Send Orders for Reprints to reprints@benthamscience.net

Medicinal Chemistry, **2020**, *16*, **1-6**

RESEARCH ARTICLE

iATP: A Sequence Based Method for Identifying Anti-tubercular Peptides

Wei Chen[1,2,*], Pengmian Feng[1] and Fulei Nie[2]

[1]*Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China;* [2] *Center for Genomics and Computational Biology, School of Life Sciences, North China University of Science and Technology, Tangshan 063000, China*

Abstract: ***Background*:** Tuberculosis is one of the biggest threats to human health. Recent studies have demonstrated that anti-tubercular peptides are promising candidates for the discovery of new anti-tubercular drugs. Since experimental methods are still labor intensive, it is highly desirable to develop automatic computational methods to identify anti-tubercular peptides from the huge amount of natural and synthetic peptides. Hence, accurate and fast computational methods are highly needed.

Methods and Results: In this study, a support vector machine based method was proposed to identify anti-tubercular peptides, in which the peptides were encoded by using the optimal *g*-gap dipeptide compositions. Comparative results demonstrated that our method outperforms existing methods on the same benchmark dataset. For the convenience of scientific community, a freely accessible web-server was built, which is available at http://lin-group.cn/server/iATP.

Conclusion: It is anticipated that the proposed method will become a useful tool for identifying anti-tubercular peptides.

ARTICLE HISTORY

Received: April 09, 2019
Revised: May 15, 2019
Accepted: August 23, 2019

DOI:
10.2174/1573406415666191002152441

Keywords: Tuberculosis; anti-tubercular peptides; *g*-gap dipeptide; support vector machine; feature selection; web-server.

1. INTRODUCTION

Tuberculosis (TB), one of the top 10 causes of death, is a serious infectious disease caused by *Mycobacterium tuberculosis* (*Mtb*) [1-3]. According to the global tuberculosis report 2018 from the World Health Organization (WHO), TB has infected approximately 10 million people and caused 1.3 million deaths worldwide last year. Therefore, there is an urgent need for the development of new drugs for the treatment of TB.

The isoniazid and rifampicin have been selected as the first-line medicines to treat TB for a long time [44]. However, the capability of these first-line medicines began to decrease due to the prevalence of drug-resistant strains of *Mtb* [5, 6]. For example, more than 0.5 million people are resistance to the most effective first-line medicine rifampicin in 2017. In such case, the second-line medicines, including cycloserine, terizidone, and ethionamide have been approved for the cure of TB [2, 4]. Unfortunately, the second-line drugs are more expensive and toxic than the first-line drugs. Moreover, the emergence of multi-drug resistant, extensively-drug resistant and totally drug resistant *Mtb* strains give rise to challenges for the cure of TB [7].

*Address correspondence to this author at Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu 611730, China; E-mails: chenweiimu@gmail.com (C.W.)

Owing to their merits of low immunogenicity, selective affinity to prokaryotic negatively charged cell envelopes, and diverse modes of action, several peptides known as anti-tubercular peptides, have been used as the pharmaceutical agents to treat TB [8]. Therefore, they become promising sources for the discovery of new anti-tubercular drugs. Since experimentally identifying anti-tubercular peptides from the huge amount of natural and synthetic peptides is still cost-ineffective, accurate and fast computational methods are highly needed for this aim. Recently, Usmani and his colleagues developed a computational method for discriminating anti-tubercular peptides from anti-bacterial peptides and non-antibacterial peptides [9]. However, the performance of their method is still unsatisfactory.

In order to improve the predictive performance for identifying anti-tubercular peptides, this study developed a support vector machine-based method, in which the peptides were encoded by using the optimal *g*-gap dipeptide. In the jackknife test, we obtained the overall accuracies of 80.69% and 87.80% for discriminating anti-tubercular peptides from anti-bacterial peptides and non-antibacterial peptides, which is better than Usmani *et al*.'s method. Moreover, a web-server was built based on the proposed method, which is freely available at http://lin-group.cn/server/iATP. It is anticipated that the proposed method will become a useful tool for identifying anti-tubercular peptides.

