Hindawi
Scientific Programming
Volume 2020, Article ID 8862468, 7 pages
https://doi.org/10.1155/2020/8862468

Research Article

SubRF_Seq: Identification of Sub-Golgi Protein Types with Random Forest with Partial Sequence Information

Qingyu Cui,[1] Yi Cao,[1] Wenzheng Bao ,[2] Bin Yang,[3] and Yuehui Chen[1]

[1]*School of Information, University of Jinan, Jinan 250024, China*
[2]*School of Information Engineering, Xuzhou University of Technology, Xuzhou 221018, China*
[3]*School of Information Science and Engineering, Zaozhuang University, Zaozhuang 277100, China*

Correspondence should be addressed to Wenzheng Bao; baowz55555@126.com

Received 7 April 2020; Revised 13 May 2020; Accepted 16 June 2020; Published 16 July 2020

Academic Editor: Chenxi Huang

In the recent years, the subject of Golgi classification has been studied intensively. It has been scientifically proven that Golgi can synthesize many substances, such as polysaccharides, and it can also combine proteins with sugars or lipids with glycoproteins and lipoproteins. In some cells (such as liver cells), the Golgi apparatus is also involved in the synthesis and secretion of lipoproteins. Therefore, the loss of Golgi protein function may have severe effects on the human body. For example, Alzheimer's disease and diabetes are related to the loss of Golgi protein function. Because the classification of Golgi proteins has a specific effect on the treatment of these diseases, many scholars have studied the classification of Golgi proteins, but the data sets they used were complete Golgi sequences. The focus of this article is whether there is redundancy in the Golgi protein classification or, in other words, whether a part of the entire Golgi protein sequence can be used to complete the Golgi protein classification. Besides, we have adopted a new method to deal with the problem of sample imbalance. After experiments, our model has certain observability.

1. Introduction

Golgi is an organelle found in eukaryotic cells [1]. The Golgi was initially defined by Camilo-Golgi in 1897 and was named after Golgi in 1898 [2–4]. Considering its large size and unique structure, the Golgi apparatus can be treated as the first organelles which are discovered and observed in detail [5–7]. As part of the inner membrane system, Golgi proteins are encapsulated in membrane vesicles [8], which are sent to their destination. Golgi is located between the secretory pathway, the lysosome, and the endocytosis pathway [9]. Golgi plays an essential role in protein secretion. Meanwhile, such an issue contains a series of related glycosylases [10]. The subcellular position of the Golgi apparatus is different from that of various eukaryotic cells. In most eukaryotic cells, the Golgi apparatus includes cis-Golgi and trans-Golgi [11, 12]. Cis-Golgi is mainly composed of vesicles and multiple vesicles form the Golgi pile. Trans-Golgi is the final vesicle structure, where proteins are encapsulated in transport vesicles and sent to the lysosome, secretory pathway, or cell surface. The Golgi apparatus is closely related in the areas of structure and function [13, 14]. Each independent Golgi stack can contain several types of enzymes. These abovementioned enzymes can process several biological issues [15].

Disorders of protein metabolism are the core link leading to the development of many neurodegenerative diseases [16]. The Golgi apparatus is an essential organelle in the material metabolic pathway and must be closely related to it. Parkinson's disease [17] and Alzheimer's disease [18] are typical of neurodegenerative diseases. Experiments have shown that β-amyloid protein plays a central role in the pathological changes of Alzheimer's disease [19], and its metabolic disorder is closely related to the loss of a certain function of the Golgi apparatus. However, in order to understand the mechanism of Golgi function, an essential step is to find a Golgi-resident [20] and use the types and functions of the Golgi-resident protein to determine the principles of the disease. For example, the cause of the diseases is likely to be a lack of a Golgi-resident protein

[21, 22], resulting in a loss of Golgi function. Therefore, it is important to correctly judge the type of Golgi apparatus [23, 24].

