Hindawi
BioMed Research International
Volume 2020, Article ID 9701734, 8 pages
https://doi.org/10.1155/2020/9701734

Research Article

Predicting Gram-Positive Bacterial Protein Subcellular Location by Using Combined Features

Feng-Min Li and Xiao-Wei Gao

College of Science, Inner Mongolia Agricultural University, Hohhot 010018, China

Correspondence should be addressed to Feng-Min Li; fmli@imau.edu.cn

Received 23 May 2020; Revised 30 June 2020; Accepted 13 July 2020; Published 3 August 2020

Guest Editor: Quan Zou

There are a lot of bacteria in the environment, and Gram-positive bacteria are the most common ones. Some Gram-positive bacteria are very harmful to the human body, so it is significant to predict Gram-positive bacterial protein subcellular location. And identification of Gram-positive bacterial protein subcellular location is important for developing effective drugs. In this paper, a new Gram-positive bacterial protein subcellular location dataset was established. The amino acid composition, the gene ontology annotation information, the hydropathy dipeptide composition information, the amino acid dipeptide composition information, and the autocovariance average chemical shift information were selected as characteristic parameters, then these parameters were combined. The locations of Gram-positive bacterial proteins were predicted by the Support Vector Machine (SVM) algorithm, and the overall accuracy (OA) reached 86.1% under the Jackknife test. The overall accuracy (OA) in our predictive model was higher than those in existing methods. This improved method may be helpful for protein function prediction.

1. Introduction

The cell is the most basic unit of life, and it contains many protein molecules. When a protein is in the right subcellular position, it can perform the right function [1]. So, studying protein subcellular location can help us better understand the biological function of proteins at the cellular level. In the postgenetic era, the amount of biological information has grown rapidly and the traditional experimental method became time-consuming and exhausting. So, the prediction of protein subcellular location based on the machine method has gradually become a hot research topic in bioinformatics [2–7].

Gram-positive bacteria are those that retain their original blue-violet color after being stained by Gram staining. Gram-positive bacteria exist widely in the human body, and they are harmful to the environment and human health. So, it is important to study the protein subcellular location of Gram-positive bacteria. There are a few researches on the protein subcellular location of Gram-positive bacteria. In 2007, Shen and Chou [8] established a Gram-positive bacteria dataset of five categories. They used the GO-PseAA discrete model and the Fusion OET-KNN method, and the overall success accuracy was 82.7% with the Jackknife test. In 2009, Shen and Chou [9] rebuilt the Gram-positive bacteria dataset with four categories: cell wall, cell membrane, cytoplasm, and extracell. The feature of gene ontology information and functional domain information were extracted, and the total success accuracy reached 82.2% with the Jackknife test. In 2012, the total success accuracy was 85.9% for the GP25 dataset constructed by Hu et al. [10]. In the 9th international conference on electrical and computer engineering, Rahman et al. [11] proposed two hybrid features, AACPPM and PAACPPM, which combined PPM with AAC and PseAAC, respectively. The accuracy of both AACPPM and PAACPPM were 73.2%. In 2017, Xiao et al. [12] took advantage of the dataset established by Shen and Chou in 2009 and applied the new algorithm, and a better result was obtained. In 2018, Xiao et al. [13] developed a new bias-reducing predictor. The results showed that this

predictor was very helpful in predicting the training dataset.

In this paper, we reconstructed the Gram-positive bacterial protein subcellular location dataset. The amino acid composition information [14], the amino acid dipeptide composition information [15, 16], the gene ontology [17] annotation information, the hydropathy dipeptide [18] composition information, and the autocovariance average chemical shift [19] information were selected as characteristic parameters, then these parameters were combined. Finally, the overall accuracy in the Jackknife test was 86.1% by using the combined parameter AAC+DC+hpDC for the Support Vector Machine.

2. Materials and Methods

2.1. Dataset. In order to collect as much desired information as possible while ensuring a high quality for the dataset, the protein sequences were collected from the Swiss-Prot [20] database at http://www.uniprot.org/. The dataset was established in strict accordance with the following criteria: (1) We conducted a search for all protein sequences with "actinobacteria" and "firmicutes" in the OC firmicutes from the UniProtKB/Swiss-Prot database. (2) Different locations of the protein in the "Subcellular Location" annotation were selected, and the ambiguous or uncertain terms, such as "By similarity" and "Probably" were removed. (3) The protein sequence of 50 aa-3000 aa in the "Sequence" information were selected. (4) Sequences annotated by two or more locations were not included. (5) Sequences annotated with "fragment," "B," "X," and "Z" were excluded. (6) To avoid any homology bias, the software CD-HIT [21] was used to winnow those sequences which have ≥25% sequence identity to any other sequence in the same subcellular location.