1573-4064/20 $65.00+.00

2. MATERIALS AND METHOD

2.1. Benchmark Dataset

In the current study, the benchmark datasets used to train and test the proposed model were constructed by Usmani *et al*. [9]. The positive dataset contains 246 experimentally verified anti-tubercular peptides that are effective against *Mycobacterium*. Two negative datasets were prepared by Usmani *et al*. [9]. One negative dataset contains 246 anti-bacterial peptides that were obtained from antimicrobial peptide database DBAASP [10]. As indicated by Usmani *et al*., they are active against Gram positive and Gram negative bacteria, and are independent of the positive dataset [9]. The other negative dataset contains 246 non-antibacterial peptides that were generated from Swiss-Prot database, none of which is identical to anti-tubercular and anti-bacterial peptides. Therefore, two benchmark datasets were formed and expressed as follows,

$$\begin{cases} S_1 = S^+ \cup S^-_{antibacterial} \\ S_2 = S^+ \cup S^-_{non-antibacterial} \end{cases} \quad (1)$$

where S^+ contains the 246 anti-tubercular peptides, $S^-_{antibacterial}$ contains the 246 anti-bacterial peptides, and $S^-_{nonantibacterial}$ contains the 246 non-antibacterial peptides, respectively. The length of the peptides in both positive and negative dataset ranges from 5 to 61 amino acids.

2.2. *g*-gap Dipeptide Composition

Instead of the proximate dipeptide composition, the *g*-gap dipeptide composition describing long-range correlation between two residues has demonstrated its effectiveness in the realm of proteomics [11]. Therefore, in the present work, the *g*-gap dipeptide composition was used to encode the peptides in the benchmark dataset.

Suppose a peptide **P** with *L* residues as given by

$$\mathbf{P}=R_1\ R_2\ R_3 \cdots R_{(L-2)}\ R_{(L-1)}\ R_L \quad (2)$$

where R_1 is the residue at the first position of the peptide, R_2 is the residue at the second position, and so forth.

According to the *g*-gap dipeptide composition, a peptide will be converted to a 400-dimensional feature vector expressed as

$$\mathbf{F} = [f_1^g\ f_2^g \cdots f_i^g \cdots f_{400}^g]^T \quad (3)$$

where f_g^i is the frequency of the *i*th *g*-gap dipeptide in the peptide and is defined as,

$$f_i^g = \frac{n_i^g}{\sum_{i=1}^{400} n_i^g} = \frac{n_i^g}{L-g-1} \quad (4)$$

where n_g^i is the number of the *i*th *g*-gap dipeptide and *g* is an integral number. It is obvious that the *g*-gap dipeptide composition represents the correlation between two residues with *g* residues interval. In the current study, considering the length distribution of the peptides in the benchmark dataset, *g* is in the range of [0, 4].

2.3. Support Vector Machine (SVM)

SVM is a powerful and popular method for classification and regression analysis, which has been widely used in computational genomics and proteomics [12-22]. Its basic idea is to transform the input data into a high dimensional feature space and then determine the optimal separating hyperplane. Owing to its effectiveness and speed in the training process, the radial basis kernel function (RBF) of SVM was used to obtain the classification hyperplane in this study. The regularization parameter *C* and kernel parameter γ of the SVM operation engine were optimized in the ranges $[2^{-5}, 2^{15}]$ and $[2^{-15}, 2^{-5}]$ with the steps of 2 and 2^{-1}, respectively.

2.4. Feature Selection

Inclusion of redundant features will cause poor prediction results and increase computational time. In order to alleviate irrelevant features, a series of effective feature selection techniques have been proposed [23-25]. To improve the prediction quality, in the current study, we performed feature selection using the ''fselect.py'' algorithm (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools), which ranks the features according to their scores. The ranked feature with a higher score indicates that it is a more highly relevant one for the target to be predicted. To determine the optimal number of features, the Incremental Feature Selection (IFS) was performed. By adding these features sequentially from the higher to lower ranks, new feature sets will be obtained [26-28]. For each feature set, an SVM model was built and evaluated in terms of accuracy by using the 5-fold cross validation test. By doing so, an IFS curve will be obtained, from which the optimal feature set is defined when the IFS curve reaches its peak.