With several years' effort, the prediction of the Golgi type has become one of the most significant hot subjects [25] in the field of computational biology and bioinformatics. Currently, simply knowing whether a protein is a Golgi-resident protein is not enough to fully explain the function of the Golgi body [26–28]. Further analysis of the specific type of Golgi-resident protein is needed. For now, some methods are applied to this subject. Ding et al. proposed the improved Mahalanobis Discriminant (MD) algorithm to predict Golgi-resident protein types in 2011 [29]. Dijk proposed the prediction of the Golgi-resident protein type of type II membrane proteins using structural information and transmembrane domain information in 2008 [30]. Jiao and Du proposed that the general form of Chou pseudoamino acids to predict the Golgi-resident protein type in 2016 [31]. Ding and Jiao used a relatively small data set with 150 Golgi proteins. Yang et al. created a new data set with 304 sub-Golgi proteins for training and 64 sub-Golgi proteins for testing classification models [21]. Ahmad and Hayat [32] proposed a Golgi protein classification model using multi-voting feature selection. Zhou [33] proposed XGBoost conditional covariance minimization based on multifeature fusion to predict Golgi protein types. Whether it is based on an amino acid feature extraction method or after multiple amino acid feature extractions and voting or multifeature fusion, they all use the complete amino acid sequence to extract features, and because they use the complete amino acid sequence to extract features, their models obtain considerable accuracy. However, we all know that the amino acid sequence of a Golgi is very long, and it will take a lot of effort to extract feature information on the entire amino acid sequence.

In this paper, we propose a new model, dubbed sub-RF_seq. In detail, if we do not use a complete protein sequence in feature extraction, some of them can also get considerable accuracy. Throughout this article, our work is summarized as follows: firstly, we propose 529 types of cutting sequences. The training set and test set are cut according to these 529 cutting types. Then, the 529 training sets are encoded. We use EAAC technology to extract features and put them into the RF classifier to train the model. Finally, we use the split to equal validation to balance the data set and test the classification effect of the Golgi apparatus. We use the random forest classifier to get the top 5 cutting sequence methods, and then put the features of these five cutting methods into other classifiers we have constructed and compare which classifier is the best classifier with the partial Golgi protein sequence.

Our workflow is as follows.

2. Methods and Materials

2.1. Data. This experiment uses a new data set created by Ahmad [12]. There are 87 cis-Golgi protein sequences and 217 trans-Golgi protein sequences in the training set. No protein has more than 40% pairing with any other protein in the data set. The 64 sub-Golgi protein sequences were independently used for testing the effect of the classifier, of which 13 were cis-Golgi protein sequences and 51 were trans-Golgi protein sequences. It should be noted that there is no connection between the training set and the test set.

Our work flow chart is shown in Figure 1. Specifically, we need to process the complete sub-Golgi protein sequence. In this step, the 304 sub-Golgi protein sequences in the training set are cut. The cutting method is to cut three positions in the front and three positions from the back to form a new protein sequence. This forms the first partial Golgi sequence. Then, the front three digits are unchanged, the back cleavage digit is increased by one, and it is added to the back cleavage 25 to form 23 new protein sequences and form 23 partial Golgi training sets. Then, the number of front-end cuts is increased by one, and the number of rear-end cuts is from 3–25, until the last front-end cut is 25 digits and the back-end cut is 25 digits. There are 23×23 different cutting methods. 23×23 incomplete Golgi protein sequences were formed. The test set adapts the same cutting methods. Then, use EAAC to extract protein sequence features, input to the classification model to train the model, and then test the effect on an independent test set.

2.2. Feature Extraction

2.2.1. Amino Acid Composition Encoding. The sequence information of the Golgi apparatus contains the types and arrangement order of 20 amino acids [34, 35]. Therefore, the feature extraction algorithm based on the amino acid composition is the simplest and most intuitive method. The amino acid composition simply represents the probability of 20 kinds of amino acids appearing in the sequence [36, 37]. It is a basic Golgi sequence feature extraction algorithm. The amino acid composition maps the Golgi sequence to a point in the 20-dimensional European space. The vector is expressed as follows:

$$V_{\text{aac}} = (v_1, v_2, v_3, \ldots, v_{20})T,$$

$$V_i = \frac{f\_i}{\sum_{j=1}^{20} f\_i}, \tag{1}$$

$$\sum_{j=1}^{20} v_j = 1.$$

Here, f_i is the number of times the ith amino acid appears in the sequence ($i = 1, 2, 3, \ldots, 20$). The amino acid composition is easy to calculate, and it is the most commonly used sequence feature extraction algorithm in Golgi classification research.

