Computational Prediction of Protein Arginine Methylation Based on Composition–Transition–Distribution Features

Ruiyan Hou, Jin Wu, Lei Xu, Quan Zou,* and Yi-Jun Wu*

Cite This: ACS Omega 2020, 5, 27470−27479

ABSTRACT: Arginine methylation is one of the most essential protein post-translational modifications. Identifying the site of arginine methylation is a critical problem in biology research. Unfortunately, biological experiments such as mass spectrometry are expensive and time-consuming. Hence, predicting arginine methylation by machine learning is an alternative fast and efficient way. In this paper, we focus on the systematic characterization of arginine methylation with composition−transition−distribution (CTD) features. The presented framework consists of three stages. In the first stage, we extract CTD features from 1750 samples and exploit decision tree to generate accurate prediction. The accuracy of prediction can reach 96%. In the second stage, the support vector machine can predict the number of arginine methylation sites with 0.36 R-squared. In the third stage, experiments carried out with the updated arginine methylation site data set show that utilizing CTD features and adopting random forest as the classifier outperform previous methods. The accuracy of identification can reach 82.1 and 82.5% in single methylarginine and double methylarginine data sets, respectively. The discovery presented in this paper can be helpful for future research on arginine methylation.

1. INTRODUCTION

Protein post-translational modifications (PTMs) supply the proteome with various functionalities including governing cellular physiology and dynamics. PTM includes acetylation, ubiquitination, sulfation, methylation, and so forth. Methylation is one of the most common PTMs, and it regulates functional diversity in the cell. Methylation often modifies nitrogen atoms in arginine and lysine residues. Protein arginine methyltransferases catalyze arginine methylation and include two types. Type I mainly catalyzes the formation of asymmetric ω-N⁹, N⁵-dimethylarginine (sDMA), and N⁵-monomethylarginine (MMA). Type II catalyzes the formation of sDMA, symmetric ω-N⁹⁵, and MMA. Hence, single methyl groups or double methyl groups can be added onto arginine amino acid residues.

With progressive research, researchers have found that protein methylation is involved in human diseases such as rheumatoid arthritis, coronary heart disease, neurotic disorders, cancer, and multiple sclerosis. Therefore, it is important to accurately predict methylation sites to understand molecular mechanisms involved in protein methylation. Conventional experiments, such as ChIP-chip, probing with methylation-specific antibodies, and mass spectrometry, can identify protein methylation sites. However, they are labor-intensive, expensive, and time-consuming. With the advent of the big data era, considerable prediction tools based on machine learning are much more desirable for their accurate and fast prediction abilities.

In fact, several prediction methylation site methods have been developed in the past 10 years. Plewczynski et al. (2005) built the web server AutoMotif to predict methylation sites based on the hypothesis that PTMs mainly occur in disordered regions. Shao et al. (2009) incorporated a Bi-profile Bayes feature extraction method with a support vector machine (SVM) algorithm to identify arginine and lysine methylation. Shien et al. (2009) developed a predictor called MASA, which combines sequence information with structural characteristics such as secondary structure and accessible surface area (ASA).

Although considerable progress has been made in the development of existing methods, they still need to be improved. Their benchmark data sets should be updated as there is an increasing availability of methylation data. Wei et al. (2017) adopted random forest (RF) algorithm and built MePred-RF to predict arginine and lysine methylation sites.

Received: August 18, 2020
Accepted: October 6, 2020
Published: October 19, 2020

This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.
only with 185 true arginine-methylated peptides, whereas this study found 1785 reviewed arginine-methylated protein sequences in the UniProt database. Some feature extraction methods require disordered, evolutionary, and structural information. These methods cannot be widely used. In addition, the predictive work of the previous research has only focused on prediction of 11 or 41 peptides rather than whole protein sequences.

To overcome the above deficiencies, we collected 1785 reviewed arginine-methylation protein sequences from UniProt to form a positive data set and then produced 10,474 negative samples. We integrated composition-transition-distribution (CTD) features and different classifiers to identify arginine methylation sequences. Then, we exploited various regression algorithms to predict how many arginine methylation sites are in an arginine-methylation protein sequence. We combined the feature extraction method described above and different classifiers to identify specific arginine methylation sites by choosing sequences around methylation sites. The overall procedure is presented in Figure 1.

