IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Identifying G-protein Coupled Receptors Using Mixed-Feature Extraction Methods and Machine Learning Methods

Chunyan Ao[1], Lin Gao[1], and Liang Yu[1*]
[1]School of Computer Science and Technology, Xidian University, Xi'an, China

*Corresponding author: Liang Yu (e-mail: lyu@xidian.edu.cn).

This work was supported by the National Key Research and Development Program of China (No. 2018YFC0910403), the National Natural Science Foundation of China (No. 61672406, 61532014, 61672407, and 61772395).

ABSTRACT G-protein-coupled receptors (GPCRs) are important protein molecules in the field of cell signaling and are widely found in various organisms. GPCRs play an important role in a variety of physiological processes and are important drug targets for a variety of diseases. Accurate prediction of GPCRs using machine learning is useful for drug design in a variety of related diseases. In this paper, we propose a method for identifying GPCRs based on mixed-feature vectors. We combine three individual features, such as 400D, N-gram and Parallel correlation pseudo amino acid composition (PC-PseAAC), using mixed-feature representation methods, which are evaluated by Random Forest, Naïve Bayes, and J48 for classification purposes. To measure the performance of this classifier, ten-fold cross-validation is used. Two dimensionality reduction methods—the max-relevance-max-distance (MRMD) and t-Distributed Stochastic Neighbor Embedding (t-SNE)—are applied to reduce the feature dimension. The 400D and PC-PseAAC feature extraction methods are combined, the random forest is used as the classifier, and the area under the curve (AUC) is up to 0.9413. Therefore, among these methods, the new feature vector obtained by combining the two features shows the best performance, and the mixed feature is better than the single feature.

INDEX TERMS G-protein-coupled receptors (GPCRs), J48, mixed-feature methods, naïve bayes, random forest.

I. INTRODUCTION

G protein-coupled receptors (GPCRs) are seven transmembrane proteins that perform reactions to transduce extracellular signals into cells. Because of their characteristic configuration of seven transmembrane alpha-helical counterclockwise beams [1], GPCRs are one of the largest membrane protein superfamilies, containing more than 800 genes in the human genome [2]. GPCRs are mainly divided into the following six classes [3]: rhodopsin-like receptors, secretin-like receptors, metabo-tropic glutamate receptors, fungal mating pheromone receptors, cyclic AMP (cAMP) receptors, and frizzled receptors.

GPCRs bind a variety of ligands, such as small molecule organic compounds, eicosanoids, peptides, and proteins [2]. GPCRs play an important role in many basic physicochemical processes, such as human vision, taste, smell, metabolism, neurotransmission, immune regulation, and cell growth [4-9]. These basic physicochemical processes are carried out by binding of GPCRs with ligand to activate a guanine-binding protein (G protein). At present, most drugs target GPCRs [10, 11], and thus GPCRs are involved in various diseases including depression, diabetes, cancer, and central nervous system diseases are widely targeted in drug development. In the future, accurate prediction of GPCRs is of great significance for drug design for various related diseases.

As the data available on GPCRs continues to increase, many methods for predicting GPCRs have been proposed. These methods mainly predict GPCRs from two aspects: one is based on statistical and machine learning algorithms, and the other is based on the extraction of different features. Statistics and machine learning algorithms mainly include Support Vector Machine (SVM) [8, 12-17], neural network [18, 19], the hidden Markov models (HMM) [4, 20, 21], Naïve Bayes [22, 23], k-nearest neighbor (KNN) [12, 24, 25], random forest [23, 26-32], etc. The main features used for GPCRs prediction are the following: pseudo amino acid composition (PseAAC) [33-36], split amino acid composition (SAAC) [37], fast Fourier transform (FFT) [37], N-gram [19, 38], amino acid composition (AAC) [12, 13, 39],

etc. In summary, based on different classifiers and different feature extraction methods to predict GPCRs, there are few useful mixed-feature representation methods to predict GPCRs.

In this paper, we propose a method of predicting GPCRs using machine learning combined with different feature extraction methods. Three single-feature extraction methods were used—400D, N-gram, and parallel correlation pseudo amino acid composition (PC-PseAAC). By combining the three single-feature methods by mixed-feature representation methods, four new hybrid feature vectors were obtained. We used three classifiers and ten-fold cross-validation to evaluate the classifier performance of each feature extraction strategy. The three classifiers are Random Forest (RF), Naïve Bayes (NB) and J48. We also used two dimensionality reduction methods—MDMR and t-SNE—to reduce the dimensions, and then used different classifiers to classify. The results are shown in the experimental section. The flow chart for predicting GPCRs is shown in Figure 1.

FIGURE 1. The flow-chart for predicting GPCRs. PC-PseAAC: parallel correlation pseudo amino acid composition; MRMD: max-relevance-max-distance; t-SNE: t-distributed stochastic neighbor embedding; RF: random forest.

II. METHODS

A. DATASET

The dataset that we mainly used for training and testing the classification approach was the same as that used in Liao et al. [27], and these data were obtained by CD-Hit to remove the sequence homology. The dataset is made up of 12881 protein sequences that can be classified into two parts: 2495 GPCRs (positive samples) and 10386 non-GPCRs (negative samples). To overcome the dataset imbalance, we randomly divided the negative samples into four groups, extracted 2495 sequences from these four parts, and averaged the results of the four experiments using the four negative experiments.

B. FEATURE EXTRACTION METHODS

In this manuscript, we extracted three single features—400D, parallel correlation pseudo amino acid composition (PC-PseAAC), and N-gram.

1) 400D

The 400D feature is based on k-skip-n-grams [40], which is a sequence-based feature [41]. In addition to the n contiguous residues considered by the k-skip-n-grams model, the model also considers the n residues with distances from 1 to k in amino acid sequence. In the k-skip-n-gram model, we take n as 2, define $S = \{A, C, D, E, \dots, Y\}$, and define f_i as a binary permutation and combination of S elements, such as: $f_1 = AA$, $f_{400} = YY$.The feature representation method calculates the feature vector set as follows:

$$T_{Skip-Gram} = \{U_{c=1}^{k} Skip(DT = c)\} \quad (1)$$

where $Skip(\mathrm{DT} = \mathrm{C}) = \{A_i A_{i+c+1} | 1 \le i \le l - c, 1 \le c \le k\}$, and l represents the length of the amino acid sequence.

$$V_i = \frac{B(f_i)}{N(T_{Skip-Gram})} \quad (2)$$

$$\mathrm{FV} = (V_1, V_2, \dots, V_i, \dots, V_{400}) \quad (3)$$

where $\mathrm{B}(f_i)$ represents number of times the f_i sequence appears in $T_{Skip-Gram}$ set, and $\mathrm{N}(T_{Skip-Gram})$ is the sum number of all elements in $T_{Skip-Gram}$ set.