After completing the above steps, we obtained 700 Gram-positive proteins, and the specific distribution is shown in Table 1.

2.2. Amino Acid Composition (AAC). The sequence information of proteins is the most basic feature information of all characteristic parameters [22]. The protein sequence consists of 20 amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y). The feature of the occurrence frequency of the 20 amino acids in the protein is important. So, the occurrence frequency of the 20 amino acids in the protein sequence can be selected as one of the characteristic parameters. The amino acid composition can be expressed as a 20-dimensional feature vector:

$$\mathrm{AAC} = [f_1, f_2, f_3, \cdots f_{20}]^T, \quad (1)$$

where $f_i = (n_i/L)(i = 1, 2, \cdots, 20)$, n_i is the occurrence number of the 20 native amino acids of the protein, L is the length of the protein, and T is the transpose operator.

2.3. Dipeptide Composition (DC). One of the main drawbacks of the amino acid composition is that it only emphasizes on overall sequence information but ignores the sequence order information. In order to make full use of the sequence information of amino acids, we proposed using the amino acid dipeptide composition information. The amino acid dipeptide information is an improvement based on the AAC parameter, and it denotes the frequency of two adjacent amino acids in a 400-dimensional vector [23–25]. The dipeptide composition can be formulated as follows:

$$\mathrm{DC} = [d_1 \cdots d_i \cdots d_{400}]^T, \quad (2)$$

where $d_i(i = 1, 2 \cdots 400)$ is the absolute occurrence frequencies of the 400 dipeptides and calculated by

$$d_i = (n_i/L - 1), \quad (3)$$

where n_i is the occurrence number of the 400 dipeptides of the protein and L is the length of the protein.

Table 1: Dataset of Gram-positive bacteria subcellular location proteins.

Subcellular location	Number of proteins
Cell wall	22
Extracell	214
Cytoplasm	252
Cell membrane	212
Total	700

2.4. Gene Ontology (GO). Gene ontology is a directed acyclic graph ontology widely used in bioinformatics, and gene ontology consists of three parts: biological process (P), molecular function (F), and cellular component (C). In the gene ontology database, we found that each AC number has a corresponding GO identification number: XXXXXXX. In this paper, since cellular component (C) contains the location information of a protein, in order to ensure the accuracy of the prediction, only biological process (P) and molecular function (F) were extracted.

The specific steps are as follows:

Step 1. The "Text" documents of all protein sequences were downloaded in Swiss-Port, and the annotation information of all biological processes (P) and molecular functions (F) was extracted.

Step 2. BLAST [26] was used to find homologous sequences of biological process (P) and molecular function (F) without annotation information. The homology threshold was set to 60%, and the E value was set to 0.001.

Step 3. The frequency of occurrence of each GO term was calculated:

$$f = \frac{N_x^i}{N_x}, \quad (4)$$

where N_x^i denotes the frequency of the ith GO terms at the x position of Gram-positive bacteria and N_x is the total

number of amino acid sequences at the x position of Gram-positive bacteria. A threshold value$T$was set; when $f > T$, the corresponding GO terms were retained.

Step 4. The GO terms of all target sequences were integrated and repeated, then 2573 GO terms were acquired. Finally, the 2573 GO terms were integrated into one vector, P_{GO}:

$$P_{\mathrm{GO}} = \{\psi_1, \psi_2, \cdots \psi_n, \cdots \psi_{2573}\}, \quad (5)$$

where ψ_n is 0 or 1, and the GO number with the corresponding location information of the proteins was set to 1; otherwise, it was 0.