2.5. Performance Evaluation

Sensitivity (Sn), specificity (Sp), accuracy (Acc) and Mathew's correlation coefficient (MCC) were used to evaluate the performance of the proposed method, which are expressed as follows [29-35].

$$\begin{cases} Sn = \dfrac{TP}{TP+FN} \times 100\% \\ Sp = \dfrac{TN}{TN+FP} \times 100\% \\ Acc = \dfrac{TP+TN}{TP+FN+TN+FP} \times 100\% \\ MCC = \dfrac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FN) \times (TN+FP)}} \end{cases} \quad (4)$$

where *TP*, *TN*, *FP* and *FN* represent true positive, true negative, false positive, and false negative, respectively.

Besides the four metrics defined above, the threshold independent measure, area under the ROC curve (auROC), has also been widely used to objectively evaluate the performance quality of a binary classifier [18, 36]. A value of 0.5 is equivalent to random prediction, while a value of 1 represents a perfect prediction. Therefore, the auROC was also used to evaluate the performance of the current method.

3. RESULT AND DISCUSSION

Table 1. Performance of different methods for identifying anti-tubercular peptides.

	Benchmark Dataset S_1				Benchmark Dataset S_2			
	Sn(%)	**Sp(%)**	**Acc(%)**	**Mcc**	**Sn(%)**	**Sp(%)**	**Acc(%)**	**Mcc**
g=0	71.54	83.74	77.64	0.56	**79.67**	**91.87**	**85.77**	**0.72**
g=1	**75.61**	**84.96**	**80.28**	**0.61**	76.83	92.28	84.55	0.70
g=2	67.48	83.33	75.41	0.51	75.2	91.06	83.13	0.67
g=3	69.11	80.08	74.59	0.49	74.39	87.80	81.10	0.63
g=4	67.48	86.18	76.83	0.55	76.42	91.46	83.94	0.69

3.1. Determining the Optimal *g*-gap Dipeptide

In Eq. (3), the *g* describes the global sequence-order effect. The greater the *g* is, the more global sequence-order information will be included. However, if *g* is too large, it will decrease the signal-to-noise ratio. Therefore, our searching for the optimal value of *g* was carried out in the range of [0, 4] with a step of 1. Accordingly, five models (*g*=0, 1, ..., 4) were built based on the benchmark dataset S_1 and S_2, respectively. Their predictive performances of the 5-fold cross validation test for identifying anti-tubercular peptides by these models were reported in Table **1**.

As shown in Table **1**, the model based on *g*=1 obtained the best performance for discriminating anti-tubercular peptides from anti-bacterial peptides, while the model based on *g*=0 obtained the best performance for discriminating anti-tubercular peptides from non-antibacterial peptides. Therefore, in the following analysis, *g* was set to 1 for discriminating anti-tubercular peptides from anti-bacterial peptides and to 0 for discriminating anti-tubercular peptides from non-antibacterial peptides.

3.2. Prediction Performance

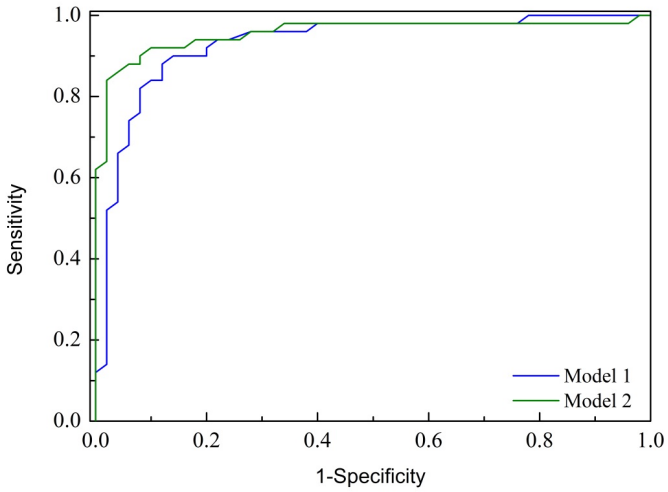

Fig. (2). A graphical illustration to show the performance of the models by means of the ROC curves obtained from the jackknife test. Model 1 is based on the benchmark dataset S_1, Model 2 is based on benchmark dataset S_2. The vertical coordinate is the true positive rate (*Sn*) while the horizontal coordinate is the false positive rate (1-*Sp*).