2. RESULTS AND DISCUSSION

2.1. Prediction of Methylarginine Proteins. We obtained data set1 according to the strategy described in the Materials and Methods section. Then, we extracted CTD features from protein sequences. After that, we employed 10-fold cross-validation to train four classifiers including k-nearest neighbor (KNN), decision tree (DT), SVM, and RF. We utilized sensitivity (SN), specificity (SP), accuracy (ACC), recall, F1-score, and area under curve (AUC) to assess the performances of four models. Table 1 shows that compared with other classification models, the ACC of DT is approximately 96% and the SP reaches 99%. The other assessment index also indicates that DT performs better than other classifiers when predicting methylarginine proteins (Table 1). It shows that the ACC of DT is superior by at least 5% over other classifiers, and its F1-score is significantly more than that of the other classifiers.

The method of feature extraction can generate 188-dimensional features. Then, which features are the most important? To solve this problem, an extremely popular method of dimension reduction was used to find several crucial features. As shown in Figure 3, it indicates that six of 188 features play vital roles in prediction of methylarginine. The six features are D120, D18, D119, D10, D1, and D135. As shown in Figure 3A, the accuracies of the top six features increase rapidly. Figure 3B,C provides the same result in F1 scores and AUC scores as the accuracies. From Figure 3D, we can see that the top six features play more important roles than other features. D1, D10, and D18 are frequencies of occurrence of arginine, lysine, and valine in the entire protein sequence. The frequencies of appearance of arginine are high as expected. D119 and D120 are features about charge property. It is believed that the charge has a high coefficient with isoelectric point. A study has shown that the isoelectric point plays an important role in arginine methylation. D135 is the feature about surface tension. According to the analysis mentioned above, charge and surface tension properties play significant roles in judging whether a protein is a methylarginine protein.

2.2. Prediction of the Number of Arginine-Methylated Sites. We used simple linear regression, nearest-neighbor regression, DT regression, and support vector regression (SVR) to judge the number of methylarginine sites in a protein. R-squared and mean squared errors were adopted to evaluate the performance of these models. As shown in Table 2, the R-squared values of the four models are −0.611, −0.33, 0.29, and 0.36. The R square of linear regression and DT regression are negative numbers, indicating that two models may not be optimal in this problem. The maximum and minimum numbers of arginine methylation sites in a protein are 30 and 1, respectively. The mean squared error of SVR is 3.43, and it only accounts for 10% of the maximum. However, it is three times the quantity of the

---

**Table 1. Results of Four Classifiers in Identifying Arginine Methylation Protein**

| classifiers | SN (%) | SP (%) | ACC (%) |
|-------------|--------|--------|---------|
| KNN         | 75.5   | 91.0   | 83.3    |
| DT          | 93.0   | 99.5   | 96.3    |
| SVM         | 87.1   | 93.9   | 90.5    |
| RF          | 91.1   | 90.8   | 91.2    |

---

**Figure 1. Roadmap of this study.**
minimum. This indicates that although SVR is the best among these four models, it needs to be further improved.

2.3. Prediction of Arginine Methylation Sites in a Protein Sequence.

In a methylarginine protein sequence, which arginine is modified by a methyl group? We hypothesize that it is related to the sequence around the central arginine. Therefore, we exploited a tool, WebLogo, to explore and represent significant differences for the motif selected in single-methylarginine, double-methylarginine, and negative samples.
The compositional preference for the arginine methylation sites is shown in Figure 4.

![Figure 4](https://example.com/fig4.png)

**Figure 4.** Compositional preference of peptide around central arginine in positive samples and negative samples.

The presented motifs are similar in single-methylarginine and double-methylarginine protein sequences (Figure 4). We can see that glycine (G) residues are enriched neighbors of the central site (R) both in single- and double-methylarginine protein sequences. However, negative samples are different from methylarginine protein sequences in position-specific preferences (Figure 4). Overall, these results indicate that amino acid residues around arginine assist in the accurate classification of true single- and double-methylarginine sites.

