, CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides, Nucl. Acids Res. 44 (2016) D1098–D1103.
[16] A.S. Ettayapuram Ramaprasad, S. Singh, P.S.R. Gajendra, S. Venkatesan, Anti-AngioPred: a server for prediction of anti-angiogenic peptides, PLoS One 10 (2015) e0136990.
[17] Y.E. Feng, Identify secretory protein of malaria parasite with modified quadratic discriminant algorithm and amino acid composition, Interdiscip. Sci. 8 (2015) 156–161.
[18] H. Lin, W. Chen, Prediction of thermophilic proteins using feature selection technique, J. Microbiol. Methods 84 (2011) 67–70.
[19] K.C. Chou, Prediction of protein cellular attributes using pseudo-amino acid composition, Proteins 43 (2001) 246–255.
[20] H. Tang, W. Chen, H. Lin, Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique, Mol. BioSyst. 12 (2016) 1269–1275.
[21] W. Chen, H. Ding, P.M. Feng, H. Lin, K.C. Chou, iACP: a sequence-based tool for identifying anticancer peptides, Oncotarget 7 (2016) 16895–16909.
[22] H. Ding, L. Luo, H. Lin, Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition, Protein Pept. Lett. 16 (2009) 351–355.
[23] X. Ma, J. Guo, X. Sun, Sequence-based prediction of RNA-binding proteins using random forest with minimum redundancy maximum relevance feature selection, Biomed. Res. Int. 2015 (2015) 425810.
[24] A. Suratanee, K. Plaimas, Identification of inflammatory bowel disease-related proteins using a reverse k-nearest neighbor search, J. Bioinform. Comput. Biol. 12 (2014) 1450017.
[25] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, H. Lin, W. Chen, K.C. Chou, iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 30 (2014) 1522–1529.
[26] G.L. Fan, X.Y. Zhang, Y.L. Liu, Y. Nang, H. Wang, DSPMP: Discriminating secretory proteins of malaria parasite by hybridizing different descriptors of Chou's pseudo amino acid patterns, J. Comput. Chem. 36 (2015) 2317–2327.
[27] H. Lin, E.Z. Deng, H. Ding, W. Chen, K.C. Chou, iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucl. Acids Res. 42 (2014) 12961–12972.
[28] A. Gautam, H. Singh, A. Tyagi, K. Chaudhary, R. Kumar, P. Kapoor, G.P. Raghava, CPPsite: a curated database of cell penetrating peptides, Database (Oxford) (2012), 2012:bas015.