In order to avoid the high-dimension problem and improve the performance of the proposed model, it is necessary to choose the optimal number of features to build a robust and efficient predictive model. Here, we take the model based on benchmark dataset S_1 as an example to show the way of how to obtain the optimal feature set. At first, the 400 1-gap dipeptides were ranked by using the 'fselect.py' algorithm. Subsequently, based on the ranked 1-gap dipeptides and 147 dipeptides were used to build the final computational models for discriminating anti-tubercular peptides from anti-bacterial peptides and discriminating anti-tubercular peptides from non-antibacterial peptides, respectively. In the jackknife test, the model obtained an accuracy of 80.69% for discriminating anti-tubercular peptides from anti-bacterial peptides, and an accuracy of 87.80% discriminating anti-tubercular peptides from non-antibacterial peptides.

Furthermore, to show the performance of the current model across the entire range of SVM decision values, the ROC curves of the two models from the jackknife test were plotted in Fig. (**2**). The auROCs for the two models are 0.86 and 0.95, respectively. All results demonstrate that our proposed models are powerful for identifying anti-tubercular peptides. Accordingly, two models (Model 1 and Model 2) based on the benchmark dataset S_1 and S_2 were built for discriminating anti-tubercular peptides from anti-bacterial peptides and non-antibacterial peptides, respectively.

3.3. Comparison with Other Methods

To further demonstrate the performance of the proposed method, a comparison was made between our proposed method and Usmani *et al.*'s method [9]. Since Usmani *et al.* evaluated their method by the 5-fold cross-validation test [9], for a fair comparison, we also evaluated our method via the 5-fold cross-validation test on benchmark dataset S_1 and S_2. To perform 5-fold cross-validation test, the benchmark dataset will be randomly divided into five parts, four of them will be used as the training set and the other part as the testing set.

It was found in Table **2**, for discriminating anti-tubercular peptides from anti-bacterial peptides, our proposed method obtained an average accuracy of 79.13% with the average auROC of 0.88 on the training dataset and an average accuracy of 77.66% with the average auROC of 0.85 on the validation dataset, which are higher than the corresponding rates

Table 2. A comparison of the proposed method with the existing method.

	Benchmark Dataset S_1				Benchmark Dataset S_2			
	T^a	V^a	T^b	V^b	T^a	V^a	T^b	V^b
Sn(%)	84.22	80.43	76.76	75.02	82.01	88.51	78.68	73.33
Sp(%)	74.01	74.90	77.27	76.73	91.05	93.19	84.64	83.75
Acc(%)	79.13	77.66	77.48	75.87	86.53	90.85	81.66	78.54
MCC	0.59	0.56	0.55	0.52	0.73	0.82	0.64	0.57
auROC	0.88	0.85	0.82	0.83	0.94	0.96	0.87	0.86

[a] Results obtained from this work by using the optimal features; T indicates training, V indicates validation;
[b] The best predictive results obtained by Usmani et al[9]; T indicates training, V indicates validation.

reported by Usmani *et al*. For discriminating anti-tubercular peptides from non-antibacterial peptides, our method yielded an average accuracy of 86.53% and 90.85% on the training and validation dataset, respectively, which are 4.67% and 13.31% higher than that of Usmani *et al*.'s method [9]. The auROC obtained by our method is also higher than that of Usmani *et al*.'s method [9] (Table **2**). The above results indicate that the proposed method is indeed quite promising or at least can play a complementary role to the existing method for identifying anti-tubercular peptides.

4. WEB-SERVER

Since user-friendly and publicly accessible web-servers [26, 27] and databases [37-40] represent the future direction for developing practically more useful models, for the convenience of the scientific community, a public-accessible web-server was provided for the proposed method. The step-by-step guide on how to use the webs-server is given below.

First, open the web-server at http://lin-group.cn/server/iATP, and its top page will be shown on the computer screen, as shown in Fig. (**3**).

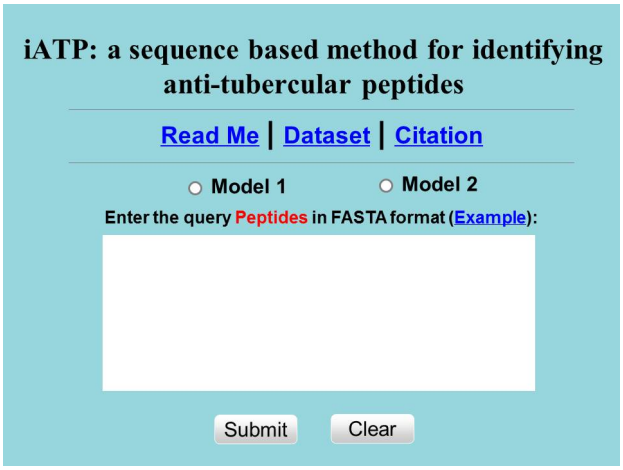

Fig. (3). The top page of the web-server. Its website address is at http://lin.uestc.edu.cn/server/iATP.