2.5. Autocovariance Average Chemical Shifts (acACS). The most important issue is how to extract features from primary sequences of a protein in a predictor. Hence, the acACS [27, 28] algorithm that uses simple secondary structure information to represent the sample of a protein was proposed. The average chemical shift of a protein is closely related to the protein's secondary structure [29] and the function of this protein. The secondary structure of the protein sequence (C, H, and E) was obtained by submitting the protein sequence to the PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) online tool, and then the secondary structure was submitted to Fan et al.'s [30] average chemical shift service website acACS (http://wlxy.imu.edu.cn/college/biostation/fuwu/acACS/index.asp) to obtain the results of the chemical shifts. For a protein P

$$P = [j_1, j_2 \cdots j_i \cdots j_L], \quad (6)$$

where L means the length of the protein sequence P and j is the 20 amino acid residues; thus, P can be expressed as follows:

$$P^i_{\mathrm{acACS}} = \left[\psi^i(0), \psi^i(1), \psi^i(2), \cdots, \psi^i(\lambda)\right], \quad \left(i = {}^{15}\mathrm{N}, {}^{13}\mathrm{C}_\alpha, {}^{1}\mathrm{H}_\alpha, {}^{1}\mathrm{H}_N, \ 0 < \lambda < L\right), \quad (7)$$

where $\psi^i(\lambda)$ represents the correlation factor of the average chemical shift for j_l with the average chemical shift for $j_{l+\lambda}$ along the protein sequence. The factor $\lambda(0 < \lambda < L)$ means the rank of correlation. The factor i can be represented in a different composition of ^{15}N, $^{13}\mathrm{C}_\alpha$, $^{1}\mathrm{H}_\alpha$, and $^{1}\mathrm{H}_N$. In order to obtain the best accuracy, an appropriate number factor λ and the best combination mode i were selected to predict the results.

2.6. Hydropathy Dipeptide Composition (hpDC). Hydropathy dipeptide composition is based on the improvement of hydrophilic and hydrophobic proteins. Firstly, 20 kinds of amino acids were divided into 6 categories [31] according to the hydrophilic and hydrophobic standards, namely, strong hydrophilic amino acids (H), strong hydrophobic amino acids (L), weak hydrophilic amino acids or weak hydrophobic amino acids (W), and three types of proline (P), glycine (G), and cysteine (C) with special chemical structures. Hydrophilic and hydrophobic dipeptide composition

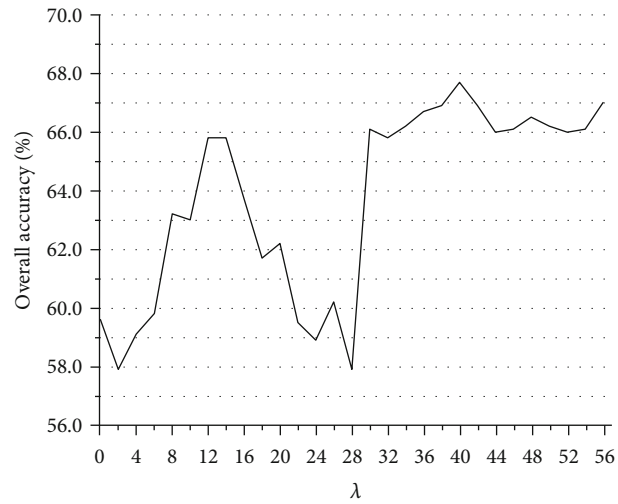

Figure 1: Predictive results with respect to the correlation factor λ of the acACS based on the Jackknife test. The best results obtained with $\lambda = 40$.

is a discrete method that uses protein sequence representation, and it can be represented as a 36-dimensional vector:

$$P_{\mathrm{hpDC}} = [q_1 \cdots q_i \cdots q_{36}]^T, \quad (8)$$

where $q_i = (n_i/L - 1) \ (i = 1, 2 \cdots 36)$ represents the occurrence frequencies of the 36 hydropathy dipeptides, while n_i denotes the occurrence number of the 36 hydropathy dipeptides of the protein and L is the length of the protein.

2.7. Support Vector Machine (SVM). The Support Vector Machine is a machine learning method to solve classification and regression problems based on statistical principles. The SVM model is a representation of the examples as points in space, mapped by a kernel function so that the examples are divided by a clear gap that is wide enough. The new examples are mapped into the same space and predicted according to which side of the gap they fall on. The radial basis kernel function (RBF) was used to obtain the best classification hyperplane. The regularization parameter C and the kernel width parameter γ were tuned via the grid search method. So far, the risk minimization of the SVM algorithm has become the latest research hotspot and it has been successfully applied to various fields [32–38], especially in the field of biological computing, such as in the prediction of protein sequence structure and in the classification of protein structure [28, 39–46]. In this paper, the LIBSVM algorithm has been used to predict various feature information, which can be downloaded from http://www.csie.ntu.edu.tw/cjlin/libsvm/.