2.2.2. Enhanced Amino Acid Content Encoding (EAAC). Chen et al. [38] proposed a new encoding method based on AAC encoding, dubbed EAAC. EAAC coding directly reflects the distribution frequency of 20 amino acid residues. EAAC coding differs from AAC coding in that EAAC coding defines a sliding window of length 8 and calculates the

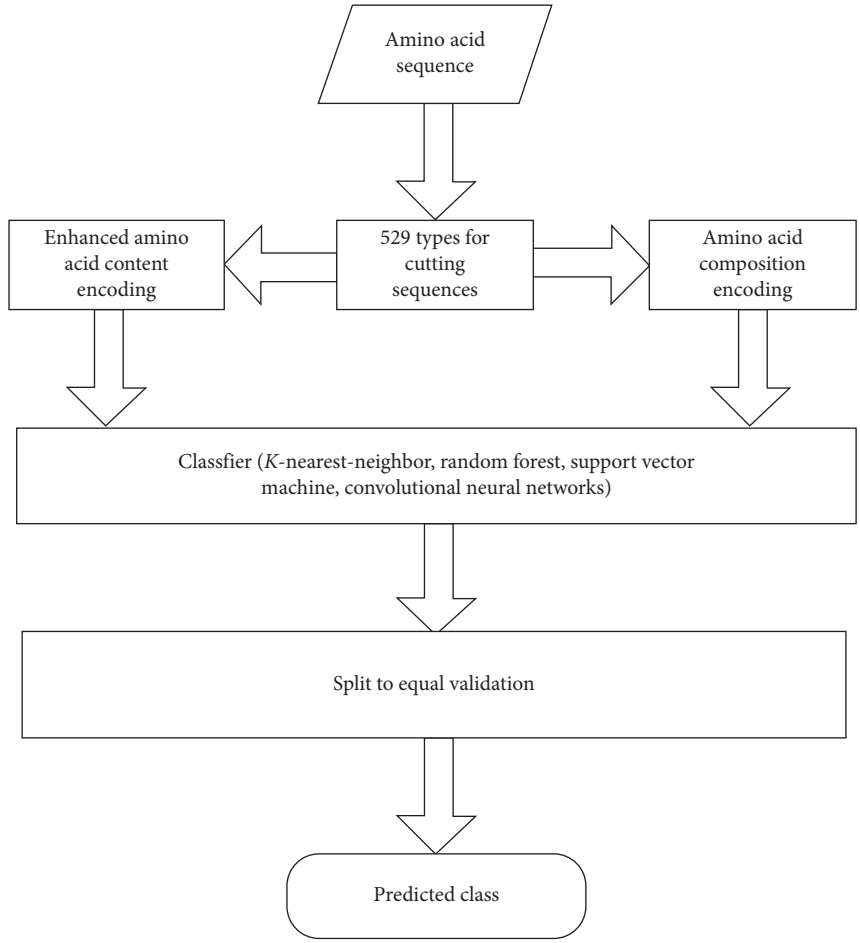

Figure 1: Illustration of the process of the proposed method.

frequency of 20 amino acid residues that appear in each 8-dimensional subsequence segment [39]. The frequency of 20 amino acid residues is continuously sliding in the window from the N-terminus to the C-terminus of each Golgi sequence in the dataset. Therefore, the vector dimension corresponding to a Golgi sequence of x residues is

$$\begin{aligned} L_s &= x - L_v + 1, \\ D_s &= L_s \times 20. \end{aligned} \tag{2}$$

Here, L_v is the size of the sliding window we defined. In EAAC encoding, the value of L_v is 8, x is the length of the Golgi sequence, and the D_s is the dimension of the feature vector.

