Hindawi
Computational and Mathematical Methods in Medicine
Volume 2022, Article ID 7493834, 7 pages
https://doi.org/10.1155/2022/7493834

Research Article

Identification of *Helicobacter pylori* Membrane Proteins Using Sequence-Based Features

Mujiexin Liu,[1] **Hui Chen**,[2] **Dong Gao**,[3] **Cai-Yi Ma**,[3] **and Zhao-Yue Zhang**[2,3]

[1]*Ineye Hospital of Chengdu University of TCM, Chengdu University of TCM, Chengdu 610084, China*
[2]*School of Healthcare Technology, Chengdu Neusoft University, 611844 Chengdu, China*
[3]*School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China*

Correspondence should be addressed to Zhao-Yue Zhang; zyzhang@uestc.edu.cn

Received 7 November 2021; Accepted 16 December 2021; Published 12 January 2022

Academic Editor: Balachandran Manavalan

Helicobacter pylori (*H. pylori*) is the most common risk factor for gastric cancer worldwide. The membrane proteins of the *H. pylori* are involved in bacterial adherence and play a vital role in the field of drug discovery. Thus, an accurate and cost-effective computational model is needed to predict the uncharacterized membrane proteins of *H. pylori*. In this study, a reliable benchmark dataset consisted of 114 membrane and 219 nonmembrane proteins was constructed based on UniProt. A support vector machine- (SVM-) based model was developed for discriminating *H. pylori* membrane proteins from nonmembrane proteins by using sequence information. Cross-validation showed that our method achieved good performance with an accuracy of 91.29%. It is anticipated that the proposed model will be useful for the annotation of *H. pylori* membrane proteins and the development of new anti-*H. pylori* agents.

1. Introduction

Helicobacter pylori (*H. pylori*) is a Gram-negative spiral-shaped bacterium that infects half of the human population worldwide. *H. pylori* causes gastric mucosa damage, chronic inflammation, and dysregulation of the gut community, increasing the risk of gastric cancer [1–3]. Attachment to the gastric mucosa is the first step in establishing bacterial colonization [4]. *H. pylori* membrane proteins such as antigen-binding adhesin (BabA), sialic acid-binding adhesin (SabA), outer inflammatory protein (OipA), and outer membrane protein Q (HopQ) can act as putative virulence factors that mediate the host-pathogen interactions, induce the release of inflammatory cytokines, and enhance the virulence property of the bacterium [4–6]. Thus, the identification of *H. pylori* membrane protein receptors contributes to the design of therapeutic drugs and vaccine development [7, 8].

Although *H. pylori* membrane proteins play a key role in attachment to and entry into host cells, only few have been described so far. There are some efforts in the prediction of membrane proteins [9, 10] for other germs like *Mycobacterial* [11] and *Chlamydiae* [12]. However, there are no machine learning-based approaches for the prediction of the *H. pylori* membrane proteins. In this study, we developed a comprehensive in silico approach for discriminating novel *H. pylori* membrane proteins using amino acid sequence-based criteria. First, the benchmark dataset was constructed based on a reliable source. Second, sequence-based feature encoding methods were used to represent protein sequences. Next, the incremental feature selection (IFS) technique with multiple feature ranking methods was applied to obtain the optimal feature set. Finally, a membrane protein prediction model was established based on the optimal feature set. The workflow can be seen in Figure 1.

2. Materials and Methods

2.1. Benchmark Dataset. An objective and strict benchmark dataset is fundamental for a robust prediction model construction [13–18]. The Universal Protein Resource (UniProt) [19] is a comprehensive resource for proteins and can be freely accessed at https://www.uniprot.org/. The

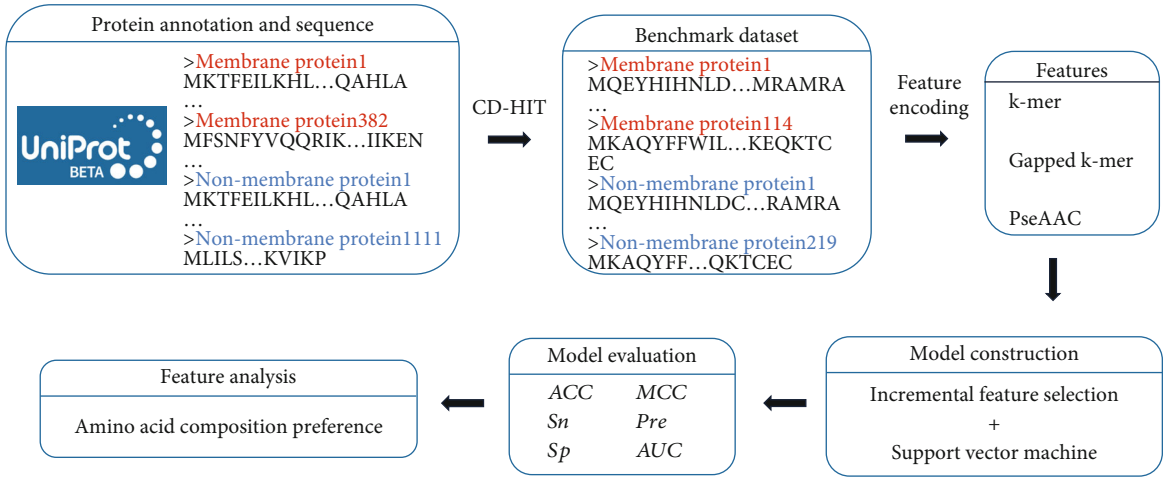

Figure 1: The workflow diagram of developing the *H. pylori* membrane protein prediction model.