In this manuscript, we took the value of n to 2 and obtained the 400D feature. The protein sequence consists of 20 amino acids, so A_1A_2 is represented as the combination of two consecutive amino acids. $f_{A_1A_2}$ represents the frequency of the combination of A_1A_2. The 400D feature is represented by 400 combined frequencies.

2) N-GRAM

N-gram [42] is a commonly used large vocabulary continuous recognition language probability model. N-gram model is widely used in bioinformatics research, such as protein identification [19, 43-46], RNA structure modeling [47], genome sequence analysis [48, 49], etc. N-gram model is often used to estimate the probability of the occurrence of a given sentence in the corpus, that is, an N-gram is a word sequence of length N. When N=1, it is called a Unigram model, that is, a unary model, also called a context-independent model; when N=2, it is called a bigram model; when N=3, it is called a trigram model or a ternary model. The probability of the whole sentence is the product of the probability of occurrence of each word.

Since the dimension of the feature space grows exponentially with N, in order to reduce the feature space and overfitting phenomenon, high accuracy is obtained, and the maximum value of N is set to 3. In this manuscript, we took an N value of 2 and accumulated the number of features generated by the model. The total number of input features in 1, 2 is equal to $20 + 20^2(420)$, and we obtained a 420-dimensional feature.

3) PARALLEL CORRELATION PSEUDO AMINO ACID COMPOSITION

Parallel correlation pseudo amino acid composition (PC-PseAAC) is a commonly used protein analysis method, and its use is based on the method of integrating continuous local sequence order information and global sequence order information into protein sequence feature vectors [50, 51]. Pseudo amino acids are widely used for protein prediction [52, 53]. According to the pseudo-amino acid composition theory [51], λ sequence correlation functions that reflect the physicochemical properties of amino acids are introduced, and a protein sequence (or peptide) is encoded into a vector of$20 + \lambda$ dimensions:

$$\mathrm{Y} = (\mathrm{Y}_1, \mathrm{Y}_2, \mathrm{Y}_3, \dots, \mathrm{Y}_{20}, \mathrm{Y}_{21}, \dots, \mathrm{Y}_{20+\lambda}) \quad (4)$$

where the first 20 dimensions of Y represent the frequency of the amino acid. In this study, we selected three features, set λ to 2, and obtained 22-dimensional feature.

C. CLASSIFICATION

In this study, we selected the following three classifier models for classification: random forest (RF), naïve Bayes (NB), and J48. The three classifiers were implemented in the data mining tool Weka [54], which is an ensemble package of multiple machine learning algorithms and is based on the Java environment.

1) RANDOM FOREST

Random forest (RF) is a powerful algorithm designed and proposed by Brieman et al. [55] and is a collection of tree predictors. RF has been widely used in many fields of bioinformatics [56-73]. RF is implemented by constructing a large number of decision trees during training and outputting the class pattern of individual trees [55]. The RF algorithm behaves similar to the ensemble algorithm [74, 75]; it consists of decision trees, and each is grown according to a subset of features selected by the stochastic feature selection technique. The feature number of each tree is determined by a number of factors, including generalization errors, classifier strength, and dependence. The prediction result of the RF algorithm is a set of results of combining all training trees with a majority voting strategy.

2) NAÏVE BAYES

Naïve Bayes (NB) [76] is a classifier based on conditional probability and is usually used to calculate conditional probabilities. The NB algorithm is one of the most commonly used algorithms because it is simple to implement and has high classification performance [77, 78]. The Naïve Bayes classifier [79] is a simple probability classifier that assumes conditional independence between variables, i.e., the presence (or absence) of a particular type of variable is independent of the presence (or absence) of any other variable. Only a small amount of training data is needed to estimate the parameters required for classification.

In the sample space F, the representation of the specimen is i, $F_i = (f_{ij}, \dots, f_{in})$, where f_{ij} represents j features in i samples, and the calculation formula of NB based on Bayes' theorem formula is as follows [79]:

$$P(y|F_i) = \frac{P(F_i|y)P(y)}{P(F_i)} \quad (5)$$

If $f_{ij}, \dots, f_{in}$ are independent of each other, we get the following formula:

$$P(F_i|y) = \prod_{j=1}^{n} P(f_{ij}|y) \quad (6)$$

$$P(F_i) = \prod_{j=1}^{n} P(f_{ij}) \quad (7)$$

3) J48

The J48 is a decision tree classifier generated by the C4.5 algorithm developed by Quinlan [80]. The decision tree is a classification algorithm based on univariate logic, and it can sort the training examples using eigenvalues based on the divide and conquer strategy. The decision tree J48 [80, 81] is tree-like graph in which the nodes in the graph test certain conditions on a set of features, and the branches divide the decision into the leaf nodes. The leaves represent the lowest level in the graph and determine the category labels.

The C4.5 algorithm [81, 82] uses the gain ratio impurity method to evaluate the segmentation properties. At each node of the tree, C4.5 selects data that most effectively splits its sample set into one or another subset rich in categories. Its standard is the normalized information gain, which is generated by the choice of attributes used to split the data. The attribute with the highest normalized information gain is selected for decision making. The J48 decision tree algorithm is used for four different M parameter values [83], which defines the minimum number of examples (M = 2, 4, 6, 8) in each node of the tree. The high value of M corresponds to the regular model and simple model.

D. FEATURE SELECTION

The max-relevance-max-distance (MRMD) is a dimension reduction method designed by Zou et al. [84]. This algorithm is automatically stopped when the maximum ACC is obtained, and the feature set after dimension reduction is obtained. In this paper, four mixed-feature vectors were obtained by the mixed-feature representation methods 400D+N-gram, 400D+ PC-PseAAC, N-gram+ PC-PseAAC, and 400D+ N-gram+ PC-PseAAC. Because the mixed-feature vector may contain redundant vectors, we used MRMD to reduce the dimension of the feature extraction algorithm, reduce the redundant vector, and improve the classification effect.