2.3. Construction of the Classifier. This experiment mainly uses a classifier of random forests. Random forests are called "representative methods for ensemble learning" [40], which is easy to implement and has relatively low overhead. Random forests are an extension of Bagging's idea [41], which is based on decision tree learning, and the algorithm further introduces random attribute selection in the training process of the decision tree [42–44]. The basic idea of random forest is to train the model with data, then get multiple decision trees, and then merge the decision trees to get more stable predictions. In random forests, the performance becomes better as the number of trees increases, and the error becomes smaller. In this experiment, we selected 1000 decision trees to build a random forest model. In addition, we also constructed KNN (K nearest neighbor classification algorithm), SVM (Support Vector Machine Algorithm), CNN (Convolutional Neural Network), and ANN (Artificial Neural Network) classifiers to compare which is in the best classifiers with the part of Golgi protein sequences.

2.4. Evaluation Methods. The positive and negative samples of the training set of this experiment are imbalanced, and the ratio of positive and negative samples is about 1 : 2. In the binary classification problem, the imbalance of positive and negative samples will have a certain impact on the classification effect. It will cause the prediction category towards the category with many samples. Therefore, for the evaluation method, we chose an SE verification method proposed by Sun et al. [45]. The advantage of this verification method is that data processing and cross-validation can be implemented at the same time.

Performance measurement is an evaluation standard for measuring the generalization ability of the model, which

reflects the needs of the task. The use of different performance metrics often leads to different evaluation results. Therefore, it is essential to choose a good set of performance indicators to predict the performance of the model. In this experiment, ACC and AUC were selected for evaluation. ACC and AUC performance indicators have evolved from the confusion matrix [46–51]. In the binary classification problem, when the real situation of data classification in the test set is a positive example, the model prediction result is a positive example, which is called the real example (TP). When the predicted outcome is a counterexample, it is a false counterexample (FN). Similarly, when the true situation of the data classification of the test set is a counterexample, there are false positive examples (FP) and true counterexamples (TN). The accuracy rate formula is

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{3}$$

The recall formula is

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4}$$

The formula for the accuracy rate (ACC) is

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{5}$$

The value of AUC is the area of the ROC curve. We often use the value of AUC as the criterion for judging the quality of the model because the ROC curve cannot intuitively see the quality of the model [52, 53]. ROC is a curve drawn with sensitivity as the vertical axis and 1 minus specificity as the horizontal axis.

The formula for sensitivity is

$$\text{Sens} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{6}$$

3. Results and Discussion

In this section, we mainly describe the effect of the 529 incomplete Golgi sequences we have defined for training the model. Besides, we chose the top 5 cutting methods for classification effects in the sub_RF_seq model for comparison experiments.

3.1. Results. In this experiment, we recorded the AUC values of 529 different cutting methods. In order to intuitively understand the classification effect of these 529 cutting methods, we made a three-dimensional histogram based on the AUC values. The X-axis represents how many bits are cut from the front end of the protein sequence, and the Y-axis represents the number of bits cut from the rear end of the protein. The X-axis represents how many bits are cut from the front end of the protein sequence, and the Y-axis represents the number of bits cut from the back end of the protein. In this way, this three-dimensional histogram shows the classification effect using incomplete Golgi protein sequences. Figure 2 shows that, among these 529 Golgi sequence cutting methods, 202 of the cutting methods have an AUC value greater than or equal to 0.6, and 426 of the cutting methods have a value greater than or equal to 0.5.

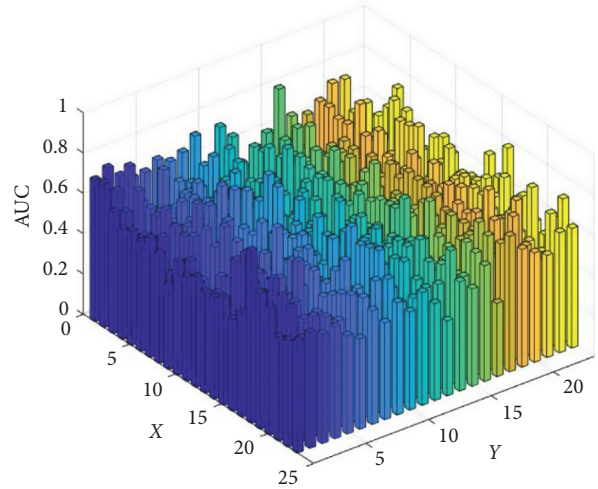

Figure 2: Three-digit histogram of an AUC value of 529 cutting methods.