3. Results

3.1. Cross-Validation. In statistical prediction, three test methods of prediction accuracy are used: the Jackknife test,

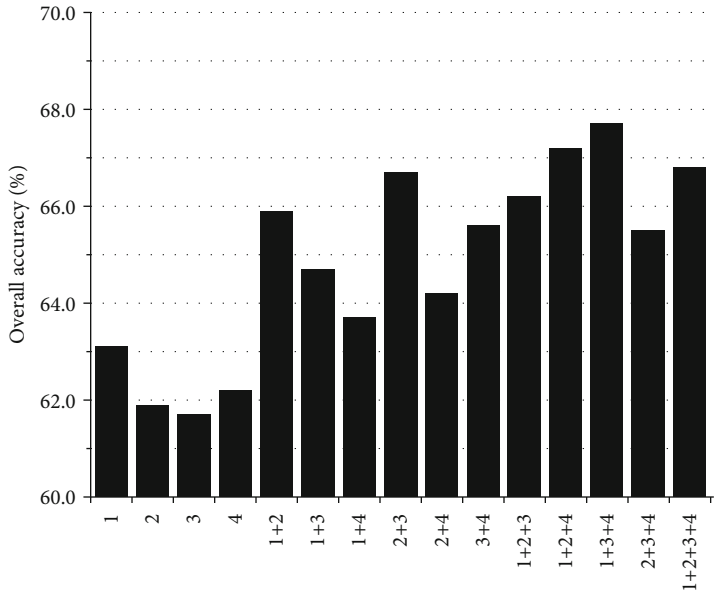

Figure 2: The combination scheme of chemical shifts. The number 1 denotes $^{1}H_{\alpha}$, 2 denotes $^{1}H_{N}$, 3 denotes ^{15}N, and 4 denotes $^{13}C_{\alpha}$.

Table 2: The predictive results based on the different information parameters in the Jackknife test.

Features		Location				OA (%)
		Cell wall	Extracell	Cytoplasm	Cell membrane	
AAC	S_n (%)	13.64	74.30	70.64	85.85	74.6%
	S_p (%)	99.85	84.77	88.62	89.34	
	MCC	0.31	0.58	0.61	0.73	
	ACC (%)	96.71	81.57	82.14	88.29	
DC	S_n (%)	0.00	70.09	70.24	84.91	72.4%
	S_p (%)	99.85	84.57	86.34	88.53	
	MCC	-0.01	0.54	0.57	0.71	
	ACC (%)	96.71	80.14	80.57	88.53	
GO	S_n (%)	0.00	75.23	71.03	66.98	68.9%
	S_p (%)	99.71	86.63	82.14	85.45	
	MCC	-0.01	0.61	0.53	0.52	
	ACC (%)	96.57	83.14	78.14	79.86	
acACS	S_n (%)	0.00	66.36	70.24	73.11	67.7%
	S_p (%)	99.95	83.13	80.36	88.53	
	MCC	-0.01	0.49	0.50	0.62	
	ACC (%)	96.85	78.00	76.71	83.86	
hpDC	S_n (%)	0.00	73.82	76.59	76.42	73.3%
	S_p (%)	99.85	86.01	82.81	91.60	
	MCC	0.07	0.59	0.59	0.69	
	ACC (%)	96.71	82.3	80.57	87.00	

the*k*-fold cross-validation test, and the independent test [8, 47–53]. In this paper, a strict and objective method for the Jackknife test was adopted to examine the performance of the proposed model. The principle of the Jackknife test is to select one from among all protein sequences as a testing set and the other remaining sequences as a training set until all protein sequences are recycled once.

3.2. Evaluation of the Predictive Performances. In order to evaluate the performance of related predictive methods and

Table 3: The predictive results based on the hybrid information in the Jackknife test.