In single-methylarginine problems, we collected data according to the method description. Then, we extracted CTD features. 10-fold cross validation was utilized to train KNN, DT, SVM, and RF. The result of four classifiers is shown in Figure 5. The average accuracies of KNN, DT, SVM, and RF are 0.772, 0.735, 0.815, and 0.821, respectively, as shown in Figure 5A. The performance of RF is the best among these classifiers. Figure 5B,C illustrates that the F1 and AUC of the RF model are 0.894 and 0.821, respectively, which are higher than those of other classifiers. However, the recall of the KNN model is the highest among the four classifiers (Figure 5D).

To further assess the performance of different classifiers in single-methylarginine proteins, receiver operating characteristic (ROC) of four classifiers was plotted. As shown in Figure 6, RF is an effective classifier to identify single-methylarginine sites. Table 1 shows the experimental data on the best classifier RF in the single-methylarginine classification problem.

In addition, we employed mRMR to reduce dimensions from 188 features to 2 features and train the same classifiers with 10-fold cross validation. We chose D7 and D169 which have the highest scores. D7 represents the frequency of histidine in the whole sequence. D169 is a feature related to solvent accessibility. The result shows that solvent accessibility is an important feature, which is consistent with the previous research results. Protein methylation prefers to appear in the areas that are intrinsically disordered and easily accessible.23 As shown in Figure 5A, the accuracies are 0.587, 0.701, 0.701, and 0.703 in KNN, DT, SVM, and RF, respectively. The accuracies are lower than those for 188-dimensional features by 23.0, 4.6, 13.9, and 14.3%, respectively. The prediction results of two-dimensional features of SVM and RF are lower by approximately 10% than those including 188-dimensional features.

In double-methylarginine problems, we performed similar operation. The average accuracies are 0.773, 0.731, 0.821, and 0.825 for KNN, DT, SVM, and RF, respectively (Figure 7A).

![Figure 5](https://example.com/fig5.png)

**Figure 5.** Performances of the different classifiers in prediction of single-methylarginine sites. Comparison of ACC (A), F1 (B), AUC (C), and recall (D) of the four models acting on 188-dimensional features under 10-fold cross-validation.
As shown in Figure 7B,C, we achieve an F1 score of 0.825 and an AUC score of 0.9 using the RF model. Obviously, the KNN classifier shows the best recall, followed by SVM and RF classifiers according to Figure 7D. The result indicates that RF...
and the SVM are optimal models to predict methylarginine sites. RF performs slightly better than SVM.

To further assess whether CTD features can effectively represent 11 peptides in double-methylarginine problems, we adopted t-distributed stochastic neighbor embedding (t-SNE) to visualize the features in two-dimensional spaces. Figure 8 represents the features of 10 benchmark data sets for double-methylarginine using our feature extraction method; from this figure, we can see that most of the positive (true double-methylarginine sites) samples are clearly separated from the negative (non-double-methylarginine sites) samples.

The prediction of arginine methylation sites has been studied previously. It can be seen from the data in Table 3 that our study outperforms previous methods. We chose the best classifier, RF, for comparison with other models. CTD-RF achieved significantly better performance than MeMo; the average ACC of MeMo is lower than those of either of the CTD-RFs (Table 3). Though MePred-RF also adopted RF as the classifier, the method of feature extraction is different between the two studies. This illustrates that the method extracting CTD features is superior to MePred in extracting arginine methylation features.

## 3. CONCLUSIONS

The purpose of the current study was to choose an optimal classifier to identify methylarginine proteins, find an excellent model to predict the number of methylarginines in a protein sequence, and determine a classifier that is suitable to determine which arginine is modified by the methyl group. The present study establishes a quantitative data set for predicting methylarginine proteins, the number of arginine methylation sites, and loci of arginine methylation. The most obvious finding from this study is that DT can sometimes surpass popular classifiers such as RF and SVM to yield excellent results in identifying methylarginine proteins. The second major finding was that the SVR model is appropriate to predict the number of arginine methylation sites. The study also identified that SVM and RF are reliable predictors of methylation sites including single-methylarginine and double-methylarginine. The performance of RF is slightly superior to SVM. A limitation of this study is that prediction result is not sufficiently accurate in predicting the number of arginine-methylated sites. Further research is needed to establish a more effective model to predict the number of arginine methylation sites.

