k-Skip-n-Gram-RF: A Random Forest Based Method for Alzheimer’s Disease Protein Identification

Lei Xu¹, Guangmin Liang ¹*, Changrui Liao², Gin-Den Chen³ and Chi-Chang Chang⁴,⁵*

¹ School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China, ² Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Optoelectronic Engineering, Shenzhen University, Shenzhen, China, ³ Department of Obstetrics and Gynecology, Chung Shan Medical University Hospital, Taichung, Taiwan, ⁴ School of Medical Informatics, Chung Shan Medical University, Taichung, Taiwan, ⁵ IT Office, Chung Shan Medical University Hospital, Taichung, Taiwan

In this paper, a computational method based on machine learning technique for identifying Alzheimer’s disease genes is proposed. Compared with most existing machine learning based methods, existing methods predict Alzheimer’s disease genes by using structural magnetic resonance imaging (MRI) technique. Most methods have attained acceptable results, but the cost is expensive and time consuming. Thus, we proposed a computational method for identifying Alzheimer disease genes by use of the sequence information of proteins, and classify the feature vectors by random forest. In the proposed method, the gene protein information is extracted by adaptive k-skip-n-gram features. The proposed method can attain the accuracy to 85.5% on the selected UniProt dataset, which has been demonstrated by the experimental results.

Keywords: Alzheimer’s disease, n-gram model, random forest, gene coding, sequence information

INTRODUCTION

Alzheimer’s disease (AD) is a common cause of dementia, and it can lead a degeneration of brain. The research shows that more than 35 million people have been affected by Alzheimer’s disease all over the world. It is predicted that there will be over 70 million people diagnosed by Alzheimer’s disease in 2030, and the number will be increased by 50% in 2050 (Brookmeyer et al., 2007).

Until now, there is no treatment for AD. As the status becoming worse, it will destroy the ability of speak and think. At last, AD will lead to die. So, it is meaningful to predict AD at an early stage. Machine learning methods have been extensively used in multiple fields of bioinformatics (Zeng et al., 2014; Wang et al., 2016; Liu Y. et al., 2017; Zhang et al., 2017a; Cheng et al., 2018; Fu et al., 2018; Liu et al., 2018; Peng et al., 2018a; Song et al., 2018), such as anticancer peptides prediction (Xu et al., 2018b), identification of antioxidant proteins (Xu et al., 2018a), disease gene identification (Jiang et al., 2017; Liu G. et al., 2017; Peng et al., 2017; Zeng et al., 2017a; Zhang et al., 2017b; Cheng et al., 2018,a,b; Liu et al., 2018a,b; Zhu et al., 2018), microRNA classification (Wei et al., 2014; Chen et al., 2016; Zeng et al., 2018; Zhang et al., 2018), protein remote homology detection (Liu et al., 2014b; Liu and Li, 2018), drug-induced hepatotoxicity prediction (Li et al., 2018; Su et al., 2018), DNA binding protein identification (Zhang and Liu, 2017; Liu, 2018), protein interaction identification (Guo et al., 2011, 2012, 2014; Ding et al., 2016, 2017a,b; Peng et al., 2018b) and so on (Li et al., 2016, 2017; Zou et al., 2016; Zeng et al., 2017a,b; Hu et al., 2018; Xue et al., 2018; Zhang and Liu, 2018; Zhang et al., 2018). In this paper, machine learning method is used to identify AD.