Features		Location				OA (%)
		Cell wall	Extracell	Cytoplasm	Cell membrane	
AAC+GO	S_n (%)	9.09	87.85	76.59	86.79	81.0%
	S_p (%)	99.56	89.10	92.86	90.78	
	MCC	0.18	0.75	0.71	0.76	
	ACC (%)	96.71	88.71	87.00	89.57	
AAC+hpDC	S_n (%)	40.91	83.18	83.33	89.60	83.9%
	S_p (%)	99.26	90.74	91.96	94.50	
	MCC	0.50	0.74	0.78	0.74	
	ACC (%)	97.14	87.26	88.94	89.24	
AAC+GO+acACS	S_n (%)	22.73	85.51	81.35	86.79	82.4%
	S_p (%)	99.56	90.74	91.74	92.21	
	MCC	0.37	0.75	0.74	0.78	
	ACC (%)	97.14	89.14	88.00	90.57	
AAC+DC+hpDC	S_n (%)	40.91	86.92	87.30	88.68	86.1%
	S_p (%)	99.71	91.15	92.86	93.65	
	MCC	0.49	0.77	0.77	0.82	
	ACC (%)	97.57	89.96	89.57	92.14	
AAC+GO+acACS+hpDC	S_n (%)	22.73	83.65	86.51	85.85	83.4%
	S_p (%)	99.56	90.71	90.63	94.67	
	MCC	0.37	0.73	0.77	0.81	
	ACC (%)	97.14	88.57	89.14	92.00	
AAC+DC+GO+hpDC	S_n (%)	36.36	83.18	82.54	91.98	84.1%
	S_p (%)	99.41	88.86	92.86	93.24	
	MCC	0.48	0.74	0.76	0.84	
	ACC (%)	97.43	88.86	89.14	92.86	
AAC+DC+GO+acACS+hpDC	S_n (%)	22.27	84.11	84.13	90.09	84.1%
	S_p (%)	99.71	90.95	92.86	93.24	
	MCC	0.44	0.74	0.78	0.82	
	ACC (%)	97.43	88.86	89.71	92.29	

the reliability of the algorithm, the sensitivity (S_n), specificity (S_p), accuracy (ACC), Matthew's correlation coefficient (MCC), and overall accuracy (OA) [54–59] were used and defined by

$$
\begin{aligned}
S_n &= \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})},\\
S_p &= \frac{\mathrm{TN}}{(\mathrm{TN}+\mathrm{FP})},\\
\mathrm{ACC} &= \frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})},\\
\mathrm{MCC} &= \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FP})},\\
\mathrm{OA} &= \sum_{i=1}^{4}\frac{\mathrm{TP}_i}{N},
\end{aligned}
\tag{9}
$$

where N is the total number of protein sequences in the dataset, TP represents the numbers of the correctly recognized positives, FN is the numbers of the positives recognized as negatives, FP means the numbers of the negatives recognized as positives, while TN is the numbers of correctly recognized negatives.

3.3. The Prediction of Gram-Positive Bacteria. In this paper, in order to investigate the effectiveness of our approaches, we have used five feature extraction strategies and the SVM is used as classification algorithm.

The autocovariance average chemical shift (acACS) vectors were formed based on protein sequence, and in order to obtain the best prediction results, we need to find the best chemically shifted atom combination and the best parameter λ. Figure 1 shows that the predicted results for λ ranges from 0 to 56, and the best λ is 40. Figure 2 shows that the prediction result was the best when the combination mode of chemically shifted atoms was $^{1}\mathrm{H}_{\alpha} + {}^{15}\mathrm{N} + {}^{13}\mathrm{C}_{\alpha}$. For gene ontology information, the first 2573-dimensional vector

Table 4: The results compared with previous methods.

Method	Validation method	OA (%)
Shen's first work[a]	Jackknife test	82.7%
Shen's second work[b]	Jackknife test	82.2%
Hu's work[c]	Jackknife test	85.9%
Julia Rahman's work[d]	8-Fold cross-validation	73.2%
This study	Jackknife test	86.1%

[a]See ref. [8]. [b]See ref. [9]. [c]See ref. [10]. [d]See ref. [11].

was obtained. Since the redundancy of data has a detrimental effect on the prediction results, we used the method of principal component analysis to reduce the vector to 854 dimensions. First of all, the 2573 GO terms were integrated into one vector, then the frequency of each GO term was counted. According to the sum of frequencies, the first 854 data was selected.

The predicted results by the Jackknife test for the different information parameters are recorded in Table 2, and the predicted results based on the combined parameter information with the Jackknife test are shown in Table 3. The results showed that the combined parameters were better than a single characteristic parameter. And the combined parameter AAC+DC+hpDC obtained the best accuracy which was 86.1%. The results indicated that the combined parameter was helpful to predict the protein subcellular location of Gram-positive bacteria. The reason that the accuracies of AAC+GO+acACS+hpDC, AAC+DC+GO+hpDC, and AAC+DC+GO+acACS+hpDC were lower than AAC +DC+hpDC was probably due to the redundancy of data.

4. Discussion

For the purpose of comparing the predictive capability of our method, the predicted results of Shen's, Hu's, and Julia Rahman's method are enumerated in Table 4. It can be seen from Table 4 that our results were superior to others. The accuracy of our method was 3.4% higher than Shen's first work, 3.9% higher than Shen's second work, 0.2% higher than Hu's work, and 12.9% higher than Julia Rahman's work.