## 4. MATERIALS AND METHODS

### 4.1. Data set Acquisition

Data set 1 was utilized to predict proteins with arginine methylation. Identification of arginine methylation proteins is a binary classification problem used to decide whether or not a protein has arginine residues modified by methyl groups. In this study, arginine methylation proteins are regarded as positive samples and non-arginine methylation proteins as negative examples. We searched “methylarginine” in the UniProt database and obtained 1785 reviewed protein sequences used as positive examples. The negative examples were obtained according to the following method. The families including positive examples were obtained and then excluded from the PFAM database. We collected the longest protein sequence from the residual families. The sequences from the remaining 10,474 protein families were regarded as negative samples.

To assure the accuracy of the experimental results, we adopted CD-HIT to filter redundant samples with a threshold of 0.9 in the positive dataset. We eliminated surplus data with a threshold of 0.7 in the negative data set. The high-quality data set contained 857 positive samples and 9627 negative samples.

The proportion of negative to positive samples was approximately 11:1, which indicated that our data set was imbalanced. To solve this problem, we utilized k-means algorithm to cluster negative samples into 857 classes. Then, we extracted the longest sequence from each class as negative samples and combined 875 positive samples with 875 negative samples to form a balanced data set.

Data set 2 was utilized to predict the number of methylarginine sites in each protein sequence. Prediction of the number of methylarginine sites can be regarded as a regression problem that needs features and target values in the data set. We extracted CTD features of 857 positive samples mentioned above. Then, we used the programming language Python (version 3.7; Python Software Foundation, Wilmington, Delaware, USA) to extract methylarginine sites from UniProt. Then, the number of methylarginine sites was calculated in Python as the target values.

Data set 3 was used to determine which site contains methylarginine in a protein sequence. According to data set 1, we obtained 875 methylarginine proteins covering 4128 experimental methylarginine sites. Deciding whether a site is methylarginine should consider amino acid residues around methylation sites. Hence, we cut out 11 amino acid residues to form a window centered at the methylarginine site and filled in the rest with the character “B” when the peptide was shorter than 11 amino acid residues to obtain 4128 sequences. These 4128 sequences include 3038 ω-N-methylarginine sequences, 13 Nω-methylarginine sequences, 973 asymmetric dimethylarginine sequences, and 104 symmetric dimethylarginine sequences. We found that several sequences belong to both the ω-N-methylarginine group and the asymmetric dimethylarginine group. Therefore, we divided 4128 sequences into 2 groups: the single-methylarginine sites group and the double-methylarginine group. The single-methylarginine sites group includes ω-N-methylarginine and Nω-methylarginine. The double-methylarginine site group includes asymmetric dimethylarginine and symmetric dimethylarginine. A total of 3051 single-methylarginine sequences and 1077 double-methylarginine sequences were obtained.

According to the grouping mentioned above, we generated two classification problems for single-methylarginine and double-methylarginine. Subsequently, we generated an equal number of negative samples for these two classification problems and selected amino acid residues of arginine but not methylarginine as centers in the 875 methylarginine sequences. 11 amino acid residues around these centers were cut off to form a window, and the rest was filled by character
features. For the bigeminal group, we can obtain 3 features that are occurrence rates of the bigeminal sequence in every category. In conclusion, we can obtain \((3 + 15 + 3) = 21\) features for each physicochemical property, and \(8 \times 21 = 168\) features can be extracted from 8 properties. Finally, \(168 + 20 = 188\) features can be obtained from the methylarginine protein sequence.

### 4.3. Classifiers

KNN algorithm is a kind of supervised machine learning which can be applied to classification and regression predictive problems. It is a simple, nonparametric method. The workflow of KNN used in classification is as follows. First, it calculates the distance between the test sample and every training sample. Second, it finds the nearest \(k\) training sample neighbors of the test sample. Third, the test sample is identified as the class that is the most frequent class in the KNNs.

The main steps of KNN utilized in regression \(^{31}\) are as follows. The distance between the test sample and every training sample is calculated. The distance can be Euclidean’s distance, Manhattan distance, and so forth. Then, it the average distance from the nearest \(k\) training to yield a location of the prediction sample is calculated. A linearity is obtained according to these locations of prediction samples.