Second, either paste or type the query peptides into the input box. The input peptide should be in the FASTA format that can be seen by clicking on the Example button.

Third, click the open circle (Model 1 or Model 2) to choose the model concerned, followed by clicking the Submit button. The predictive results are shown in a new page.

CONCLUSION

In this work, we developed a support vector machine based in *silico* model to identify anti-tubercular peptides by using the *g*-gap dipeptide compositions. The feature selection technique was utilized to select the optimal *g*-gap dipeptide compositions for improving the performance of the models. In the jackknife test, the accuracy of 80.69% was obtained for discriminating anti-tubercular peptides from anti-bacterial peptides, while an accuracy of 87.80% was obtained for discriminating anti-tubercular peptides from non-antibacterial peptides. To further access the effectiveness of the proposed model, we have compared its performances with the state-of-the-art predictor for the same purpose. The cross validation results demonstrate that our method is promising and outperforms the existing predictor. For the convenience of the scientific community, a freely accessible web server for the proposed method was established at http://lin-group.cn/server/iATP. It has not escaped our notice that deep learning has become a popular method because of its wonderful performance [41-48]. Thus, in the future, we will apply it in this field.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are base of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

Not applicable.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

This work was supported by the National Nature Scientific Foundation of China (31771471, 61772119), Natural Science Foundation for Distinguished Young Scholar of Hebei Province (No. C2017209244).

REFERENCES

[1] Padhi, A.; Sengupta, M.; Sengupta, S.; Roehm, K.H.; Sonawane, A. Antimicrobial peptides and proteins in mycobacterial therapy: current status and future prospects. *Tuberculosis (Edinb.),* **2014**, *94*(4), 363-373. [http://dx.doi.org/10.1016/j.tube.2014.03.011] [PMID: 24813349]

[2] Khusro, A.; Aarti, C.; Agastian, P. Anti-tubercular peptides: A quest of future therapeutic weapon to combat tuberculosis. *Asian Pac. J. Trop. Med.,* **2016**, *9*(11), 1023-1034. [http://dx.doi.org/10.1016/j.apjtm.2016.09.005] [PMID: 27890360]

[3] Teng, T.; Liu, J.; Wei, H. Anti-mycobacterial peptides: from human to phage. *Cell. Physiol. Biochem.,* **2015**, *35*(2), 452-466. [http://dx.doi.org/10.1159/000369711] [PMID: 25613372]

[4] De Leon Rodriguez, L.M.; Kaur, H.; Brimble, M.A. Synthesis and bioactivity of antitubercular peptides and peptidomimetics: an update. *Org. Biomol. Chem.,* **2016**, *14*(4), 1177-1187. [http://dx.doi.org/10.1039/C5OB02298C] [PMID: 26645944]

[5] Silva, J.P.; Appelberg, R.; Gama, F.M. Antimicrobial peptides as novel anti-tuberculosis therapeutics. *Biotechnol. Adv.,* **2016**, *34*(5), 924-940. [http://dx.doi.org/10.1016/j.biotechadv.2016.05.007] [PMID: 27235189]

[6] Eldholm, V.; Balloux, F. Antimicrobial Resistance in Mycobacterium tuberculosis: The Odd One Out. *Trends Microbiol.,* **2016**, *24*(8), 637-648. [http://dx.doi.org/10.1016/j.tim.2016.03.007] [PMID: 27068531]

[7] Gandhi, N.R.; Nunn, P.; Dheda, K.; Schaaf, H.S.; Zignol, M.; van Soolingen, D.; Jensen, P.; Bayona, J. Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. *Lancet,* **2010**, *375*(9728), 1830-1843. [http://dx.doi.org/10.1016/S0140-6736(10)60410-2] [PMID: 20488523]