Gram-positive bacteria exist widely in nature and could cause many diseases, so studying Gram-positive bacteria subcellular location could solve the many problems of disease. In this paper, the dataset of protein subcellular location of Gram-positive bacteria was reconstructed, and the subcellular location of Gram-positive bacterial protein was predicted. The method in this paper had the advantages of a simple algorithm and an automatic process. The results showed that the combined parameter can improve the prediction accuracy of protein subcellular location of Gram-positive bacteria.

The protein data used to support the findings of this study are included within the supplementary information file.

Data Availability

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

Conflicts of Interest

The authors declare that there is no conflict of interest.

Authors' Contributions

Gao XW conceived the selection of feature parameters and carried out the computation by SVM. Li FM analysed the results and wrote the manuscript. All authors reviewed the manuscript.

Acknowledgments

This work was supported by the Natural Science Foundation of Inner Mongolia of China (2019MS03015) and the National Natural Science Foundation of China (31360206).

Supplementary Materials

The protein data used to support the findings of this study are included within the supplementary information file. *(Supplementary Materials)*

References

[1] Y. Fujiwara and M. Asogawa, "Prediction of subcellular localizations using amino acid composition and order," *Genome informatics*, vol. 12, pp. 103–112, 2001.

[2] M. Suzuki, R. J. Youle, and N. Tjandra, "Structure of Bax: coregulation of dimer formation and intracellular localization," *Cell*, vol. 103, no. 4, pp. 645–654, 2000.

[3] M. L. Liu, W. Su, Z. X. Guan et al., "An overview on predicting protein subchloroplast localization by using machine learning methods," *Current Protein & Peptide Science*, vol. 21, 2020.

[4] H. Ding and D. Li, "Identification of mitochondrial proteins of malaria parasite using analysis of variance," *Amino Acids*, vol. 47, no. 2, pp. 329–333, 2015.

[5] T. H. Zhang and S. W. Zhang, "Advances in the prediction of protein subcellular locations with machine learning," *Current Bioinformatics*, vol. 14, no. 5, pp. 406–421, 2019.

[6] S. Wan, Y. Duan, and Q. Zou, "HPSLPred: an ensemble multilabel classifier for human protein subcellular location prediction with imbalanced source," *Proteomics*, vol. 17, no. 17-18, pp. 17-18, 2017.

[7] L. Wei, Y. Ding, R. Su, J. Tang, and Q. Zou, "Prediction of human protein subcellular localization using deep learning," *Journal of Parallel & Distributed Computing*, vol. 117, pp. 212–217, 2018.

[8] H. B. Shen and K. C. Chou, "Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins," *Protein Engineering, Design & Selection*, vol. 20, no. 1, pp. 39–46, 2007.

[9] H. B. Shen and K. C. Chou, "Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins," *Protein and Peptide Letters*, vol. 16, no. 12, pp. 1478–1484, 2009.

[10] Y. Hu, T. Li, J. Sun et al., "Predicting gram-positive bacterial protein subcellular localization based on localization motifs," *Journal of Theoretical Biology*, vol. 308, pp. 135–140, 2012.

[11] J. Rahman, M. N. I. Mondal, M. K. B. Islam, M. A. M. Hasan, and S. M. S. Amin, "Gram-positive bacterial protein

subcellular localization prediction using features fusion strategy," in *2016 9th International Conference on Electrical and Computer Engineering (ICECE)*, pp. 291–294, 2016.

[12] X. Xiao, X. Cheng, S. Su, Q. Mao, and K. C. Chou, "pLoc-mGpos: incorporate key gene ontology information into general PseAAC for predicting subcellular localization of gram-positive bacterial proteins," *Natural Science*, vol. 9, no. 9, pp. 330–349, 2017.

[13] X. Xiao, X. Cheng, G. Chen, Q. Mao, and K. C. Chou, "pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC," *Genomics*, vol. 111, no. 4, pp. 886–892, 2019.

[14] S. H. Li, J. Zhang, Y. W. Zhao et al., "iPhoPred: a predictor for identifying phosphorylation sites in human protein," *Ieee Access*, vol. 7, pp. 177517–177528, 2019.

[15] W. Yang, X. J. Zhu, J. Huang, H. Ding, and H. Lin, "A brief survey of machine learning methods in protein sub-Golgi localization," *Current Bioinformatics*, vol. 14, no. 3, pp. 234–240, 2019.