In addition, we used the random forest classifier to select the top 5 cutting methods for Golgi classification. The values of AUC and ACC for these five cutting methods are shown in Table 1.

3.2. Discussion

3.2.1. Comparison of Model Effects under Different Classifiers. We put the cutting sequence of the top 5 classification effects in the model into the SVM, KNN, CNN, and ANN classifiers and compared which classifiers used the partial Golgi sequence to achieve the best Golgi classification. From Table 2, we found that the RF classifier performs better than several other classifiers. For example, under the premise of a certain cutting sequence method, EAAC coding is selected for the feature coding method. In the 20 + 3 Golgi sequence, the value of ACC in the RF classifier is as high as 82.81%, and the value of AUC is as high as 0.854. The values are better than several other classifiers. However, the classification effect of partial Golgi sequences in other classifications is still considerable. In Table 2, the AUC and Acc values of most classifiers are above 70%, which further to confirm that there is a certain redundancy in the Golgi sequence when it is used to determine the Golgi types.

3.2.2. Classification Effect under Different Encoding Methods. In this experiment, we chose two encoding methods, EAAC and AAC, to see the effect of different amino acid sequence encoding methods of the classification effect. In order to explore the classification effect under different encoding methods, we controlled the variable classifier. Only the RF classifier is selected. From Table 3, we can see the AUC and ACC values of the five cutting methods under the EAAC and AAC. Table 3 shows that, in the EAAC encoding mode, the values of Acc and AUC are higher than those in the AAC

Table 1: Sub-Golgi protein sequence cutting methods ranked by the AUC value.

Classifier	Encoding schemes	Cutting method	AUC	ACC
RF	EAAC	20 + 3	0.854449	0.828125
		4 + 17	0.849170	0.859375
		18 + 25	0.782805	0.828125
		20 + 11	0.782805	0.796875
		11 + 11	0.773002	0.828125

Table 2: Comparison of the effects of different classifiers.

Cutting	RF		SVM		CNN		KNN		ANN	
	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC
20 + 3	82.81	0.8544	67.19	0.6448	70.31	0.7179	62.5	0.549	78.13	0.4042
4 + 17	85.94	0.8492	76.56	0.7858	76.56	0.7451	67.19	0.6531	78.13	0.6456
18 + 25	82.81	0.7828	65.63	0.6365	78.13	0.7602	60.94	0.7451	68.75	0.4781
20 + 11	79.69	0.7828	68.75	0.7873	81.25	0.736	62.5	0.7572	60.03	0.7188
11 + 11	82.81	0.773	70.31	0.7451	73.44	0.7315	56.25	0.5038	68.75	0.4962

Table3: Classification effect under different encoding methods.

Cutting	Classifier	EAAC encoding		AAC encoding	
		Acc (%)	AUC	Acc (%)	AUC
20 + 3	RF	82.81	0.8544	65.63	0.4434
4 + 17		85.94	0.8492	70.31	0.4894
18 + 25		82.81	0.7828	71.88	0.7549
20 + 11		79.69	0.7828	68.75	0.586
11 + 11		82.81	0.7730	64.06	0.4563

Table 4: Performance of imbalance of positive and negative samples of the data set on the classification effect.

Cutting	Classifier	Encoding	SEV		10-flod CV	
			Acc (%)	AUC	Acc (%)	AUC
20 + 3	RF	EAAC	82.81	0.8544	78.13	0.7813
4 + 17			85.94	0.8492	67.95	0.6161
18 + 25			82.81	0.7828	79.69	0.4615
20 + 11			79.69	0.7828	77.75	0.7681
11 + 11			82.81	0.7730	80.06	0.5716

encoding method, which directly proves our guess that different encoding methods will affect the classification effect of the model.

3.2.3. Performance of Imbalance of Positive and Negative Samples of the Data Set on the Classification Effect. Due to the imbalance of the positive and negative samples in the data set we used, we used both the SEV verification method and the 10-fold cross-validation method to verify the classification effect of the model. The SEV verification method can deal with the imbalance of the positive and negative samples of the data set, and the 10-fold cross-validation does not have the effect of data preprocessing. Table 4 proves that processing the imbalance of the data set will improve the model's effectiveness. Using SEV is nearly 18% higher than a simple 10-fold cross-validation.