DT is a vital supervised machine learning method covering both classification and regression. \(^{32}\) DT can create a training model that can learn simple decision rules from prior data to predict the category or value of the target variable. \(^{33}\) DT builds a tree by using the attributes of training samples. \(^{34}\) The tree would grow leaves rather than nodes if all training samples are in the same class. Otherwise, the DT would select discriminatory attributes as new nodes. All of the training samples are divided into several groups and establish the branches of the DT. There are several groups here forming several branches. On the basis of branches obtained in the previous procedure, the procedures are repeated to build a tree. \(^{35}\)

DT begins with the root of the tree to predict a class label of test samples. Then, it compares the values of root attribute with the values of samples attribute and chooses the eligible branch to jump to the next node. Several essential algorithms of DT include CART, ID3, and C4.5. In this study, we used the default CART tree algorithm in the scikit-learn data mining package of Python.
SVM is one of the prevailing supervised machine learning models that was adopted to solve regression and classification problems. The object of the SVM algorithm is building a model that can assign test samples to one class or the other in classification problems. Each data point is plotted in N-dimensional spaces. The aim of SVM is to find a hyperplane in an N-dimensional space (where N is the number of features) that classifies the data points. The objective of the SVM algorithm is building a model that can find an optimal line or hyperplane to fit the data. Compared to the ordinary least squares, SVR minimizes coefficients.

RF is made up of numerous individual trees that work as an ensemble. In RF, we stochastically choose “M” features from a total of “n” features. The best split point is used to calculate the node “b” among the “M” features. Then, we use the best split to split the node into daughter nodes and repeat these steps until reaching “l” number of nodes. Through repeating the above procedures “n” times, we construct a forest with “n” number of DTs. Finally, bagging is adopted to combine the outputs of “n” number of DTs into a RF.

**4.4. Measurement.** In a classification model, a confusion matrix is an essential table that can visualize the performance of an algorithm. We show a confusion matrix for a binary classification model, a confusion matrix, and the performance of the model in our present study. Generally, higher assessment scores reflect better models.

There are two main metrics employed to evaluate the regression model. Mean squared error can measure how close a fitted line is to data points. It calculates the distance vertically from the point to the corresponding y value and then squares the value. All of these values for all data points are added up and averaged. The result is called mean squared error. R-squared is also called the coefficient of determination and can measure how close the data are to the fitted regression line. R-squared denotes the percentage of variance in the dependent variable that the independent variables explain collectively. The R-squared score varies between 0 and 1. R-squared is calculated as follows.

\[
R^2 = 1 - \frac{\sum_{i=0}^{m} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{m} (y_i - \bar{y})^2}
\]

where m is numbers of data points, \(y_i\) is the true value, and \(\hat{y}_i\) is the predicted value.

**DATA AVAILABILITY**

Datasets used in this paper are available at the website: https://github.com/Jenny-Jason/Arginine-methylationprediction-with-CTDfeatures.

**AUTHOR INFORMATION**

**Corresponding Authors**

*Quan Zou — Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China; [orcid.org/0000-0001-6406-1142; Email: zouquan@nclab.net]

*Yi-Jun Wu — Laboratory of Molecular Toxicology, State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China; [orcid.org/0000-0003-1064-2886; Email: wuyj@ioz.ac.cn]

**Authors**

*Ruiyan Hou — Laboratory of Molecular Toxicology, State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China; [orcid.org/0000-0002-1880-2664]

*Jin Wu — School of Management, Shenzhen Polytechnic, Shenzhen 518055, China

*Lei Xu — School of Electronic and Engineering, Shenzhen Polytechnic, Shenzhen 518055, China

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c03972

**Author Contributions**

*R.H. and J.W. contributed equally to this work.

**Notes**

The authors declare no competing financial interest.
ACKNOWLEDGMENTS

The authors are grateful to Shixin Jin for his assistance with data collection and Yujia Xiang and Shida He for their helpful discussions. This work was supported in part by the grant from the National Natural Science Foundation of China (no. 31672366).