[16] J. X. Tan, S. H. Li, Z. M. Zhang et al., "Identification of hormone binding proteins based on machine learning methods," *Mathematical Biosciences and Engineering*, vol. 16, no. 4, pp. 2466–2480, 2019.

[17] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[18] F.-M. Li and X.-Q. Wang, "Identifying anticancer peptides by using improved hybrid compositions," *Scientific Reports*, vol. 6, no. 1, 2016.

[19] G.-L. Fan and Q.-Z. Li, "Predicting protein submitochondria locations by combining different descriptors into the general form of Chou's pseudo amino acid composition," *Amino Acids*, vol. 43, no. 2, pp. 545–555, 2012.

[20] B. Boeckmann, A. Bairoch, R. Apweiler et al., "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," *Nucleic Acids Research*, vol. 31, no. 1, pp. 365–370, 2003.

[21] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.

[22] M. A. Andrade, S. I. O'Donoghue, and B. Rost, "Adaptation of protein surfaces to subcellular location," *Journal of Molecular Biology*, vol. 276, no. 2, pp. 517–525, 1998.

[23] M. Reczko and H. Bohr, "The DEF data base of sequence based protein fold class predictions," *Nucleic Acids Research*, vol. 22, no. 17, pp. 3616–3619, 1994.

[24] H. Tang, Y. W. Zhao, P. Zou et al., "HBPred: a tool to identify growth hormone-binding proteins," *International Journal of Biological Sciences*, vol. 14, no. 8, pp. 957–964, 2018.

[25] W. Chen, F. Nie, and H. Ding, "Recent advances of computational methods for identifying bacteriophage virion proteins," *Protein and Peptide Letters*, vol. 27, no. 4, pp. 259–264, 2020.

[26] A. A. Schaffer, L. Aravind, T. L. Madden et al., "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements," *Nucleic Acids Research*, vol. 29, no. 14, pp. 2994–3005, 2001.

[27] W. Shi, M. Punta, J. Bohon et al., "Characterization of metalloproteins by high-throughput X-ray absorption spectroscopy," *Genome Research*, vol. 21, no. 6, pp. 898–907, 2011.

[28] X. J. Zhu, C. Q. Feng, H. Y. Lai, W. Chen, and L. Hao, "Predicting protein structural classes for low-similarity sequences by evaluating different features," *Knowledge-Based Systems*, vol. 163, pp. 787–793, 2019.

[29] S. P. Mielke and V. V. Krishnan, "Protein structural class identification directly from NMR spectra using averaged chemical shifts∗," *Bioinformatics*, vol. 19, no. 16, pp. 2054–2064, 2003.

[30] G. L. Fan, Y. L. Liu, and Y. C. Zuo, "acACS: improving the prediction accuracy of protein subcellular locations and protein classification by incorporating the average chemical shifts composition," *The Scientific World Journal*, vol. 2014, Article ID 864135, 9 pages, 2014.

[31] Y. L. Chen and Q. Z. Li, "Prediction of the subcellular location of apoptosis proteins," *Journal of Theoretical Biology*, vol. 245, no. 4, pp. 775–783, 2007.

[32] B. Manavalan and J. Lee, "SVMQA: support-vector-machine-based protein single-model quality assessment," *Bioinformatics*, vol. 33, no. 16, pp. 2496–2503, 2017.

[33] B. Manavalan, T. H. Shin, and G. Lee, "PVP-SVM: sequence-based prediction of phage virion proteins using a Support Vector Machine," *Frontiers in Microbiology*, vol. 9, p. 476, 2018.

[34] L. Wei, R. Su, S. Luan et al., "Iterative feature representations improve N4-methylcytosine site prediction," *Bioinformatics*, vol. 35, no. 23, pp. 4930–4937, 2019.

[35] C. Meng, S. Jin, L. Wang, F. Guo, and Q. Zou, "AOPs-SVM: a sequence-based classifier of antioxidant proteins using a Support Vector Machine," *Frontiers in Bioengineering and Biotechnology*, vol. 7, 2019.

[36] H. Bu, J. Hao, J. Guan, and S. Zhou, "Predicting enhancers from multiple cell lines and tissues across different developmental stages based on SVM method," *Current Bioinformatics*, vol. 13, no. 6, pp. 655–660, 2018.