The t-Distributed Stochastic Neighbor Embedding (t-SNE) [85] is a dimensionality reduction technology suitable for visualization of high-dimensional datasets, which uses heavy-tailed distribution in low-dimensional space to alleviate the congestion and optimization problems of SNE. In this manuscript, we obtained four mixed-feature vectors by three single features through the mixed feature representation methods. These feature vectors are visualized in a two-dimensional feature space using t-distributed random neighbor embedding (t-SNE).

III. EXPERIMENT

A. MEASUREMENT

The performance of the predictive classifier is usually verified by three cross-validation methods—the independent dataset test, the jack-knife test, and the k-fold cross-validation test [86-107]. In this study, we used ten-fold cross-validation to evaluate the performance of the classifier. Ten-fold cross-validation divides the dataset into ten parts, and takes 9 of them as training data and 1 as test data for experimentation.

Parameters such as the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity (Sn), specificity (Sp), accuracy (Acc) and precision are commonly used for performance evaluation and are computed as follows:

$$\mathrm{Sn} = \frac{TP}{TP+FN} \quad (8)$$

$$\mathrm{Sp} = \frac{TN}{TN+FP} \quad (9)$$

$$\mathrm{Acc} = \frac{TP+TN}{TP+FP+TN+FN} \quad (10)$$

$$\mathrm{Precision} = \frac{TP}{TP+FP} \quad (11)$$

where TN, TP, FN, and FP represent the number of true

negative, true positive, false negative, and false positive values, respectively.

B. PERFORMANCE OF SINGLE-FEATURE EXTRACTION MENTHOD

In this manuscript, we employed the single-feature extraction method to classify the GPCRs sequences. The classification results using different classifiers are shown in Table 1.

TABLE 1. GPCRs identification results using single-feature representation methods.

According to the experimental results in Table 1, 400D has the best accuracy at 0.8644, followed by PC-PseAAC and N-gram (N=2). The performance of the random forest is best when using the 400D feature extraction method, and the performance parameters for AUC, Acc, precision, Sn, and Sp values are 0.9408, 0.8644, 0.8653, 0.8378, and 0.8910, respectively. The Sp value is not the highest, but the other values are the best. The performance of J48 is the worst, and the AUC value is 0.7620, Acc is 0.7935, precision is 0.7935, Sn is 0.7913, and Sp is 0.7958. Under the three feature extraction methods, the NB performance is the best under the PC-PseAAC feature extraction method, the AUC is 0.8460. The experimental results show that the RF has better classification effects than J48 and NB (Figure 2).

FIGURE 2. Comparison of AUC and Precision based on different feature extraction methods and different classifiers.

From the experimental results and comparison of the AUC and precision values among the three feature extraction methods and three classifiers, the random forest classifier has the best classification effect among different feature methods, followed by NB and J48. To more clearly and intuitively represent the comparison of the three feature extraction methods and the three different classifiers, we use bold to show the best method and classifier in Table 1.

C. PERFORMANCE OF MIXED-FEATURE EXTRACTION METHODS

In this section, we combined the three feature extraction methods in various ways, and obtained four new feature vectors. The obtained feature vectors are 400D+N-gram, 400D+PC-PseAAC, PC-PseAAC+N-gram, and 400D+N-gram+PC-PseAAC. These features are classified by different classifiers, and the obtained classification results are shown in Table 2.

TABLE 2. Classification result of the mixed-feature methods.

The results of 400D combined with other features are presented in Table 2. 400D has an AUC value of up to 0.9413 and a minimum of 0.7668, compared to the 400D individual feature classification results in Table 1. The highest Acc value is 0.8647, and the lowest value is 0.7911. The values are higher than when using the 400D feature method alone, while the precision value is reduced. The highest value of precision is 0.8662, and the lowest value is 0.7910.

The results of the PC-PseAAC feature combined with other features are shown in Table 2. The new features of PC-PseAAC combined with N-gram for all classifiers have worse AUC and precision values than using three individual features. The performance of the three classifiers is reduced. When 400D was combined with the PC-PseAAC, the J48 classifier has the worst performance, and all parameter values are lower than when using the PC-PseAAC feature method.

The results for the N-gram feature combined with the other two features are shown in Table 2. The experimental results show that the performance of the three classifiers is better than that of the N-gram feature alone, and all the parameters improved. The AUC value ranged from 0.9408 to 0.7683. The Acc value is up to 0.8647 and as low as 0.7795. The highest value of precision is 0.8658, and the lowest value is 0.7798.

Finally, we combined the three individual features, and the classification results are shown in Table 2. The RF classification has the best classification effect, and the obtained AUC value is up to 0.9412 compared with the three features alone. The performance of the NB classifier degraded. The classification effect of the J48 classifier is better than the classification effect of the PC-PseAAC individual feature.

In summary, the RF classifier has the best classification effect, and J48 has the worst. We use bold font to show the best parameters for each feature classification in Table 2. When 400D was combined with PC-PseAAC, RF is used as the classifier to predict GPCRs with the best results, and the AUC value is 0.9413.

D. PERFORMANCE OF EMPLOY MRMD AND T-SNE TO REDUCE THE DIMENSION

In Section 3.3, we combined the three features through mixed-feature representation methods to obtain four new feature vector sets. We used the MRMD method to reduce the dimension of the new feature vector. The data after dimension reduction was classified by various classifiers. The classification results are shown in Table 3 and Figure 3.

TABLE 3. Classification result of the reduction the features.

According to the experimental results, although the classification result of the RF classifier was lower than that of the mixed feature after the dimension reduction by MRMD, RF is the best classifier, and J48 is the worst classifier. From the experimental results, the highest AUC value is 0.9410 and the lowest value is 0.7665, and the highest value of Acc is 0.8639 and the lowest value is 0.7830. The highest precision value is 0.8653, and the lowest value is 0.7830. We present the results of the best classifier for each feature in bold font in Table 3.

Figure 3 shows the results intuitively. When using dimensionality reduction data for classification, the

classification effect of the NB classifier improves, and the AUC values of the four features increases.

FIGURE 3. Comparison of classification results of mixed features and features after dimensionality reduction. Blue indicates the classification results after dimensionality reduction, and orange indicates the classification results of mixed features.