Because the structural features of brain is related to AD, the structural brain information is described by structural magnetic resonance imaging (MRI) data. Most existing works use machine learning methods, such as ensemble classifier, deep learning method to classify AD and non-AD
samples. However, most existing works are limited by the expensive cost on money and time. For the purpose of identifying AD efficiently and effectively, a method called k-skip-n-gramRF, which is based on gene coding information of proteins, is proposed to recognize AD samples. In this paper, adaptive k-skip-2-gram is used to extract the information from the protein sequences, and then the samples are classified by random forest (RF) classifier. Consequently the classification accuracy can attain the accuracy to 85.5% using the select data set from Uniprot database. In our proposed method, adaptive k-skip-n-gram describes the correlation information of both adjacent and non-adjacent residues based on traditional n-gram model (Wei et al., 2017a). The idea of our proposed method is shown in Figure 1. As Figure 1 shown, the protein peptides are extracted by k-skip-n-gram method. Each sequence is transferred into a vector. The training vectors are used to train the parameters of random forest. The performance of methods is evaluated by testing vectors. The testing vectors are labeled by trained random forest classifier.

In the proposed method, adaptive n-gram-k-skip model is used to represent the gene coding information by a 400-dimensional vector. Then an ensemble classifier named random forest (RF) is used to classify the samples. In the experiments, the accuracy of the proposed classifier is 85.5%, which is competitive to existing works with low cost. In other words, the experimental results demonstrated that the proposed methods can be utilized to identify AD samples. The contributions of our work include:

- A computational model for predicting Alzheimer's disease is proposed in the paper. The experimental results demonstrate that the classification accuracy of the prediction model is 85.5%, which is competitive to some existing works with low cost and fast speed.
- Different from previous work using MRI data, the gene coding information of proteins is considered to identify Alzheimer’s disease protein. Each protein sequence is represented by a 400-dimensional vector, which the information of distance is considered.
- In our work, random forest is used to classify the AD protein peptides and non-AD protein peptides. Random forest is an ensemble classification method based on bagging, which is used to predict AD peptides in the work.

The rest of the paper is organized as follows. Section Materials and Methods introduces the dataset and the proposed method (k-skip-n-gram-RF) for identifying AD peptides. The results of AD prediction are described in Section Results and Discussions. The conclusion is made in Section Conclusions.

---

**FIGURE 1** | The flow chart of proposed method.
MATERIALS AND METHODS

Benchmark Dataset
The used data is selected from the UniProt database. The data set \( S \) is composed of positive samples \( S^+ \) and negative samples \( S^- \). The positive sample set is represented by Alzheimer's disease (AD) samples, and the negative sample set is represented by non-AD samples.

Positive Data Set
The positive data set contains AD samples. The samples are built by the sequences which are labeled by “Alzheimer’s disease.” As a result, 310 AD samples are selected from the UniProt database. To avoid the overestimation of the performance, the sequences with more than 60% similarity are removed. Thus, there are 279 positive samples left.

Negative Data Set
The data labeled with “non Alzheimer’s disease” are chosen then there are 312 non-AD samples. The proteins which are confirmed as non Alzheimer’s disease are also selected in the negative data set. After CD-HIT program (Fu et al., 2012), 1,743 negative samples are left in the benchmark data set for experiments.

In the experiments, the benchmark data set is divided into training data set and testing data set. The training data are used for train the classifier, and the testing data are used for the performance evaluation.

Random Forest
Random forest (Ho, 1995) is an ensemble classifier by combining decision trees together. Due to its effectiveness, random forest has been widely used in many bioinformatics problems (Deng and Chen, 2015; Liu, 2018). The key idea will be introduced briefly here.

The key element in random forest is decision tree. The decision trees are built based on bagging. Bagging is a sampling method. The used samples will be put back into the data set for reusing. In other words, a sample may be used more than one time for building data set. For example, there is a data set with \( n \) samples. If \( m \) decision trees are needed, \( m \) data sets will be built by bagging for training. Each node on the decision tree is represented by a feature used for classification (Quinlan, 1986).