4. Conclusions

In the past, when determining the type of Golgi apparatus, many people used the entire Golgi protein sequence in encoding; a complete Golgi protein sequence has a large number of amino acids, which is very time-consuming when encoding. In this article, we present subRF_seq, which can complete the classification of Golgi using a part of the Golgi protein sequence and has a considerable classification effect. We cut the data set, extract the feature vector from the cut sequence, and finally, train it in a random forest to distinguish trans-Golgi and cis-Golgi. Also, in the binary classification problem, the proportion of positive and negative samples of many training sets cannot reach 1 : 1, which will cause the problem of falsely high AUC values. Our model can effectively overcome this problem. We also used other classifiers and feature extraction techniques to prove our ideas, and the results show that our ideas of using part of the Golgi sequence in feature extraction is feasible because the values of AUC and ACC are considerable in different classifiers and encoding methods. The experimental results prove that Golgi proteins can still be distinguished by using partial Golgi sequences, In other words, there is a certain degree of redundancy in Golgi protein classification on Golgi classification. If we use part of the Golgi sequence in Golgi classification, it will significantly reduce the time.

Data Availability

To data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the grants from the National Science Foundation of China (nos. 61902337 and 61702445)

Jiangsu Provincial Natural Science Foundation, China (no. SBK2019040953), and Natural Science Fund for Colleges and Universities in Jiangsu Province no. 19KJB520016.

References

[1] G. Griffiths and K. Simons, "The trans Golgi network: sorting at the exit site of the Golgi complex," *Science*, vol. 234, no. 4775, pp. 438–443, 1986.

[2] M. Gribskov, I. Mellman, and K. Simons, "The golgi complex: in vitro veritas?," *Cell*, vol. 68, no. 5, pp. 829–840, 1991.

[3] Q. Zou and Q. Liu, "Advanced machine learning techniques for bioinformatics," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 16, no. 4, pp. 1182-1183, 2019.

[4] I. Mishqat, "Camillo Golgi's black reaction for staining neurons," Embryo Project Encyclopedia, Washington, DC, USA, 2017.

[5] S. Bassnett, "The fate of the Golgi apparatus and the endoplasmic reticulum during lens fiber cell differentiation," *Investigative Ophthalmology & Visual Science*, vol. 36, no. 9, pp. 1793–1803, 1995.

[6] J. Rothman, "The Golgi apparatus: two organelles in tandem," *Science*, vol. 213, no. 4513, p. 1212, 1981.

[7] W. He, L. Wei, and Q. Zou, "Research progress in protein posttranslational modification site prediction," *Briefings in Functional Genomics*, vol. 18, no. 4, pp. 220–229, 2019.

[8] M. Rao and C. R. Alving, "Delivery of lipids and liposomal proteins to the cytoplasm and Golgi of antigen-presenting cells," *Advanced Drug Delivery Reviews*, vol. 41, no. 2, pp. 171–188, 2000.

[9] Z. Wang, H. Ding, and Q. Zou, "Identifying cell types to interpret scRNA-seq data: how, why and more possibilities," *Briefings in Functional Genomics*, 2020.

[10] L. Yuan, F. Guo, L. Wang, and Q. Zou, "Prediction of tumor metastasis from sequencing data in the era of genome sequencing," *Briefings in Functional Genomics*, vol. 18, no. 6, pp. 412–418, 2019.

[11] B. H. Hummer, D. Maslar, M. S. Gutierrez, N. F. D. Leeuw, and C. S. Asensio, "Differential sorting behavior for soluble and transmembrane cargoes at the trans-Golgi network in endocrine cells," *Molecular Biology of the Cell*, vol. 31, no. 3, pp. 157–166, 2020.

[12] J. Ahmad, F. Javed, and M. Hayat, "Intelligent computational model for classification of sub-Golgi protein using oversampling and Fisher feature selection methods," *Artificial Intelligence in Medicine*, vol. 78, pp. 14–22, 2017.