REFERENCES

(1) Mann, M.; Jensen, O. N. Proteomic analysis of post-translational modifications. Nat. Biotechnol. 2003, 21, 255–261.
(2) Bannister, A. J.; Kouzarides, T. Reversing histone methylation. Nature 2005, 436, 1103–1106.
(3) Pahlich, S.; Zakaryan, R. P.; Gehring, H. Protein arginine methylation: Cellular functions and methods of analysis. Biochim. Biophys. Acta 2006, 1764, 1890–1903.
(4) Suzuki, A.; Yamada, R.; Yamamoto, K. Citrullination by peptidylarginine deiminase in rheumatoid arthritis. Ann. N.Y. Acad. Sci. 2007, 1108, 323–339.
(5) Chen, X.; Niromand, F.; Liu, Z.; Zankl, A.; Katus, H. A.; Jahn, L.; Tiefenbacher, C. P. Expression of nitric oxide related enzymes in coronary heart disease. Basic Res. Cardiol. 2006, 101, 346–353.
(6) Longo, V. D.; Kennedy, B. K. Sirtuins in aging and age-related disease. Cell 2006, 126, 257–268.
(7) Liu, C.; Chyr, J.; Zhao, W.; Xu, Y.; Ji, Z.; Tan, H.; Soto, C.; Zhou, X. Genome-wide association and mechanistic studies indicate that immune response contributes to Alzheimer’s disease development. Front. Genet. 2018, 9, 410.
(8) Wang, Y.; Zhang, S.; Li, F.; Zhou, Y.; Zhang, Y.; Wang, Z.; Zhang, R.; Zhu, J.; Ren, Y.; Tan, Y.; Qin, C.; Li, Y.; Li, X.; Chen, Y.; Zhu, F. Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. Nucleic Acids Res. 2020, 48, D1031–D1041.
(9) Yin, J.; Sun, W.; Li, F.; Hong, J.; Li, X.; Zhou, Y.; Lu, Y.; Liu, M.; Zhang, X.; Chen, N.; Jin, X.; Xue, J.; Zeng, S.; Yu, L.; Zhu, F. VARIDT 1.0: variability of drug transporter database. Nucleic Acids Res. 2020, 48, D1171.
(10) Mastronardi, F. G.; Wood, D. D.; Mei, J.; Rajmakers, R.; Tseveleki, V.; Dosch, H.-M.; Probert, L.; Casacca-Bonnefille, P.; Moscarello, M. A. Increased citrullination of histone H3 in multiple sclerosis brain and animal models of demyelination: a role for tumor necrosis factor-induced peptidylarginine deiminase 4 translocation. J. Neurosci. 2006, 26, 11387–11396.
(11) Johnson, D. S.; Li, W.; Gordon, D. B.; Bhattacharjee, A.; Curry, B.; Ghosh, J.; Brizuela, L.; Carroll, J. S.; Brown, M.; Flicek, P.; Koch, C. M.; Dunham, I.; Bieda, M.; Xu, X.; Farnham, P. J.; Kapranov, P.; Nis, D. A.; Gingeras, T. R.; Zhang, X.; Holster, H.; Jiang, N.; Green, R. D.; Song, J. S.; McGuire, S. A.; Anton, E.; Nguyen, L.; Trinklein, N. D.; Ye, Z.; Ching, K.; Hawkins, D.; Ren, B.; Scacheri, P. C.; Rozowsky, J.; Karpikov, A.; Euskirchen, G.; Weissman, S.; Gerstein, M.; Snyder, M.; Yang, A.; Moqtaderi, Z.; Hirsch, H.; Shulha, H. P.; Fu, Y.; Weng, Z.; Struhl, K.; Myers, R. M.; Lieb, J. D.; Liu, X. S. Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. Genome Res. 2008, 18, 393–403.
(12) Boisvert, F.-M.; Côté, J.; Boulanger, M.-C.; Richard, S. A proteomic analysis of arginine-methylated protein complexes. Mol. Cell. Proteomics 2003, 2, 1319–1330.
(13) Ong, S.-E.; Mittler, G.; Mann, M. Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. Nat. Methods 2004, 1, 119–126.
(14) Zhang, F.; Ma, A.; Wang, Z.; Ma, Q.; Liu, B.; Huang, L.; Wang, Y. A central edge selection based overlapping community detection algorithm for the detection of overlapping structures in protein-protein interaction networks. Molecules 2018, 23, 2633.
(15) Plewczynski, D.; Tkaczy, A.; Wyrwicz, L. S.; Rychlewski, L. AutoMotif server: prediction of single residue post-translational modifications in proteins. Bioinformatics 2005, 21, 2525–2527.