The distance between residues \( R_i \) and \( R_j \) is calculated as Equation (3). For example, if there is \( i = 2 \) and \( j = 3 \), the distance between \( A_2 \) and \( A_3 \) is 0. The distance between \( A_4 \) and \( A_2 \) is 1, because \( A_2 \) and \( A_4 \) are separated by \( A_3 \).

\[
\text{Dis} = |j - i - 1|
\]  

\[\text{FIGURE 2} \quad \text{The voting process of random forest.}\]

The attribute with the maximum entropy gain is selected. Random forest is built based on the decision tree, so the features with larger information gains will be selected first in the training process. Because random forest is an ensemble classifier, the decision is made by voting process shown as Figure 2. As Figure 2 shown, the sample will be assigned to the class with the maximum votes. In our problem, the decision trees are trained by protein sequences, and the input of Figure 2 is protein peptide.

Sequence Representation
The sequence information of each protein peptides is encoded into a 400-dimensional feature vector by adaptive k-skip-2-gram model (Wei et al., 2017a). The k-skip-2-gram method is proposed based on k-skip-n-gram. The key idea of k-skip-n-gram is the distance information integrated into traditional n-gram model (Liu et al., 2014a). However, the maximum value of \( k \) is the length of the shortest amino acid, which is small for long peptides. Thus, for the purpose of mining more relation between peptides, adaptive k-skip-n-gram method is proposed in and used in our method.

There is a peptide sequence \( S \), denoted by \( R_1 R_2 \ldots R_n \), where \( n \) is the length of the sequence. In n-gram model, the occurrence frequencies of any \( n \) consecutive amino acids are measured. The amino acid set is denoted as \( L \), where \( L_i \) is the \( i \)th element in \( L \). The n-gram features can be calculated as follows:

\[
F_{n, \text{gram}} = \frac{N(T_{m1} T_{m2} \ldots T_{mn})}{N(T_s)}
\]

Where \( N(T_i) \) is denoted as the number of all items in the set \( T_i \), and \( T_s \) represents the number of segments with \( n \) consecutive amino acids in the peptide \( S \). Each position has 20 possible amino acids, so there are \( 20^n \) dimensions with the length of \( n \) peptides in n-gram model. It is obvious that the features are sparse. Thus, n-gram-k-skip is proposed to overcome the sparse problem of n-gram model. The distance information is considered in k-skip-n-gram model.

In k-skip-n-gram model, distance between residues \( R_i \) and \( R_j \) is calculated as Equation (3). For example, if there is \( i = 2 \) and \( j = 3 \), the distance between \( A_2 \) and \( A_3 \) is 0. The distance between \( A_4 \) and \( A_2 \) is 1, because \( A_2 \) and \( A_4 \) are separated by \( A_3 \).

\[
\text{Dis} = |j - i - 1|
\]  

\[\text{FIGURE 2} \quad \text{The voting process of random forest.}\]

The used data is selected from the UniProt database. The data set \( S \) is composed of positive samples \( S^+ \) and negative samples \( S^- \). The positive sample set is represented by Alzheimer’s disease (AD) samples, and the negative sample set is represented by non-AD samples.

The key idea will be introduced briefly here.

The used samples will be put back into the data set for reusing. In other words, a sample may be used more than one time for building data set. For example, there is a data set with \( n \) samples. If \( m \) decision trees are needed, \( m \) data sets will be built by bagging for training. Each node on the decision tree is represented by a feature used for classification (Quinlan, 1986).

The features used on different levels of the tree are selected briefly here.

The attribute with the maximum entropy gain is selected. Random forest is built based on the decision tree, so the features with larger information gains will be selected first in the training process. Because random forest is an ensemble classifier, the decision is made by voting process shown as Figure 2. As Figure 2 shown, the sample will be assigned to the class with the maximum votes. In our problem, the decision trees are trained by protein sequences, and the input of Figure 2 is protein peptide.

Sequence Representation
The sequence information of each protein peptides is encoded into a 400-dimensional feature vector by adaptive k-skip-2-gram model (Wei et al., 2017a). The k-skip-2-gram method is proposed based on k-skip-n-gram. The key idea of k-skip-n-gram is the distance information integrated into traditional n-gram model (Liu et al., 2014a). However, the maximum value of \( k \) is the length of the shortest amino acid, which is small for long peptides. Thus, for the purpose of mining more relation between peptides, adaptive k-skip-n-gram method is proposed in and used in our method.