[8] Abedinzadeh, M.; Gaeini, M.; Sardari, S. Natural antimicrobial peptides against Mycobacterium tuberculosis. *J. Antimicrob. Chemother.,* **2015**, *70*(5), 1285-1289. [http://dx.doi.org/10.1093/jac/dku570] [PMID: 25681127]

[9] Usmani, S.S.; Bhalla, S.; Raghava, G.P.S. Prediction of Antitubercular Peptides From Sequence Information Using Ensemble Classifier and Hybrid Features. *Front. Pharmacol.,* **2018**, *9*, 954. [http://dx.doi.org/10.3389/fphar.2018.00954] [PMID: 30210341]

[10] Gogoladze, G.; Grigolava, M.; Vishnepolsky, B.; Chubinidze, M.; Duroux, P.; Lefranc, M.P.; Pirtskhalava, M. DBAASP: database of antimicrobial activity and structure of peptides. *FEMS Microbiol. Lett.,* **2014**, *357*(1), 63-68. [http://dx.doi.org/10.1111/1574-6968.12489] [PMID: 24888447]

[11] Pan, Y.; Gao, H.; Lin, H.; Liu, Z.; Tang, L.; Li, S. Identification of Bacteriophage Virion Proteins Using Multinomial Naïve Bayes with g-Gap Feature Tree. *Int. J. Mol. Sci.,* **2018**, *19*(6)E1779 [http://dx.doi.org/10.3390/ijms19061779] [PMID: 29914091]

[12] Chen, W.; Feng, P.M.; Deng, E.Z.; Lin, H.; Chou, K.C. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem.,* **2014**, *462*, 76-83. [http://dx.doi.org/10.1016/j.ab.2014.06.022] [PMID: 25016190]

[13] Feng, P.M.; Chen, W.; Lin, H.; Chou, K.C. iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.,* **2013**, *442*(1), 118-125. [http://dx.doi.org/10.1016/j.ab.2013.05.024] [PMID: 23756733]

[14] Su, Z.D.; Huang, Y.; Zhang, Z.Y.; Zhao, Y.W.; Wang, D.; Chen, W.; Chou, K.C.; Lin, H. iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics,* **2018**, *34*(24), 4196-4204. [http://dx.doi.org/10.1093/bioinformatics/bty508] [PMID: 29931187]

[15] Zhu, X.J.; Feng, C.Q.; Lai, H.Y.; Chen, W.; Lin, H. Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Base. Syst.,* **2019**, *163*, 787-793. [http://dx.doi.org/10.1016/j.knosys.2018.10.007]

[16] Manavalan, B.; Shin, T.H.; Lee, G. PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. *Front. Microbiol.,* **2018**, *9*, 476. [http://dx.doi.org/10.3389/fmicb.2018.00476] [PMID: 29616000]

[17] Chen, W.; Feng, P.M.; Lin, H.; Chou, K.C. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. *BioMed Res. Int.,* **2014**, *2014*623149 [http://dx.doi.org/10.1155/2014/623149] [PMID: 24967386]

[18] Chen, W.; Yang, H.; Feng, P.; Ding, H.; Lin, H. iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics,* **2017**, *33*(22), 3518-3523. [http://dx.doi.org/10.1093/bioinformatics/btx479] [PMID: 28961687]

[19] Li, D.; Ju, Y.; Zou, Q. Protein Folds Prediction with Hierarchical Structured SVM. *Curr. Proteomics,* **2016**, *13*(2), 79-85. [http://dx.doi.org/10.2174/1570164613021605140009401]

[20] Wang, S.P.; Zhang, Q.; Lu, J.; Cai, Y.D. Analysis and Prediction of Nitrated Tyrosine Sites with the mRMR Method and Support Vector Machine Algorithm. *Curr. Bioinform.,* **2018**, *13*(1), 3-13. [http://dx.doi.org/10.2174/1574893611666160608075753]

[21] Zhang, N.; Sa, Y.; Guo, Y.; Lin, W.; Wang, P.; Feng, Y.M. Discriminating Ramos and Jurkat Cells with Image Textures from Diffraction Imaging Flow Cytometry Based on a Support Vector Machine. *Curr. Bioinform.,* **2018**, *13*(1), 50-56. [http://dx.doi.org/10.2174/1574893611666160608102537]