Next, we obtained datasets of four mixed vectors by combining three features using t-distributed random neighbor embedding (t-SNE) for dimensionality reduction and visualizing them in 2D feature space. The results are shown in Figure 4. Figure 4A represents a combination of 400D features and N-gram features, which were obtained by dimensionality reduction visualization, and the classifier was logistic regression. Figure 4B shows the fusion of the 400D feature and the PC-PseAAC feature. The classifier was also a logistic regression. After the dimension reduction, a visualization was obtained. The combination of the PC-PseAAC and N-gram features with the NB classifier after dimensionality reduction is shown in Figure 4C. The last Figure 4D is a visualization of the three features, which was obtained by t-SNE dimensionality reduction with the logistic regression classifier. In summary, the distribution of the 2D features space in the three features combined with t-SNE dimensionality reduction is better than that of the other three combinations.

FIGURE 4. The visualization of four mixed vectors. The red points are negative examples, and the blue points are positive examples.

E. COMPARISON WITH OTHER METHOD

In this paper, the experimental data are derived from the research of Liao et al. [27]. For the purpose of demonstrating the accuracy of our method, the proposed method is compared with literature method [27].

According to Table 4, the method proposed can obtain high AUC value based on three different feature extract methods. When RF is used as the classifier to predict GPCRs, the AUC value obtained by our method range from 0.9408 to 0.9370. The average AUC obtained by literature method was 0.9282. The results indicate that our method is superior to the literature method.

Table 4. Comparison of the AUC of our method and literature method.

IV. CONCLUSION

To date, studies have shown that GPCRs are found only in eukaryotes and are involved in many cellular signal transduction processes. Therefore, many drugs target GPCRs. In this study, we used multiple classifiers to combine different features to classify GPCRs. A hybrid combination of three feature extraction techniques was used, and then MRMD and t-SNE were used to reduce the dimension of the mixed-feature vector. From the classification results extracted from a single feature, it is shown that the combination of the 400D feature and the random forest obtained the highest AUC of 0.9408, and the classification effect is the best. Of the four new feature vectors obtained by the mixed-features methods, the combination of 400D and PC-PseAAC after MRMD dimension reduction with the random forest classifier has the highest AUC value of 0.9410. The visualization is obtained by t-SNE dimensionality reduction, and the high-dimensional data are mapped to the two-dimensional space. The experimental results show that the best classifier for predicting GPCRs is RF. Additionally, both dimensionality reduction hybrid feature and nondimension reduction hybrid feature exhibited better performance than single features. This finding indicates that the mixed combination of different feature extraction methods improves the overall performance of GPCRs prediction. In the future, we hope to propose more advanced classification algorithms and feature selection methods to improve performance and establish an online service website.

REFERENCES

[1] P. K. Papasaikas, P. G. Bagos, Z. I. Litou, and S. J. Hamodrakas, "A Novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden Markov models," *SAR and QSAR in Environmental Research,* vol. 14, no. 5-6, pp. 413-420, 2003/10/01 2003.

[2] M. C. Lagerström and H. B. Schiöth, "Structural diversity of G protein-coupled receptors and significance for drug discovery," *Nature reviews Drug discovery,* vol. 7, no. 4, pp. 339-357, 2008.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology,* vol. 215, no. 3, pp. 403-410, 1990.

[4] H.-S. Eo, J. P. Choi, S.-J. Noh, C.-G. Hur, and W. Kim, "A combined approach for the classification of G protein-coupled receptors and its application to detect GPCR splice variants," *Computational biology and chemistry,* vol. 31, no. 4, pp. 246-256, 2007.

[5] J. M. Baldwin, "Structure and function of receptors coupled to G proteins," *Current opinion in cell biology,* vol. 6, no. 2, pp. 180-190, 1994.

[6] R. Lefkowitz, "The superfamily of heptahelical receptors," *Nature cell biology,* vol. 2, no. 7, p. E133, 2000.

[7] K.-C. Chou and D. W. Elrod, "Bioinformatical analysis of G-protein-coupled receptors," *Journal of proteome research,* vol. 1, no. 5, pp. 429-433, 2002.

[8] R. Karchin, K. Karplus, and D. Haussler, "Classifying G-protein coupled receptors with support vector machines," *Bioinformatics,* vol. 18, no. 1, pp. 147-159, 2002.

[9] T. E. Hébert and M. Bouvier, "Structural and functional aspects of G protein-coupled receptor oligomerization," *Biochemistry and cell biology,* vol. 76, no. 1, pp. 1-11, 1998.

[10] K. H. Lundstrom and M. L. Chiu, *G protein-coupled receptors in drug discovery*. CRC Press, 2005.

[11] M. Bhasin and G. Raghava, "GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors," *Nucleic acids research,* vol. 32, no. suppl_2, pp. W383-W389, 2004.

[12] M. Naveed and A. U. Khan, "GPCR-MPredictor: multi-level prediction of G protein-coupled receptors using genetic ensemble," *Amino Acids,* vol. 42, no. 5, pp. 1809-1823, 2012.

[13] G. Nie, Y. Li, F. Wang, S. Wang, and X. Hu, "A novel fractal approach for predicting G-protein–coupled receptors and their subfamilies with support vector machines," *Bio-medical materials engineering,* vol. 26, no. s1, pp. S1829-S1836, 2015.

[14] Y. Shen, J. Tang, and F. Guo, "Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC," *Journal of Theoretical Biology,* vol. 462, pp. 230-239, 2019/02/07/ 2019.

[15] J. He, T. Fang, Z. Zhang, B. Huang, X. Zhu, and Y. Xiong, "PseUI: Pseudouridine sites identification based on RNA

sequence information," *BMC Bioinformatics,* vol. 19, no. 1, p. 306, Aug 29 2018.

[16] Y. Xiong, J. Liu, W. Zhang, and T. Zeng, "Prediction of heme binding residues from protein sequences with integrative sequence profiles," *Proteome Sci,* vol. 10 Suppl 1, p. S20, Jun 21 2012.

[17] Y. Zhao, F. Wang, and L. Juan, "MicroRNA Promoter Identification in Arabidopsis Using Multiple Histone Markers," *Biomed Res Int,* vol. 2015, p. 861402, 2015.

[18] A. Khan, "Identifying GPCRs and their types with Chou's pseudo amino acid composition: An approach from multi-scale energy representation and position specific scoring matrix," *Protein Peptide Letters,* vol. 19, no. 8, pp. 890-903, 2012.