There is a peptide sequence \( S \), denoted by \( R_1 R_2 \ldots R_n \), where \( n \) is the length of the sequence. In n-gram model, the occurrence frequencies of any \( n \) consecutive amino acids are measured. The amino acid set is denoted as \( L \), where \( L_i \) is the \( i \)th element in \( L \). The n-gram features can be calculated as follows:

\[
F_{n, \text{gram}} = \frac{N(T_{m1} T_{m2} \ldots T_{mn})}{N(T_s)}
\]

Where \( N(T_i) \) is denoted as the number of all items in the set \( T_i \), and \( T_s \) represents the number of segments with \( n \) consecutive amino acids in the peptide \( S \). Each position has 20 possible amino acids, so there are \( 20^n \) dimensions with the length of \( n \) peptides in n-gram model. It is obvious that the features are sparse. Thus, n-gram-k-skip is proposed to overcome the sparse problem of n-gram model. The distance information is considered in k-skip-n-gram model.

In k-skip-n-gram model, distance between residues \( R_i \) and \( R_j \) is calculated as Equation (3). For example, if there is \( i = 2 \) and \( j = 3 \), the distance between \( A_2 \) and \( A_3 \) is 0. The distance between \( A_4 \) and \( A_2 \) is 1, because \( A_2 \) and \( A_4 \) are separated by \( A_3 \).

\[
\text{Dis} = |j - i - 1|
\]  

\[\text{FIGURE 2} \quad \text{The voting process of random forest.}\]

| TABLE 1 | The performance evaluation of k-skip-n-gram-RF. |
|---------|-----------------|-----------------|-----------------|
|         | Sn       | Sp       | Acc    |
| k-skip-n-gram-RF | 0.855   | 0.855   | 0.855  |
In the k-skip-n-gram model, the sequence information of n residues within distance k is calculated, which means that the only residues within distance k are considered. The calculation of k-skip-n-gram is shown as Equation (4). In Equation (4), N(T_{SkipG}) is denoted as all the elements in T_{SkipG}. The calculation of T_{SkipG} is shown as Equation (5). In Equation (5), Skip(DT = z) = \{A_i A_{i+z+1} \ldots A_{i+z+n−1} | 1 \leq z \leq L − 1, 1 \leq z \leq k\}. When n equals 1, the model is reduced to n-gram model. For the purpose of avoiding overfitting problem, n is constrained less than 3. Thus, only the case of n equals to 2 is analyzed. The model is considered as k-skip-2-gram. The elements of k-skip-2-gram include R_1 R_2, R_2 R_3, ..., R_{n−1} R_n, R_1 R_3, ..., R_{n−2} R_n, ..., R_1 R_n, which all of them are two amino acids pair within distance k. The number of the 2-item combination is 400. Thus, the number of features extracted by k-skip-2-gram is 400. The 20^n dimensional vector is reduced to a 400 dimensional vector.

\[
fv = \left\{ \frac{N(L_{m1}L_{m2} \ldots L_{mn})}{N(T_{SkipG})} \right\}
\]

\[
T_{SkipG} = (\cup_{z=1}^{k} \text{Skip}(DT = z))
\]

The amino acids distance within k is calculated. K is the minimum sequence length of the peptides. The length of some sequences is sometimes short. If k is small, the features will be limited in local information. In adaptive k-skip-n-gram, k is the length of each sequence. When the information of varying distances of sequences is described, adaptive k-skip-n-gram is more flexible than k-skip-n-gram.

![Performance Evaluation](image)

The performance of our proposed method is reported in Section The Performance of Proposed Method. The performance of our method compared with other feature selection methods is described in Section The Comparison of Performance Evaluation on Feature Extraction Methods. We also compared random forest with other classifiers, and the performance evaluation comparison is shown in Section The Comparison of Performance Evaluation on Other Classifiers.