382 *H. pylori* membrane protein sequences and 1111 nonmembrane protein sequences were obtained from the UniProt. If a sequence contains nonstandard letters, the sequence was removed from the dataset. To avoid the influence of sequence similarity [20], CD-HIT [21] with 0.3 sequence identity was used to exclude highly similar membrane proteins. Finally, 114 (29.8% of the original) membrane proteins and 219 (19.7% of the original) nonmembrane proteins remained in the benchmark dataset.

2.2. Feature Encoding. Generally, feature encoding plays a crucial role for machine learning in model construction [22–28]. The feature encoding method determines the degree of sequence information mining. In this work, k-mer amino acid composition [29–31], gapped k-mer method [32], and pseudo-amino acid composition (PseAAC) [33–39] were used to formulate sequences.

Let the protein $\mathbf{S}$ be expressed as follows:

$$\mathbf{S} = R_1 R_2 R_3 R_4 R_5 \cdots R_i R_{i+1} \cdots R_L, \tag{1}$$

where L denotes the length of the protein sequence and R_i is the i-th amino acid.

By using k-mer amino acid composition, a primary protein sequence $\mathbf{S}$ can be transferred into a vector $\mathbf{V}_k$ with 20^k elements according to the following formula:

$$\mathbf{V}_k = \left[f_1^{k-\text{mer}} \; f_2^{k-\text{mer}} \bullet\bullet\bullet f_i^{k-\text{mer}} \bullet\bullet\bullet f_{20^k}^{k-\text{mer}} \right]^{\mathbf{T}}, \tag{2}$$

where the symbol $\mathbf{T}$ means the transposition of a vector and $f_i^{k-\text{mer}}$ is the normalized frequency of the i-th k-mer amino acid component occurring in $\mathbf{S}$ and can be calculated by

$$f_i^{k-\text{mer}} = \frac{n_i}{\sum_{i=1}^{20^k} n_i} = \frac{n_i}{L-k+1}, \tag{3}$$

where n_i means the number of occurrences of the i-th k-mer amino acid component in the sequence $\mathbf{S}$.

With the increase of k, one protein sequence may have many k-mers absent, and its feature vector will contain a large number of zero values. To overcome this sparse problem, gapped k-mer (k-mer with g gap) was used. For example, "GG" with 3 gaps constitute the patterns "GNNNG," where N represent any kind of amino acid. By using the gapped k-mer method, a primary protein sequence $\mathbf{S}$ can be transferred into a vector $\mathbf{V}_g$ with 20^{k-g} elements according to the following formula:

$$\mathbf{V}_g = \left[f_1^{gk-\text{mer}} \; f_2^{gk-\text{mer}} \bullet\bullet\bullet f_i^{gk-\text{mer}} \bullet\bullet\bullet f_{20^{k-g}}^{gk-\text{mer}} \right]^{\mathbf{T}}, \tag{4}$$

where the $f_i^{gk-\text{mer}}$ is the normalized frequency of the i-th k-mer with g gap amino acid component occurring in $\mathbf{S}$.

PseAAC can represent a protein sequence in a discrete model without completely losing its sequence-order information. A primary protein sequence $\mathbf{S}$ can be transferred into a vector $\mathbf{V}_p$ with PseAAC according to the following formula:

$$\mathbf{V}_p = [x_1 \cdots x_{20} \, x_{20+1} \cdots x_{20+\lambda}]^{\mathbf{T}}, \tag{5}$$

$$x_i = \begin{cases} \dfrac{f_i}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \Theta_j}, & 1 \le i \le 20, \\ \dfrac{\omega \Theta i - 20}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \Theta_j}, & 20+1 \le i \le 20+\lambda, \end{cases} \tag{6}$$

where f_i is the normalized frequency of i-th amino acid, and Θ_j is the j-th sequence correlation factor that can be calculated by the product of the six physicochemical property numerical values between amino acids at different positions. ω is the weight factor for short range and long range.