[19] M. Li, C. Ling, Q. Xu, and J. Gao, "Classification of G-protein coupled receptors based on a rich generation of convolutional neural network, N-gram transformation and multiple sequence alignments," *Amino acids,* vol. 50, no. 2, pp. 255-266, 2018.

[20] B. Qian, O. S. Soyer, R. R. Neubig, and R. A. Goldstein, "Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs," *FEBS letters,* vol. 554, no. 1-2, pp. 95-99, 2003.

[21] P. L. Martelli, P. Fariselli, L. Malaguti, and R. Casadio, "Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks," *Protein Engineering,* vol. 15, no. 12, pp. 951-953, 2002.

[22] J. Cao, R. Panetta, S. Yue, A. Steyaert, M. Young-Bellido, and S. Ahmad, "A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins," *Bioinformatics,* vol. 19, no. 2, pp. 234-240, 2003.

[23] J. cheol Jeong, X. Lin, and X.-w. Chen, "On position-specific scoring matrix for protein function prediction," *IEEE/ACM transactions on computational biology bioinformatics,* vol. 8, no. 2, pp. 308-315, 2010.

[24] J. Dongardive and S. Abraham, "Protein sequence classification based on n-gram and k-nearest neighbor algorithm," in *Computational Intelligence in Data Mining—Volume 2*: Springer, 2016, pp. 163-171.

[25] X. Xiao, J.-L. Min, P. Wang, and K.-C. Chou, "iGPCR-Drug: A web server for predicting interaction between GPCRs and drugs in cellular networking," *PloS one,* vol. 8, no. 8, p. e72234, 2013.

[26] Z.-L. Peng, J.-Y. Yang, and X. Chen, "An improved classification of G-protein-coupled receptors using sequence-derived features," *BMC bioinformatics,* vol. 11, no. 1, p. 420, 2010.

[27] Z. Liao, Y. Ju, and Q. Zou, "Prediction of G protein-coupled receptors with SVM-prot features and random forest," *Scientifica,* vol. 2016, 2016.

[28] Y. Ding, J. Tang, and F. Guo, "Identification of drug-target interactions via multiple information integration," *Information Sciences,* vol. 418-419, pp. 546-560, 2017/12/01/ 2017.

[29] L. Cheng *et al.*, "InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk," *BMC Genomics,* vol. 19, no. Suppl 1, p. 919, Jan 19 2018.

[30] L. Cheng, Y. Hu, J. Sun, M. Zhou, and Q. Jiang, "DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function," *Bioinformatics,* vol. 34, no. 11, pp. 1953-1956, Jun 1 2018.

[31] G. Wang, Y. Wang, M. Teng, D. Zhang, L. Li, and Y. Liu, "Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon gamma-stimulated HeLa cells," *PLoS One,* vol. 5, no. 7, p. e11794, Jul 26 2010.

[32] L. Yu, S. Yao, L. Gao, and Y. Zha, "Conserved Disease Modules Extracted From Multilayer Heterogeneous Disease and Gene Networks for Understanding Disease Mechanisms and Predicting Disease Treatments," *Frontiers in Genetics,* vol. 9, Jan 18 2019, Art. no. 745.

[33] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics,* vol. 21, no. 1, pp. 10-19, 2004.

[34] A. Khan, "G-protein-coupled receptor prediction using pseudo-amino-acid composition and multiscale energy representation of different physiochemical properties," *Analytical biochemistry,* vol. 412, no. 2, pp. 173-182, 2011.

[35] Z.-u. Rehman, M. T. Mirza, A. Khan, and H. Xhaard, "Predicting G-Protein-Coupled Receptors families using different physiochemical properties and pseudo amino acid composition," in *Methods in enzymology*, vol. 522: Elsevier, 2013, pp. 61-79.

[36] A. K. Tiwari, "Prediction of G-protein coupled receptors and their subfamilies by incorporating various sequence features into Chou's general PseAAC," *Computer methods programs in biomedicine,* vol. 134, pp. 197-213, 2016.

[37] A. Khan, "Prediction of GPCRs with pseudo amino acid composition: employing composite features and grey incidence degree based classification," *Protein peptide letters,* vol. 18, no. 9, pp. 872-878, 2011.

[38] M. Li, C. Ling, and J. Gao, "An efficient CNN-based classification on G-protein Coupled Receptors using TF-IDF and N-gram," in *2017 IEEE Symposium on Computers and Communications (ISCC)*, 2017, pp. 924-931: IEEE.

[39] B. Liu, X. Gao, and H. Zhang, "BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches," *Nucleic Acids Research,* vol. 47, no. 20, pp. e127-e127, 2019.

[40] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks, "A closer look at skip-gram modelling," in *LREC*, 2006, pp. 1222-1225.

[41] L. Wei, J. Tang, and Q. Zou, "SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides," *BMC genomics,* vol. 18, no. 7, p. 742, 2017.

[42] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational linguistics,* vol. 18, no. 4, pp. 467-479, 1992.

[43] Q. Dong, K. Wang, and X. Liu, "Identifying the missing proteins in human proteome by biological language model," *BMC systems biology,* vol. 10, no. 4, p. 113, 2016.

[44] L. Xu, G. Liang, L. Wang, and C. Liao, "A Novel Hybrid Sequence-Based Model for Identifying Anticancer Peptides," *Genes,* vol. 9, no. 3, p. 158, 2018.

[45] L. Jiang, Y. Ding, J. Tang, and F. Guo, "MDA-SKF: Similarity Kernel Fusion for Accurately Discovering miRNA-Disease Association," *Frontiers in Genetics,* vol. 9, p. 618, 12/01 2018.

[46] L. Xu, G. Liang, C. Liao, G.-D. Chen, and C.-C. Chang, "k-Skip-n-Gram-RF: A Random Forest Based Method for Alzheimer's Disease Protein Identification," *Frontiers in Genetics,* vol. 10, Feb 12 2019, Art. no. 33.

[47] I. Salvador and J.-M. Benedi, "RNA modeling by combining stochastic context-free grammars and n-gram models," *International Journal of Pattern Recognition Artificial Intelligence in Medicine,* vol. 16, no. 03, pp. 309-315, 2002.

[48] A. Tomović, P. Janičić, and V. Kešelj, "n-Gram-based classification and unsupervised hierarchical clustering of genome sequences," *Computer methods programs in biomedicine,* vol. 81, no. 2, pp. 137-153, 2006.