### RESULTS AND DISCUSSIONS

In the experiments, the data set is divided into training set and testing set. The training set is used for learning parameters, and the testing set is for performance evaluation. The performance of our proposed method is reported in Section The Performance of Proposed Method. The performance of our method compared with other feature selection methods is described in Section The Comparison of Performance Evaluation on Feature Extraction Methods. We also compared random forest with other classifiers, and the performance evaluation comparison is shown in Section The Comparison of Performance Evaluation on Other Classifiers.

### The Performance of Proposed Method

The experimental results of our proposed method are reported in Table 1. Table 1 shows that the accuracy of our proposed method is 0.855, which means that the proposed method can classified the 85.5% samples correctly in the benchmark data set. Sp(specificity) describes the performance for identifying AD samples. 85.5% AD samples of all the positive samples in the data set will be recognized in the experiment. Moreover, 85.5% non-AD samples of the negative samples can be classified correctly by the proposed method. The experimental results demonstrated that the method is practical.

### The Comparison of Performance Evaluation on Feature Extraction Methods

For the purpose of showing the effective performance of our proposed method, the feature select method of our proposed
method is compared with information theory. Information theory is a feature selection method representing.

The comparison results are shown in Table 2. The metrics of accuracy, sp and sn of k-skip-n-gram method performs better than that of information theory. The accuracy of the proposed method (k-skip-n-gram-RF) is better than that of information theory based random forest. The accuracy of information theory method is 0.715, while the accuracy of k-skip-n-gram is 0.855. For the problem of Alzheimer’s disease protein prediction, k-skip-n-gram performs better than information theory when random forest is used.

The Comparison of Performance Evaluation on Other Classifiers

To demonstrate the performance of our classifier, the classification methods are compared with other classification method, such as naive bayes (Peter Norvig, 1995), LibD3C (Lin et al., 2014), Adaboost (Rojas, 2009), and bagging. The mentioned methods are shown as followings.

Naive bayes is probability method. The sample is labeled by the class with the maximum probability.

LibD3C is an ensemble based method. k-means is integrated into the method for classifier selection.

Adaboost and bagging are ensemble classification algorithms. The difference between them is the building strategy of sample set. Bagging reuses the samples during classification. In Adaboost, the samples classified to the wrong class, the weight will be increased. The samples which are classified into the correct class, the weight will be decreased.

The comparison of our proposed method with other classifiers on Sn, Sp and accuracy is shown in Figure 3. Random forest performs better than other classifiers on Sn, Sp and Acc. Random forest is an ensemble classifier with competitive performance. The accuracy of naïve bayes and is 0.801 and 0.812. The accuracy of bagging and LibD3C is 0.83 and 0.837. When the features of proteins are extracted by k-skip-n-gram method, the accuracy of random forest is 0.855, which is better than that of other methods.

As the results shown, LibD3C preforms better than naïve bayes, Adaboost and bagging.

The feature selection method and classifiers are compared. The k-skip-n-gram method can represent more accurately than information theory. Random forest performs better than other classifiers. Thus, the experimental results demonstrate that our proposed method can provide competitive results.

CONCLUSIONS

In this paper, we study the problem of Alzheimer’s disease prediction by using gene coding information. MRI data are usually used to identifying Alzheimer’s disease, but the data are complex and with expensive cost. Different from previous work, the information is extracted from the protein peptides by k-skip-n-gram model for Alzheimer’s disease prediction. Finally, random forest is used to classify Alzheimer’s disease (AD) samples and non-AD samples. The accuracy of our proposed method is 85.5%, which means that it is meaningful for Alzheimer’s disease detection with low cost. In the literature, most work have provided web server for protein classification, and we will develop our web server for Alzheimer’s disease proteins classification. Moreover, to further improve the prediction performance, there are still many aspects can be explored. For example, other effective machine learning algorithms, such as ensemble learning algorithms and deep learning algorithms, have recently showed they can achieve better performance than traditional algorithms (Mrozek et al., 2009a; Wei et al., 2017b, 2018a; Wang et al., 2018). On the other hand, feature representation learning has demonstrated that it can exploit more informative features and improve the performance in multiple bioinformatics problems (Mrozek et al., 2009b; Momot et al., 2010; Liu et al., 2015; Wei et al., 2018b, 2019; Tang et al., 2019).