(16) Shao, J.; Xu, D.; Tsai, S.-N.; Wang, Y.; Ngai, S.-M. Computational identification of protein methylation sites through bi-profile Bayes feature extraction. PLoS ONE 2009, 4, No. e4920.
(17) Shien, D.-M.; Lee, T.-Y.; Chang, W.-C.; Hsu, J. B.-K.; Horng, J.-T.; Hsu, P.-C.; Wang, T.-Y.; Huang, H.-D. Incorporating structural characteristics for identification of protein methylation sites. J. Comput. Chem. 2009, 30, 1532–1543.
(18) Wei, L.; Xing, P.; Shi, G.; Ji, Z.; Zou, Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. IEEE/ACM Trans. Comput. Biol. Bioinf. 2017, 16, 1264–1274.
(19) Pawan, K.; Joseph, J.; Ashutosh, P.; Dinesh, G. PRMenPred: A protein arginine methylation prediction tool. PLoS ONE 2017, 12, No. e0183318.
(20) Uhlmann, T.; Geoghegan, V. L.; Thomas, B.; Ridlova, G.; Trudgian, D. C.; Acuto, O. A method for large-scale identification of protein arginine methylation. Mol. Cell. Proteomics 2012, 11, 1489–1499.
(21) Crooks, G. E.; Hon, G.; Chandonia, J. M.; Brenner, S. WebLogo: a sequence logo generator. Genome Res. 2004, 14, 1188–1190.
(22) Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 2005, 27, 1226–1238.
(23) Liu, F.; Li, C.; Wang, M.; Webb, G. I.; Zhang, Y.; Whistock, J. C.; Song, J. GlycoMine: a machine learning-based approach for predicting N- and O-linked glycosylation in the human proteome. Bioinformatics 2015, 31, 1411–1419.
(24) van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. J. Mach. Learn. Res. 2008, 9, 2579–2605.
(25) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012, 28, 3150–3152.
(26) Xu, L.; Liang, G.; Liao, C.; Chen, G.-D.; Chang, C.-C. An efficient classifier for Alzheimer’s disease gene identification. Molecules 2018, 23, 3140.
(27) Chu, Y.; Kaushik, A. C.; Wang, X.; Wang, W.; Zhang, Y.; Shan, X.; Salalub, D. R.; Xiong, Y.; Wei, D.-Q. DTV-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. Briefings Bioinf. 2019, bbz152.
(28) Tan, J.-X.; Li, S. H.; Li, S.-H.; Zhang, Z.-M.; Chen, C.-X.; Chen, W.; Tang, H.; Lin, H. Identification of hormone binding proteins based on machine learning methods. Math. Biosci. Eng. 2019, 16, 2466–2480.
(29) Tang, J.; Fu, J.; Wang, Y.; Li, B.; Li, Y.; Yang, Q.; Cui, X.; Hong, J.; Li, X.; Chen, Y.; Xue, W.; Zha, F. ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. Briefings Bioinf. 2020, 21, 621–636.
(30) Yu, B.; Qiu, W.; Chen, C.; Ma, A.; Jiang, J.; Zhou, H.; Ma, Q. SubMit-oXBoost: predicting protein submicrotubular localization by fusing multiple feature information and eXtreme gradient boosting. Bioinformatics 2020, 36, 1074–1081.
(31) Liao, Y.; Vemuri, V. R. Use of k-nearest neighbor classifier for intrusion detection. Comput. Secur. 2002, 21, 439–448.
(32) Cheng, L.; Hu, Y.; Sun, J.; Zhou, M.; Jiang, Q. DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. Bioinformatics 2018, 34, 1953–1956.
(33) Friedel, M. A.; Brodley, C. E. Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. 1997, 61, 399–409.
(34) Habibi, S.; Ahmadi, M.; Alizadeh, S. Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. Glob. J. Health Sci. 2015, 7, 304.
(35) Zou, Q.; Qiu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting diabetes mellitus with machine learning techniques. Front. Genet. 2018, 9, 515.
(36) Xu, H.; Zeng, W.; Zhang, D.; Zeng, X. MOEA/HD: A multiobjective evolutionary algorithm based on hierarchical decomposition. *IEEE Trans. Cybern.* 2019, 49, 517–526.