[22] Yang, S.; Gu, J. Feature selection based on mutual information and redundancy-synergy coefficient. *J. Zhejiang Univ. Sci.,* **2004**, *5*(11), 1382-1391. [http://dx.doi.org/10.1631/jzus.2004.1382] [PMID: 15495331]

[23] Jiao, Y.S.; Du, P.F. Prediction of Golgi-resident protein types using general form of Chou's pseudo-amino acid compositions: Approaches with minimal redundancy maximal relevance feature selection. *J. Theor. Biol.,* **2016**, *402*, 38-44. [http://dx.doi.org/10.1016/j.jtbi.2016.04.032] [PMID: 27155042]

[24] Zou, Q.; Zeng, J.C.; Cao, L.J.; Zeng, X.X. A Novel Features Ranking Metric with Application to Scalable Visual and Bioinformatics Data Classification. *Neurocomputing,* **2016**, *173*, 346-354. [http://dx.doi.org/10.1016/j.neucom.2014.12.123]

[25] Zou, Q.; Wan, S.; Ju, Y.; Tang, J.; Zeng, X. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.,* **2016**, *10*(4)(Suppl. 4), 114. [http://dx.doi.org/10.1186/s12918-016-0353-5] [PMID: 28155714]

[26] Yang, H.; Lv, H.; Ding, H.; Chen, W.; Lin, H. iRNA-2OM: A Sequence-Based Predictor for Identifying 2′-O-Methylation Sites in Homo sapiens. *J. Comput. Biol.,* **2018**, *25*(11), 1266-1277. [http://dx.doi.org/10.1089/cmb.2018.0004] [PMID: 30113871]

[27] Tang, H.; Zhao, Y.W.; Zou, P.; Zhang, C.M.; Chen, R.; Huang, P.; Lin, H. HBPred: a tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.,* **2018**, *14*(8), 957-964. [http://dx.doi.org/10.7150/ijbs.24174] [PMID: 29989085]

[28] Dao, F.Y.; Lv, H.; Wang, F.; Feng, C.Q.; Ding, H.; Chen, W.; Lin, H. Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. *Bioinformatics,* **2018**. [http://dx.doi.org/10.1093/bioinformatics/bty943] [PMID: 30428009]

[29] Yu, C.Y.; Li, X.X.; Yang, H.; Li, Y.H.; Xue, W.W.; Chen, Y.Z.; Tao, L.; Zhu, F. Assessing the Performances of Protein Function Prediction Algorithms from the Perspectives of Identification Accuracy and False Discovery Rate. *Int. J. Mol. Sci.,* **2018**, *19*(1), 183. [http://dx.doi.org/10.3390/ijms19010183] [PMID: 29316706]

[30] Chen, W.; Feng, P.; Liu, T.; Jin, D. Recent advances in machine learning methods for predicting heat shock proteins. *Curr. Drug Metab.,* **2018**. [http://dx.doi.org/10.2174/1389200219666181031105916] [PMID: 30378494]

[31] Manavalan, B.; Shin, T.H.; Lee, G. DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction us-

ing the optimal features selected by random forest. *Oncotarget,* **2017**, *9*(2), 1944-1956.
[PMID: 29416743]

[32] Feng, P.M.; Ding, H.; Chen, W.; Lin, H. Naïve Bayes classifier with feature selection to identify phage virion proteins. *Comput. Math. Methods Med.,* **2013**, *2013*530696
[http://dx.doi.org/10.1155/2013/530696] [PMID: 23762187]

[33] Feng, P.M.; Lin, H.; Chen, W. Identification of antioxidants from sequence information using naïve Bayes. *Comput. Math. Methods Med.,* **2013**, *2013*567529
[http://dx.doi.org/10.1155/2013/567529] [PMID: 24062796]

[34] Feng, C.Q.; Zhang, Z.Y.; Zhu, X.J.; Lin, Y.; Chen, W.; Tang, H.; Lin, H. iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics,* **2018**.
[http://dx.doi.org/10.1093/bioinformatics/bty827] [PMID: 30247625]

[35] Du, P.F.; Li, T.T.; Wang, X.; Xu, C. SubChlo-GO: Predicting Protein Subchloroplast Locations with Weighted Gene Ontology Scores. *Curr. Bioinform.,* **2013**, *8*(2), 193-199.
[http://dx.doi.org/10.2174/1574893611308020007]