AUTHOR CONTRIBUTIONS

LX initially drafted the manuscript and did most of the codes work and the experiments. CL and G-DC collected the features, analyzed the experiments and revised the paper. GL and C-GC revised to draft the manuscript. All authors designed the work, read and approved the final manuscript and are agree to be accountable for all aspects of the work.

ACKNOWLEDGMENTS

This work was supported by the natural science foundation of Guangdong province (grant no. 2018A0303130084), the Science and Technology Innovation Commission of Shenzhen (grant no. JCYJ201708181906219062, JCYJ201605232113602609), the Grant of Shenzhen Polytechnic (grant no. 601822K19011), and National Nature Science Foundation of China (grant nos. 61575128, 61771331).

REFERENCES

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., and Arrighi, H. M. (2007). O1-02-01: Forecasting the global prevalence and burden of Alzheimer’s disease. Alzheimer’s Dement. 3:5168. doi: 10.1016/j.jalz.2007.04.381
Chen, J., Wang, X., and Liu, B. (2016). iMiRNA-SSF: improving the identification of microRNA precursors by combining negative sets with different distributions. Sci. Rep. 6:19062. doi: 10.1038/srep19062
Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018a). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. Bioinformatics. 34, 1953–1956. doi: 10.1093/bioinformatics/bty002
Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., et al. (2018b). InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. BMC Genomics 19(Suppl. 1):919. doi: 10.1186/s12864-017-4338-6
Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Res. 47, D140–D144. doi: 10.1093/nar/gky1051
Wang, X., Zeng, X., Ju, Y., Jiang, Y., Zhang, Z., and Chen, W. (2016). A classification method for microarrays based on diversity. Curr. Bioinform. 11, 590–597. doi: 10.2174/1574893609666140820224436
Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018a). Prediction of human protein subcellular localization using deep learning. J. Parallel Distrib. Comput. 117, 212–217. doi: 10.1016/j.jpdc.2017.08.009
Wei, L., Hu, J., Li, F., Song, J., Su, R., and Zou, Q. (2018b). Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. Brief. Bioinform. doi: 10.1093/bib/bby107. [Epub ahead of print].
Wei, L., Hu, J., Li, F., Song, J., Su, R., and Zou, Q. (2018b). Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. Brief. Bioinform. doi: 10.1093/bib/bby107. [Epub ahead of print].
Wang, X., Zeng, X., Ju, Y., Jiang, Y., Zhang, Z., and Chen, W. (2016). A classification method for microarrays based on diversity. Curr. Bioinform. 11, 590–597. doi: 10.2174/1574893609666140820224436
Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018a). Prediction of human protein subcellular localization using deep learning. J. Parallel Distrib. Comput. 117, 212–217. doi: 10.1016/j.jpdc.2017.08.009
Wei, L., Hu, J., Li, F., Song, J., Su, R., and Zou, Q. (2018b). Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. Brief. Bioinform. doi: 10.1093/bib/bby107. [Epub ahead of print].
Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and promising identification of human microRNAs by incorporating a high-quality negative set. IEEE/ACM Trans. Comput. Biol. Bioinform. 11, 192–201. doi: 10.1109/TCBB.2013.146
Wei, L., Su, R., Wang, B., Li, X., and Zou, Q. (2019). Integration of deep feature representations and handcrafted features to improve the prediction of N6-methyladenosine sites. Neurocomputing 324, 3–9. doi: 10.1016/j.neucom.2018.04.082