[49] L. Xu, G. Liang, C. Liao, G.-D. Chen, and C.-C. Chang, "An Efficient Classifier for Alzheimer's Disease Genes Identification," *Molecules,* vol. 23, no. 12, p. 3140, 2018.

[50] D. Georgiou, T. Karakasidis, and A. Megaritis, "A short survey on genetic sequences, Chou's pseudo amino acid composition and its combination with fuzzy set theory," *Open Bioinform. J,* vol. 7, pp. 41-48, 2013.

[51] K. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Structure, Function, Bioinformatics,* vol. 43, no. 3, pp. 246-255, 2001.

[52] A. K. Tiwari, "Prediction of G-protein coupled receptors and their subfamilies by incorporating various sequence features into Chou's general PseAAC," *Computer methods and programs in biomedicine,* vol. 134, pp. 197-213, 2016.

[53] G. Pan, L. Jiang, J. Tang, and F. Guo, "A novel computational method for detecting DNA methylation sites with DNA sequence information and physicochemical properties,"

International journal of molecular sciences, vol. 19, no. 2, p. 511, 2018.

[54] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, no. 1, pp. 10-18, 2009.

[55] L. Breiman, "Random Forests," *Machine Learning,* journal article vol. 45, no. 1, pp. 5-32, October 01 2001.

[56] Y. Li *et al.*, "Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features," *Scientific Reports,* Article vol. 4, p. 5765, 07/21/online 2014.

[57] B. Liu, F. Yang, D.-S. Huang, and K.-C. Chou, "iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC," *Bioinformatics,* vol. 34, no. 1, pp. 33-40, 2017.

[58] J. Jia, Z. Liu, X. Xiao, B. Liu, and K.-C. Chou, "iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset," *Analytical biochemistry,* vol. 497, pp. 48-56, 2016.

[59] L. Wei, J. Tang, and Q. Zou, "Local-DPP: An Improved DNA-binding Protein Prediction Method by Exploring Local Evolutionary Information," *Information Sciences,* vol. 384, pp. 135-144, 2017.

[60] L. Wei, P. Xing, R. Su, G. Shi, Z. Ma, and Q. Zou, "CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency," *Journal of Proteome Research,* vol. 16, no. 5, pp. 2044-2053, 2017.

[61] L. Wei, P. Xing, J. Tang, and Q. Zou, "PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only," *IEEE Transactions on NanoBioscience,* vol. 16, no. 4, pp. 240-247, 2017.

[62] L. Cheng *et al.*, "MetSigDis: a manually curated resource for the metabolic signatures of diseases," *Brief Bioinform,* vol. 20, no. 1, pp. 203-209, Jan 18 2019.

[63] L. Cheng, C. Qi, H. Zhuang, T. Fu, and X. Zhang, "gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions," *Nucleic Acids Res,* Oct 4 2019.

[64] Y. Chu *et al.*, "DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features," *Brief Bioinform,* Dec 23 2019.

[65] Y. Xiong, Q. Wang, J. Yang, X. Zhu, and D. Q. Wei, "PredT4SE-Stack: Prediction of Bacterial Type IV Secreted Effectors From Protein Sequences Using a Stacked Ensemble Method," *Front Microbiol,* vol. 9, p. 2571, 2018.

[66] C. Jia, Y. Zuo, and Q. Zou, "O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique," *Bioinformatics,* vol. 34, no. 12, pp. 2029-2036, Jun 15 2018.

[67] X. Zeng, Y. Zhong, W. Lin, and Q. Zou, "Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods," *Briefings in Bioinformatics,* vol. DOI: 10.1093/bib/bbz080, 2019.

[68] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, "Prediction and Validation of Disease Genes Using HeteSim Scores," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 14, no. 3, pp. 687-695, MAY-JUN 2017 2017.

[69] X. Zeng, X. Zhang, and Q. Zou, "Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks," *Briefings in Bioinformatics,* vol. 17, no. 2, pp. 193-203, MAR 2016 2016.

[70] L. Cheng *et al.*, "LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse," *Nucleic Acids Res,* vol. 47, no. D1, pp. D140-D144, Jan 8 2019.

[71] L. Yu, J. Zhao, and L. Gao, "Predicting Potential Drugs for Breast Cancer based on miRNA and Tissue Specificity," *International Journal of Biological Sciences,* vol. 14, no. 8, pp. 971-980, 2018 2018.

[72] L. Yu *et al.*, "Prediction of Novel Drugs for Hepatocellular Carcinoma Based on Multi-Source Random Walk," *Ieee-Acm Transactions on Computational Biology and Bioinformatics,* vol. 14, no. 4, pp. 966-977, Jul-Aug 2017.

[73] Q. Jiang, G. Wang, T. Zhang, and Y. Wang, "Predicting human microRNA-disease associations based on support vector machine," in *2010 IEEE International Conference On Bioinformatics and Biomedicine (BIBM)*, 2010, pp. 467-472: IEEE.

[74] P. Zhu, Q. Hu, Q. Hu, C. Zhang, and Z. Feng, "Multi-view label embedding," *Pattern Recognition,* vol. 84, pp. 126-135, 2018/12/01/ 2018.

[75] P. Zhu, Q. Hu, Y. Han, C. Zhang, and Y. Du, "Combining neighborhood separable subspaces for classification via sparsity regularized optimization," *Information Sciences,* vol. 370-371, pp. 270-287, 2016/11/20/ 2016.

[76] L. Deng and Z. Chen, "An Integrated Framework for Functional Annotation of Protein Structural Domains," *IEEE/ACM Transactions on Computational Biology Bioinformatics,* vol. 12, no. 4, pp. 902-913.

[77] V. Narayanan, I. Arora, and A. Bhatia, "Fast and accurate sentiment classification using an enhanced Naive Bayes model," in *International Conference on Intelligent Data Engineering and Automated Learning*, 2013, pp. 194-201: Springer.

[78] M. Sahami, "Learning Limited Dependence Bayesian Classifiers," in *KDD*, 1996, vol. 96, no. 1, pp. 335-338.

[79] Z. Chen *et al.*, "Feature selection with redundancy-complementariness dispersion," *Knowledge-Based Systems,* vol. 89, pp. 203-217, 2015/11/01/ 2015.

[80] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," *Machine Learning,* vol. 16, no. 3, pp. 235-240, 1994.