[37] L. Chao, L. Wei, and Q. Zou, "SecProMTB: a SVM-based classifier for secretory proteins of *Mycobacterium tuberculosis* with imbalanced data set," *Proteomics*, vol. 19, 2019.

[38] Y. Wang, F. Shi, L. Cao et al., "Morphological segmentation analysis and texture-based Support Vector Machines classification on mice liver fibrosis microscopic images," *Current Bioinformatics*, vol. 14, no. 4, pp. 282–294, 2019.

[39] H. Yang, W. Yang, F. Y. Dao et al., "A comparison and assessment of computational method for identifying recombination hotspots in *Saccharomyces cerevisiae*," *Briefings in Bioinformatics*, 2019.

[40] H. Y. Lai, Z. Y. Zhang, Z. D. Su et al., "iProEP: a computational predictor for predicting promoter," *Molecular Therapy - Nucleic Acids*, vol. 17, pp. 337–346, 2019.

[41] H. Ding, W. Yang, H. Tang et al., "PHYPred: a tool for identifying bacteriophage enzymes and hydrolases," *Virologica Sinica*, vol. 31, no. 4, pp. 350–352, 2016.

[42] H. Y. Lai, C. Q. Feng, Z. Y. Zhang, H. Tang, W. Chen, and H. Lin, "A brief survey of machine learning application in cancerlectin identification," *Current Gene Therapy*, vol. 18, no. 5, pp. 257–267, 2018.

[43] K. Liu and W. Chen, "iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications," *Bioinformatics*, vol. 36, no. 11, pp. 3336–3342, 2020.

[44] N. Stephenson, E. Shane, J. Chase et al., "Survey of machine learning techniques in drug discovery," *Current Drug Metabolism*, vol. 20, no. 3, pp. 185–193, 2019.

[45] H. Tang, R. Z. Cao, W. Wang, T. S. Liu, L. M. Wang, and C. M. He, "A two-step discriminated method to identify thermophilic proteins," *International Journal of Biomathematics*, vol. 10, no. 4, 2017.

[46] L. Yu, F. Xu, and L. Gao, "Predict new therapeutic drugs for hepatocellular carcinoma based on gene mutation and expression," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 8, 2020.

[47] H. Lin, Z. Y. Liang, H. Tang, and W. Chen, "Identifying Sigma70 promoters with novel pseudo nucleotide composition," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1316–1321, 2019.

[48] W. Chen, P. Feng, T. Liu, and D. Jin, "Recent advances in machine learning methods for predicting heat shock proteins," *Current Drug Metabolism*, vol. 20, no. 3, pp. 224–228, 2019.

[49] S. Basith, B. Manavalan, T. Hwan Shin, and G. Lee, "Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening," *Medicinal Research Reviews*, vol. 40, no. 4, pp. 1276–1314, 2020.

[50] V. Boopathi, S. Subramaniyam, A. Malik, G. Lee, B. Manavalan, and D. C. Yang, "mACPpred: a Support Vector Machine-based meta-predictor for identification of anticancer peptides," *International Journal of Molecular Sciences*, vol. 20, no. 8, 2019.

[51] M. M. Hasan, N. Schaduangrat, S. Basith, G. Lee, W. Shoombuatong, and B. Manavalan, "HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation," *Bioinformatics*, vol. 36, no. 11, pp. 3350–3356, 2020.

[52] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation," *Bioinformatics*, vol. 35, no. 16, pp. 2757–2765, 2019.

[53] L. Yu, S. Y. Yao, L. Gao, and Y. H. 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, 2019.

[54] F. Y. Dao, H. Lv, H. Zulfiqar et al., "A computational platform to identify origins of replication sites in eukaryotes," *Briefings in Bioinformatics*, 2020.

[55] S. Basith, B. Manavalan, T. H. Shin, and G. Lee, "SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome," *Molecular Therapy–Nucleic Acids*, vol. 18, pp. 131–141, 2019.

[56] B. Manavalan, S. Basith, T. H. Shin, D. Y. Lee, L. Wei, and G. Lee, "4mCpred-EL: an ensemble learning framework for identification of DNA N(4)-methylcytosine sites in the mouse genome," *Cell*, vol. 8, pp. 1–14, 2019.

[57] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees," *Computational and Structural Biotechnology Journal*, vol. 17, pp. 972–981, 2019.

[58] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation," *Molecular Therapy - Nucleic Acids*, vol. 16, pp. 733–744, 2019.

[59] L. Yu and L. Gao, "Human pathway-based disease network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 4, pp. 1240–1249, 2019.