[13] S. Deng, H. Liu, K. Qiu, H. You, Q. Lei, and W. Lu, "Role of the Golgi apparatus in the blood-brain barrier: golgi protection may be a targeted therapy for neurological diseases," *Molecular Neurobiology*, vol. 55, no. 6, pp. 4788–4801, 2018.

[14] J. Villeneuve, J. Duran, M. Scarpa, L. Bassaganyas, J. V. Galen, and V. Malhotra, "Golgi enzymes do not cycle through the endoplasmic reticulum during protein secretion or mitosis," *Molecular Biology of the Cell*, vol. 28, no. 1, pp. 141–151, 2017.

[15] Y. Hou, J. Dai, J. He, A. J. Niemi, X. Peng, and N. Ilieva, "Intrinsic protein geometry with application to non-proline cis peptide planes," *Journal of Mathematical Chemistry*, vol. 57, no. 1, pp. 263–279, 2019.

[16] 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.

[17] J. M. V. D. Elsen, D. A. Kuntz, and D. R. Rose, "Structure of Golgi α-mannosidase II: a target for inhibition of growth and metastasis of cancer cells," *The EMBO Journal*, vol. 20, no. 12, pp. 3008–3017, 2001.

[18] S. Hoyer, "Is sporadic Alzheimer disease the brain type of non-insulin dependent diabetes mellitus? A challenging hypothesis," *Journal of Neural Transmission*, vol. 105, no. 4, pp. 415–422, 1998.

[19] D. D. Elsberry and M. T. Rise, "Method of treating movement disorders by brain infusion," U.S. Patent No. 6,042,579, U.S. Patent and Trademark Office. Washington, DC, USA, 2000.

[20] B. Radau, A. Otto, E.-C. Müller, and P. Westermann, "Protein kinase Cα-dependent phosphorylation of Golgi proteins," *Electrophoresis*, vol. 21, no. 13, pp. 2684–2687, 2000.

[21] R. Yang, C. Zhang, R. Gao, and L. Zhang, "A novel feature extraction method with feature selection to identify Golgi-resident protein types from imbalanced data," *International Journal of Molecular Sciences*, vol. 17, no. 2, p. 218, 2016.

[22] A. E. Cuadra, S.-H. Kuo, Y. Kawasaki, D. S. Bredt, and D. M. Chetkovich, "AMPA receptor synaptic targeting regulated by stargazin interactions with the Golgi-resident PDZ protein nPIST," *Journal of Neuroscience*, vol. 24, no. 34, pp. 7491–7502, 2004.

[23] I. J. Goldstein, C. E. Hollerman, and E. E. Smith, "Protein-carbohydrate interaction. II. Inhibition studies on the interaction of concanavalin A with polysaccharides," *Biochemistry*, vol. 4, no. 5, pp. 876–883, 1965.

[24] H. Ding, S.-H. Guo, E.-Z. Deng et al., "Prediction of Golgi-resident protein types by using feature selection technique," *Chemometrics and Intelligent Laboratory Systems*, vol. 124, pp. 9–13, 2013.

[25] Z. Yuan and R. D. Teasdale, "Prediction of Golgi type II membrane proteins based on their transmembrane domains," *Bioinformatics*, vol. 18, no. 8, pp. 1109–1115, 2002.

[26] P. Cosson, M. Amherdt, J. E. Rothman, and L. Orci, "A resident Golgi protein is excluded from peri-Golgi vesicles in NRK cells," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12831–12834, 2002.

[27] Y. Jiao, P. Du, and X. Su, "Predicting Golgi-resident proteins in plants by incorporating *N*-terminal transmembrane domain information in the general form of Chou's pseudoamino acid compositions," in *Proceedings of the 2014 8th International Conference on Systems Biology (ISB)*, pp. 226–229, Qingdao, China, 2014, October.

[28] C. Y. L. Yuen, P. Wang, B.-H. Kang, K. Matsumoto, and D. A. Christopher, "A non-classical member of the protein disulfide isomerase family, PDI7 of *Arabidopsis thaliana*, localizes to the cis-Golgi and endoplasmic reticulum membranes," *Plant and Cell Physiology*, vol. 58, no. 6, pp. 1103–1117, 2017.