(37) He, J.; Fang, T.; Zhang, Z.; Huang, B.; Zhu, X.; Xiong, Y. PseUI: Pseudouridine sites identification based on RNA sequence information. *BMC Bioinf.* 2018, 19, 306.

(38) Xu, L.; Liang, G.; Shi, S.; Liao, C. SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. *Int. J. Mol. Sci.* 2018, 19, 1773.

(39) Xu, L.; Liang, G.; Wang, L.; Liao, C. A novel hybrid sequence-based model for identifying anticancer peptides. *Genes* 2018, 9, 158.

(40) Lai, H.-Y.; Zhang, Z.-Y.; Su, Z.-D.; Su, W.; Ding, H.; Chen, W.; Lin, H. iProEP: A computational predictor for predicting promoter. *Mol. Ther.–Nucleic Acids* 2019, 17, 337–346.

(41) Lv, H.; Diao, F.-Y.; Zhang, D.; Guan, Z.-X.; Yang, H.; Su, W.; Liu, M.-L.; Ding, H.; Chen, W.; Lin, H. idNA-MS: An integrated computational tool for detecting DNA modification sites in multiple genomes. *iScience* 2020, 23, 100991.

(42) Yang, Q.; Li, B.; Tang, J.; Cui, X.; Wang, Y.; Li, X.; Hu, J.; Chen, Y.; Xue, W.; Lou, Y.; Qiu, Y.; Zhu, F. Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. *Briefings Bioinf.* 2020, 21, 1058–1068.

(43) Stephenson, N.; Shane, E.; Chase, J.; Rowland, J.; Ries, D.; Justice, N.; Zhang, J.; Chan, L.; Cao, R. Survey of machine learning techniques in drug discovery. *Curr. Drug Metab.* 2019, 20, 185–193.

(44) Xu, L.; Liang, G.; Liao, C.; Chen, G.-D.; Chang, C.-C. K-skip-n-gram: A random Forest based method for Alzheimer’s disease protein identification. *Front. Genet.* 2019, 10, 33.

(45) Zeng, X.; Zhu, S.; Liu, X.; Zhou, Y.; Nussinov, R.; Cheng, F. deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 2019, 35, 5191–5198.

(46) Wang, H.; Ding, Y.; Tang, J.; Guo, F. Identification of membrane protein types via multivariate information fusion with Hilbert–Schmidt independence criterion. *Neurocomputing* 2020, 383, 257–269.

(47) Yang, Q.; Wang, Y.; Zhang, Y.; Li, F.; Xia, W.; Zhou, Y.; Qiu, Y.; Li, H.; Zhu, F. NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. *Nucleic Acids Res.* 2020, 48, W436–W448.

(48) Xu, Y.; Guo, M.; Liu, X.; Wang, C.; Liu, Y.; Liu, G. Identify bilayer modules via pseudo-3D clustering: applications to miRNA-gene bilayer networks. *Nucleic Acids Res.* 2016, 44, No. e152.

(49) Xu, Y.; Wang, Y.; Luo, J.; Zhao, W.; Zhou, X. Deep learning of the splicing (epi)genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision. *Nucleic Acids Res.* 2017, 45, 12100–12112.

(50) Cheng, L.; Yang, H.; Zhao, H.; Pei, X.; Shi, H.; Sun, J.; Zhang, Y.; Wang, Z.; Zhou, M. MetSigDis: a manually curated resource for the metabolic signatures of diseases. *Briefings Bioinf.* 2019, 20, 203–209.

(51) Ding, Y.; Tang, J.; Guo, F. Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 2019, 325, 211–224.

(52) Hong, J.; Luo, Y.; Zhang, Y.; Ying, J.; Xue, W.; Xie, T.; Tao, L.; Zhu, F. Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. *Briefings Bioinf.* 2020, 21, 1437–1447.

(53) Chen, W.; Nie, F.; Ding, H. Recent advances of computational methods for identifying bacteriophage virion proteins. *Protein Pept. Lett.* 2020, 27, 259–264.

(54) Li, Y. H.; Li, X. X.; Hong, J. J.; Wang, Y. X.; Fu, J. B.; Yang, H.; Yu, C. Y.; Li, F. C.; Hu, J.; Xue, W. W.; Jiang, Y. Y.; Chen, Y. Z.; Zhu, F. Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. *Briefings Bioinf.* 2020, 21, 649–662.

(55) Zeng, X.; Wang, W.; Chen, C.; Yen, G. G. A consensus community-based particle swarm optimization for dynamic community detection. *IEEE Trans. Cybern.* 2020, 50, 2502–2513.