[36] Jiao, Y.; Du, P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant. Biol.,* **2016**, *4*(4), 320-330.
[http://dx.doi.org/10.1007/s40484-016-0081-2]

[37] Zhang, T.; Tan, P.; Wang, L.; Jin, N.; Li, Y.; Zhang, L.; Yang, H.; Hu, Z.; Zhang, L.; Hu, C.; Li, C.; Qian, K.; Zhang, C.; Huang, Y.; Li, K.; Lin, H.; Wang, D. RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res.,* **2017**, *45*(D1), D135-D138.
[PMID: 27543076]

[38] Yi, Y.; Zhao, Y.; Li, C.; Zhang, L.; Huang, H.; Li, Y.; Liu, L.; Hou, P.; Cui, T.; Tan, P.; Hu, Y.; Zhang, T.; Huang, Y.; Li, X.; Yu, J.; Wang, D. RAID v2.0: an updated resource of RNA-associated interactions across organisms. *Nucleic Acids Res.,* **2017**, *45*(D1), D115-D118.
[http://dx.doi.org/10.1093/nar/gkw1052] [PMID: 27899615]

[39] Liang, Z.Y.; Lai, H.Y.; Yang, H.; Zhang, C.J.; Yang, H.; Wei, H.H.; Chen, X.X.; Zhao, Y.W.; Su, Z.D.; Li, W.C.; Deng, E.Z.; Tang, H.; Chen, W.; Lin, H. Pro54DB: a database for experimentally verified sigma-54 promoters. *Bioinformatics,* **2017**, *33*(3), 467-469.
[PMID: 28171531]

[40] Feng, P.; Ding, H.; Lin, H.; Chen, W. AOD: the antioxidant protein database. *Sci. Rep.,* **2017**, *7*(1), 7449.
[http://dx.doi.org/10.1038/s41598-017-08115-6] [PMID: 28784999]

[41] Peng, L.; Peng, M.M.; Liao, B.; Huang, G.H.; Li, W.B.; Xie, D.F. The Advances and Challenges of Deep Learning Application in Biological Big Data Processing. *Curr. Bioinform.,* **2018**, *13*(4), 352-359.
[http://dx.doi.org/10.2174/1574893612666170707095707]

[42] Patel, S.; Tripathi, R.; Kumari, V.; Varadwaj, P. DeepInteract: Deep Neural Network Based Protein-Protein Interaction Prediction Tool. *Curr. Bioinform.,* **2017**, *12*(6), 551-557.
[http://dx.doi.org/10.2174/1574893611666160815150746]

[43] Cao, R.Z.; Bhattacharya, D.; Hou, J.; Cheng, J.L. *DeepQA: improving the estimation of single protein model quality with deep belief networks*; BMC Bioinform, **2016**, p. 17.

[44] Zou, Q.; Xing, P.; Wei, L.; Liu, B. Gene2vec: Gene Subsequence Embedding for Prediction of Mammalian N6-Methyladenosine Sites from mRNA. *RNA,* **2018**.
[http://dx.doi.org/10.1261/rna.069112] [PMID: 30425123]

[45] Yu, L.; Sun, X.; Tian, S.W.; Shi, X.Y.; Yan, Y.L. Drug and Non-drug Classification Based on Deep Learning with Various Feature Selection Strategies. *Curr. Bioinform.,* **2018**, *13*(3), 253-259.
[http://dx.doi.org/10.2174/1574893612666170125124538]

[46] Wei, L.; Su, R.; Wang, B.; Li, X.; Zou, Q. Integration of deep feature representations and handcrafted features to improve the prediction of N 6 -methyladenosine sites. *Neurocomputing,* **2019**, *324*, 3-9.
[http://dx.doi.org/10.1016/j.neucom.2018.04.082]

[47] Long, H.X.; Wang, M.; Fu, H.Y. Deep Convolutional Neural Networks for Predicting Hydroxyproline in Proteins. *Curr. Bioinform.,* **2017**, *12*(3), 233-238.
[http://dx.doi.org/10.2174/1574893612666170221152848]

[48] Wei, L.; Ding, Y.; Su, R.; Tang, J.; Zou, Q. Prediction of human protein subcellular localization using deep learning. *J. Parallel Distrib. Comput.,* **2018**, *117*, 212-217.
[http://dx.doi.org/10.1016/j.jpdc.2017.08.009]