[81] Quinlan and J. R., "Induction of decision trees," *Machine Learning,* vol. 1, no. 1, pp. 81-106.

[82] S. P. Shi, J. D. Qiu, X. Y. Sun, S. B. Suo, S. Y. Huang, and R. P. Liang, "PMeS: prediction of methylation sites based on enhanced feature encoding scheme," *PLoS One,* vol. 7, no. 6, p. e38772, 2012.

[83] J. R. Quinlan, "Improved Use of Continuous Attributes in C4.5," *Journal of Artificial Intelligence Research,* vol. 4, no. 1, pp. 77-90, 1996.

[84] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing,* vol. 173, pp. 346-354, 2016.

[85] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research,* vol. 9, no. Nov, pp. 2579-2605, 2008.

[86] F. Ali and M. Hayat, "Classification of membrane protein types using Voting Feature Interval in combination with Chou's Pseudo Amino Acid Composition," *Journal of Theoretical Biology,* vol. 384, pp. 78-83, 2015/11/07/ 2015.

[87] P. Chaudhary, A. N. Naganathan, and M. M. Gromiha, "Prediction of change in protein unfolding rates upon point mutations in two state proteins," *Biochimica Et Biophysica Acta-Proteins and Proteomics,* vol. 1864, no. 9, pp. 1104-1109, Sep 2016.

[88] C. Shen, L. Jiang, Y. Ding, J. Tang, and F. Guo, "LPI-KTASLP: Prediction of lncRNA-Protein Interaction by Semi-Supervised Link Learning with Multivariate Information," *IEEE Access,* vol. 7, pp. 13486 - 13496, 2019.

[89] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via multiple information integration with centered kernel alignment," *Neurocomputing,* vol. 325, pp. 211-224, 2019/01/24/ 2019.

[90] R. Su, H. Wu, B. Xu, X. Liu, and L. Wei, "Developing a Multi-Dose Computational Model for Drug-induced Hepatotoxicity Prediction based on Toxicogenomics Data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 16, no. 4, pp. 1231-1239, JUL-AUG 2019 2019.

[91] X. Zeng, W. Lin, M. Guo, and Q. Zou, "A comprehensive overview and evaluation of circular RNA detection tools," *Plos Computational Biology,* vol. 13, no. 6, p. e1005420, 2017.

[92] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, "Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier," *Artificial Intelligence in Medicine,* vol. 83, pp. 67-74, 2017.

[93] L. Cheng *et al.*, "Computational Methods for Identifying Similar Diseases," *Mol Ther Nucleic Acids,* vol. 18, pp. 590-604, Sep 28 2019.

[94] L. Cheng, H. Zhuang, S. Yang, H. Jiang, S. Wang, and J. Zhang, "Exposing the Causal Effect of C-Reactive Protein on the Risk of Type 2 Diabetes Mellitus: A Mendelian Randomization Study," *Front Genet,* vol. 9, p. 657, 2018.

[95] L. Cheng and Y. Hu, "Human Disease System Biology," *Curr Gene Ther,* Nov 1 2018.

[96] X. Zhu, J. He, S. Zhao, W. Tao, Y. Xiong, and S. Bi, "A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae," *Brief Funct Genomics,* vol. 18, no. 6, pp. 367-376, Nov 19 2019.

[97] X. Shan *et al.*, "Prediction of CYP450 Enzyme-Substrate Selectivity Based on the Network-Based Label Space Division Method," *J Chem Inf Model,* vol. 59, no. 11, pp. 4577-4586, Nov 25 2019.

[98] M. Zhang *et al.*, "MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters," *Bioinformatics,* vol. 35, no. 17, pp. 2957-2965, Sep 1 2019.

[99] L. Dou, X. Li, H. Ding, L. Xu, and H. Xiang, "Is There Any Sequence Feature in the RNA Pseudouridine Modification Prediction Problem?," *Molecular Therapy - Nucleic Acids,* vol. 19, pp. 293-303, 2020/03/06/ 2020.

[100] X. Zhang, Q. Zou, A. Rodriguez-Paton, X. J. I. A. T. o. C. B. Zeng, and Bioinformatics, "Meta-path methods for prioritizing candidate disease miRNAs," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 16, no. 1, pp. 283-291, 2019.

[101] X. Lin, Z. Quan, Z.-J. Wang, H. Huang, and X. J. B. i. B. Zeng, "A novel molecular representation with BiGRU neural networks for learning atom," *Briefings in Bioinformatics,* vol. Doi: 10.1093/bib/bbz125, 2019.

[102] X. Zeng, N. Ding, A. Rodríguezpatón, and Z. J. B. M. G. Quan, "Probability-based collaborative filtering model for predicting gene–disease associations," *BMC Medical Genomics,* vol. 10, no. 5, p. 76, 2017.

[103] G. Wang *et al.*, "MeDReaders: a database for transcription factors that bind to methylated DNA," *Nucleic Acids Res,* vol. 46, no. D1, pp. D146-D151, Jan 4 2018.

[104] G. Wang *et al.*, "Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells," *BMC Genomics,* vol. 9 Suppl 2, p. S22, Sep 16 2008.

[105] L. Yu, J. Zhao, and L. Gao, "Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome," *Artificial Intelligence in Medicine,* vol. 77, pp. 53-63, Mar 2017.

[106] B. Liu, "BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches," *Briefings in bioinformatics,* vol. 20, no. 4, pp. 1280-1294, 2019.

[107] X. Zeng, L. Liu, L. Lü, and Q. Zou, "Prediction of potential disease-associated microRNAs using structural perturbation method," *Bioinformatics,* vol. 34, no. 14, pp. 2425-2432, 2018.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2983105, IEEE Access

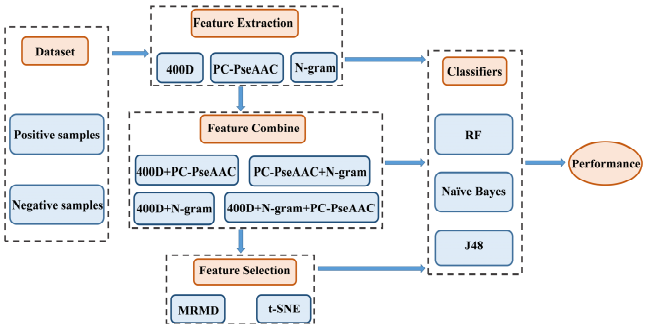

FIGURE 1. The flow-chart for predicting GPCRs. PC-PseAAC: parallel correlation pseudo amino acid composition; MRMD: max-relevance-max-distance; t-SNE: t-distributed stochastic neighbor embedding; RF: random forest.