Wei, L., Tang, J., and Zou, Q. (2017a). SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. BMC Genomics 18:1. doi: 10.1186/s12864-017-4128-1
Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017b). Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. Artif. Intell. Med. 83, 67–74. doi: 10.1016/j.artmed.2017.03.001
Xu, L., Liang, G., Shi, S., and Liao, C. (2018a). SeqSVM: a sequence-based support vector machine method for identifying antioxidant proteins. Int. J. Mol. Sci. 19:1773. doi: 10.3390/ijms19061773
Xu, L., Liang, G., Wang, L., and Liao, C. (2018b). A novel hybrid sequence-based model for identifying anticancer peptides. Genes. 9:158. doi: 10.3390/genes9030158
Xue, W., Yang, F., Wang, P., Zheng, G., Chen, Y., Yao, X., et al. (2018). What contributes to serotonin-norepinephrine reuptake inhibitors’ dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. ACS Chem. Neurosci. 9, 1128–1140. doi: 10.1021/acschemneuro.7b00490
Zeng, X., Ding, N., Rodriguez-Patón, A., and Zou, Q. (2017a). Probability-based collaborative filtering model for predicting gene–disease associations. BMC Med. Genomics 10:76. doi: 10.1186/s12920-017-0313-y
Zeng, X., Lin, W., Guo, M., and Zou, Q. (2017b). A comprehensive overview and evaluation of circular RNA detection tools. PLoS Comput. Biol. 13:e1005420. doi: 10.1371/journal.pcbi.1005420
Zeng, X., Liu, L., Lü, L., and Zou, Q. (2018). Prediction of potential disease-associated microRNAs using structural perturbation method. Bioinformatics 34, 2425–2432. doi: 10.1093/bioinformatics/bty112
Zeng, X., Pan L., and Pérez-Jiménez M. J. (2014). Small universal simple spiking neural P systems with weights. Sci. China Inform. Sci. 57, 1–11. doi: 10.1007/s11432-013-4848-z
Zhang, J., and Liu, B. (2017). PSFM-DBT: identifying DNA-binding proteins by combing position specific frequency matrix and distance-bigram transformation. Int. J. Mol. Sci. 18:1856. doi: 10.3390/ijms18091856
Zhang, J., and Liu, B. (2018). Identification of DNA-binding proteins via a voting strategy. Curr. Proteomics. 15, 363–373. doi: 10.2174/1570164615666180718150317
Zhang, J., Zhang, Z., Chen, Z., and Deng, L. (2017b). Integrating multiple heterogeneous networks for novel LncRNA-disease association inference. IEEE/ACM Trans Comput Biol Bioinform. doi: 10.1109/TCBB.2017.2701379. [Epub ahead of print].
Zhang, X., Zou, Q., Rodriguez-Paton, A., and Zeng, X. (2018). Meta-path methods for prioritizing candidate disease miRNAs. IEEE/ACM Trans. Comput. Biol. Bioinform. 16, 283–291. doi: 10.1109/TCBB.2017.2776280
Zhang, Z., Zhang, J., Fan, C., Tang, Y., and Deng, L. (2017a). KATZLGO: large-scale prediction of LncRNA functions by using the KATZ measure based on multiple networks. IEEE/ACM Trans. Comput. Biol. Bioinform. doi: 10.1109/TCBB.2017.2704587. [Epub ahead of print].
Zhu, F., Li, X., Yang, S., and Chen, Y. (2018). Clinical success of drug targets prospectively predicted by in silico study. Trends Pharmacol. Sci. 39, 229–231. doi: 10.1016/j.tips.2017.12.002
Zou, Q., Li, J., Song, L., Zeng, X., and Wang, G. (2016). Similarity computation strategies in the microRNA-disease network: a survey. Brief. Funct. Genomics 15, 55–64. doi: 10.1093/bfgp/elv024

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Xu, Liang, Liao, and Chang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.