[29] H. Ding, L. Liu, F.-B. Guo, J. Huang, and H. Lin, "Identify Golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition," *Protein and Peptide Letters*, vol. 18, no. 1, pp. 58–63, 2011.

[30] A. D. V. Dijk, D. Bosch, C. J. T. Braak, A. R. V. D. Krol, and R. C. H. J. V. Ham, "Predicting sub-Golgi localization of type II membrane proteins," *Bioinformatics*, vol. 24, no. 16, pp. 1779–1786, 2008.

[31] Y.-S. Jiao and P.-F. Du, "Predicting Golgi-resident protein types using pseudo amino acid compositions: approaches with positional specific physicochemical properties," *Journal of Theoretical Biology*, vol. 391, pp. 35–42, 2016.

[32] J. Ahmad and M. Hayat, "MFSC: multi-voting based feature selection for classification of Golgi proteins by adopting the

general form of Chou's PseAAC components," *Journal of Theoretical Biology*, vol. 463, pp. 99–109, 2019.

[33] H. Zhou, C. Chen, M. Wang, Q. Ma, and B. Yu, "Predicting golgi-resident protein types using conditional covariance minimization with XGBoost based on multiple features fusion," *IEEE Access*, vol. 7, pp. 144154–144164, 2019.

[34] A. Raina and A. Datta, "Molecular cloning of a gene encoding a seed-specific protein with nutritionally balanced amino acid composition from Amaranthus," *Proceedings of the National Academy of Sciences*, vol. 89, no. 24, pp. 11774–11778, 1992.

[35] J. Adachi, P. J. Waddell, W. Martin, and M. Hasegawa, "Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA," *Journal of Molecular Evolution*, vol. 50, no. 4, pp. 348–358, 2000.

[36] V. J. Vivekanand and J. Ramana, "Prediction of lysosomal membrane proteins using machine learning techniques," 2014.

[37] S.-Y. Kung and M.-W. Mak, "Feature selection for self-supervised classification with applications to microarray and sequence data," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 297–309, 2008.

[38] Z. Chen, N. He, Y. Huang, W. T. Qin, X. Liu, and L. Li, "Integration of a deep learning classifier with a random forest approach for predicting malonylation sites," *Genomics, Proteomics & Bioinformatics*, vol. 16, no. 6, pp. 451–459, 2018.

[39] H. Neumann, K. Wang, L. Davis, M. Garcia-Alai, and J. W. Chin, "Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome," *Nature*, vol. 464, no. 7287, pp. 441–444, 2010.

[40] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[41] G. Biau, "Analysis of a random forests model," *Journal of Machine Learning Research*, vol. 13, no. 2012, pp. 1063–1095, 2012.

[42] A. C. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble Machine Learning*, pp. 157–175, Springer, Boston, MA, USA, 2012.

[43] A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof, "On-line random forests," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 1393–1400, Kyoto, Japan, 2009, September.

[44] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, 2012.

[45] J. Sun, Y. Cao, D. Wang, W. Bao, and Y. Chen, "K_net: lysine malonylation sites identification with neural network," *IEEE Access*, vol. 8, pp. 47304–47311, 2019.

[46] M. Tahir and M. Hayat, "iNuc-STNC: a sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC," *Molecular BioSystems*, vol. 12, no. 8, pp. 2587–2593, 2016.

[47] Y. Jiao and P. Du, "Performance measures in evaluating machine learning based bioinformatics predictors for classifications," *Quantitative Biology*, vol. 4, no. 4, pp. 320–330, 2016.

[48] I. S. MacKenzie, T. Kauppinen, and M. Silfverberg, "Accuracy measures for evaluating computer pointing devices," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 9–16, Seattle, WA, USA, 2001, March.

[49] L. J. Siegel, H. J. Siegel, and P. H. Swain, "Performance measures for evaluating algorithms for SIMD machines," *IEEE Transactions on Software Engineering*, vol. 8, no. 4, pp. 319–331, 1982.

[50] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 14, no. 2, pp. 1137–1145, Montreal, Canada, 1995, August.

[51] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[52] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.

[53] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240, Pittsburgh, PA, USA, 2006, June.