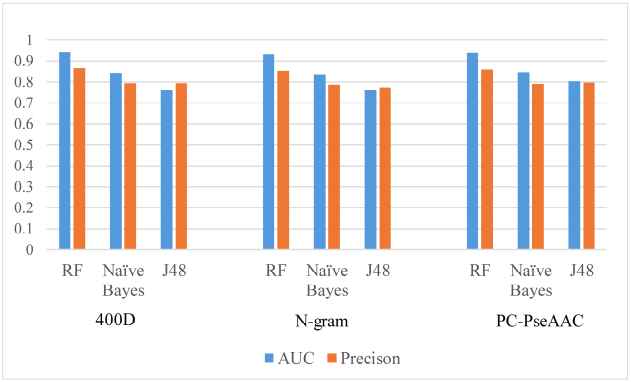

FIGURE 2. Comparison of AUC and Precision based on different feature extraction methods and different classifiers.

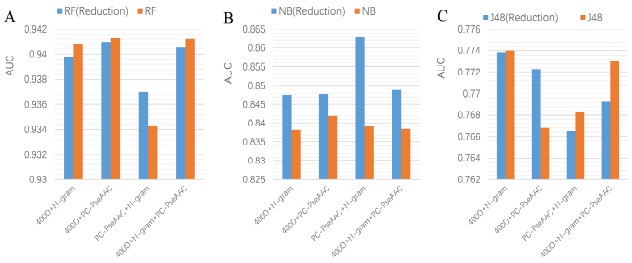

FIGURE 3. Comparison of classification results of mixed features and features after dimensionality reduction. Blue indicates the classification results after dimensionality reduction, and orange indicates the classification results of mixed features.

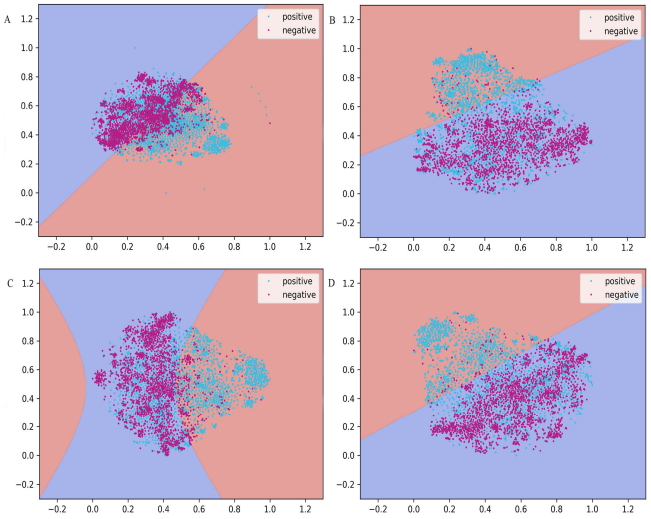

FIGURE 4. The visualization of four mixed vectors. The red points are negative examples, and the blue points are positive examples.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2983105, IEEE Access

TABLE 1. GPCRs identification results using single-feature representation methods.

Method	Classifier	AUC	Acc	Precision	Sn	Sp
400D	**RF**	**0.9408**	**0.8644**	**0.8653**	**0.8378**	**0.8910**
	Naïve Bayes	0.8430	0.7879	0.7923	0.7560	0.8200
	J48	0.7620	0.7935	0.7935	0.7913	0.7958
N-gram (N=2)	**RF**	**0.9315**	**0.8507**	**0.8515**	**0.8318**	**0.8698**
	Naïve Bayes	0.8378	0.7826	0.7865	0.7408	0.8245
	J48	0.7630	0.7737	0.7740	0.7813	0.7663
PC-PseAAC	**RF**	**0.9370**	**0.8564**	**0.8580**	**0.8202**	**0.8916**
	Naïve Bayes	0.8460	0.7837	0.7895	0.7150	0.8524
	J48	0.8033	0.7969	0.7970	0.7865	0.8072

TABLE 2. Classification result of the mixed-feature methods.

Method	Classifier	AUC	Acc	Precision
400D+N-gram	**RF**	**0.9408**	**0.8647**	**0.8658**
	Naïve bayes	0.8383	0.7922	0.7958
	J48	0.7740	0.7911	0.7910
400D+PC-PseAAC	**RF**	**0.9413**	**0.8628**	**0.8662**
	Naïve bayes	0.8420	0.7879	0.7923
	J48	0.7668	0.7916	0.7915
PC-PseAAC+N-gram	**RF**	**0.9343**	**0.8555**	**0.8565**
	Naïve bayes	0.8393	0.7865	0.7903
	J48	0.7683	0.7795	0.7798
400D+N-gram+PC-PseAAC	**RF**	**0.9412**	**0.8637**	**0.8648**
	Naïve bayes	0.8385	0.7927	0.7963
	J48	0.773	0.7894	0.7893

TABLE 3. Classification result of the reduction the features.

Method	Classifier	AUC	Acc	Precision
400D+N-gram	**RF**	**0.9398**	**0.8630**	**0.8638**
	Naïve bayes	0.8475	0.7994	0.8030
	J48	0.7738	0.7906	0.7905
400D+PC-PseAAC	**RF**	**0.9410**	**0.8639**	**0.8653**
	Naïve bayes	0.8478	0.7871	0.7913
	J48	0.7723	0.7921	0.7920
PC-PseAAC+N-gram	**RF**	**0.9370**	**0.8578**	**0.8585**
	Naïve bayes	0.8630	0.7996	0.8050
	J48	0.7665	0.7830	0.7830
400D+N-gram+PC-PseAAC	**RF**	**0.9405**	**0.8616**	**0.8628**
	Naïve bayes	0.8488	0.7981	0.8023
	J48	0.7693	0.7863	0.7865

TABLE 4. Comparison of the AUC of our method and literature method.

Method	feature	Classifier	AUC
The paper method	400D	RF	**0.9408**
	N-gram	RF	0.9315
	PC-PseAAC	RF	0.9370
Literature method	SVM-Prot	RF	0.9282