2.3. Feature Selection and Modeling. To exclude noise and improve computational efficiency, feature selection is an indispensable step [23, 40–45]. Binomial distribution is one of the wonderful feature selection techniques that have been successfully applied in many works [46–48]. The high binomial distribution score indicates that the presence of the k

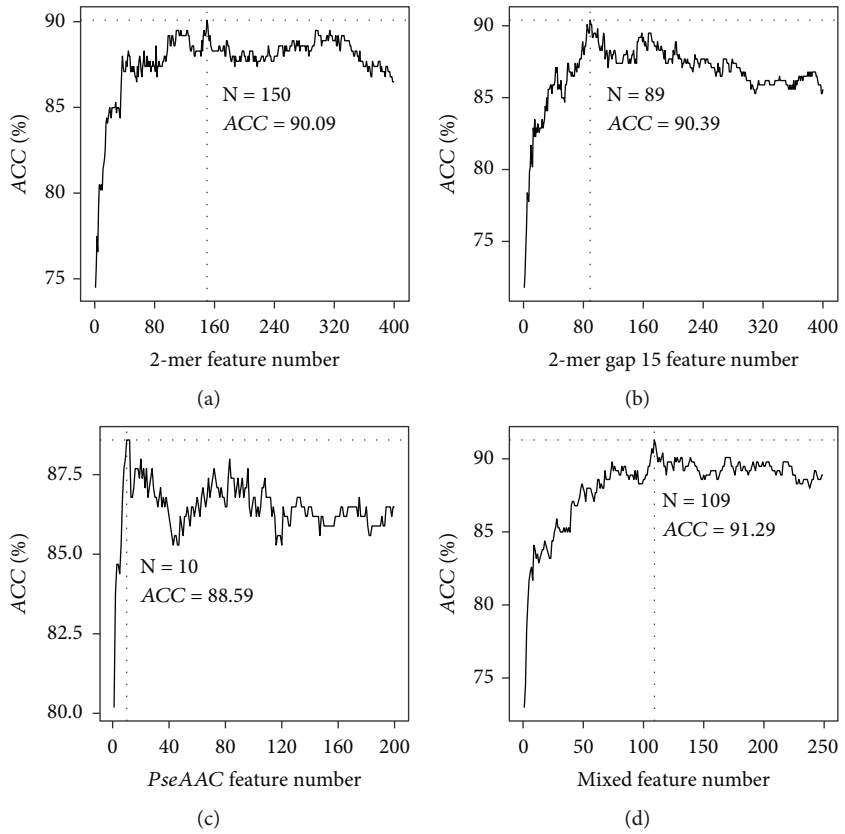

Figure 2: The IFS curves for (a) 2-mer features, (b) gapped 2-mer features, (c) PseAAC features, and (d) merged features.

-mer amino acid in a membrane protein sequence is not accidental. Analysis of variance (ANOVA) tests the ratio of the variance between groups and the variance within the groups to analyse the differences among group means [30]. The high ANOVA score means there is a big feature difference between the membrane protein group and the non-membrane protein group. In this study, binomial distribution was used on k-mer features, and ANOVA was used on gapped k-mer and PseAAC features to winnow out the irrelevant features. Then, ANOVA was used to reprune all the redundant features.

After ranking the features according to their statistical scores, the IFS strategy with support vector machine (SVM) was adopted to determine the optimal feature set [49–53]. SVM is a classification algorithm that finds the optimal classification hyperplane in the high-dimensional feature space. The IFS strategy added features one by one to the feature set from a higher-ranked to a lower-ranked score. Once a new feature set was composed, LIBSVM [54] with 5-fold cross-validation was performed to train and test prediction models. The optimal feature set is defined based on the principle that the prediction model based on such features could achieve maximum accuracy. Finally, an SVM model was constructed based on the optimal feature subset for the membrane protein prediction.

2.4. Performance Evaluation Metrics. In order to assess the capability of the binary prediction method, six indexes, namely, accuracy (ACC), sensitivity (Sn), specificity (Sp),

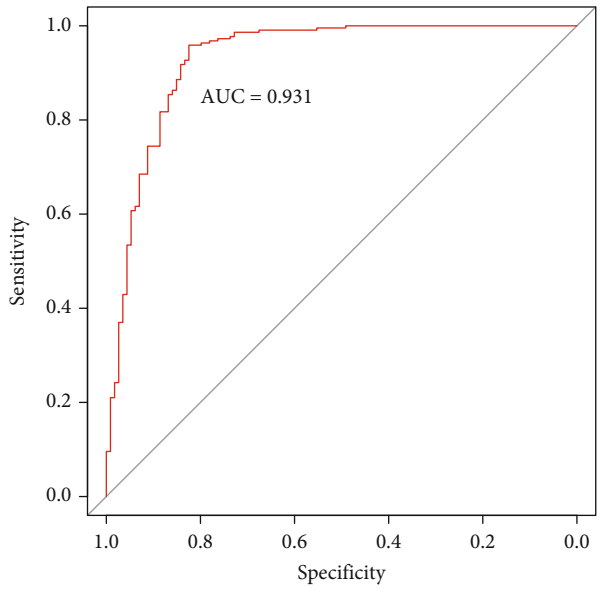

Figure 3: The ROC curves of the 5-fold cross-validation test.

precision (Pre), Matthew's correlation coefficient (MCC), and the area under the receiver operating characteristic curve (AUC) [55–60], were used and formulated as

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \quad (7)$$

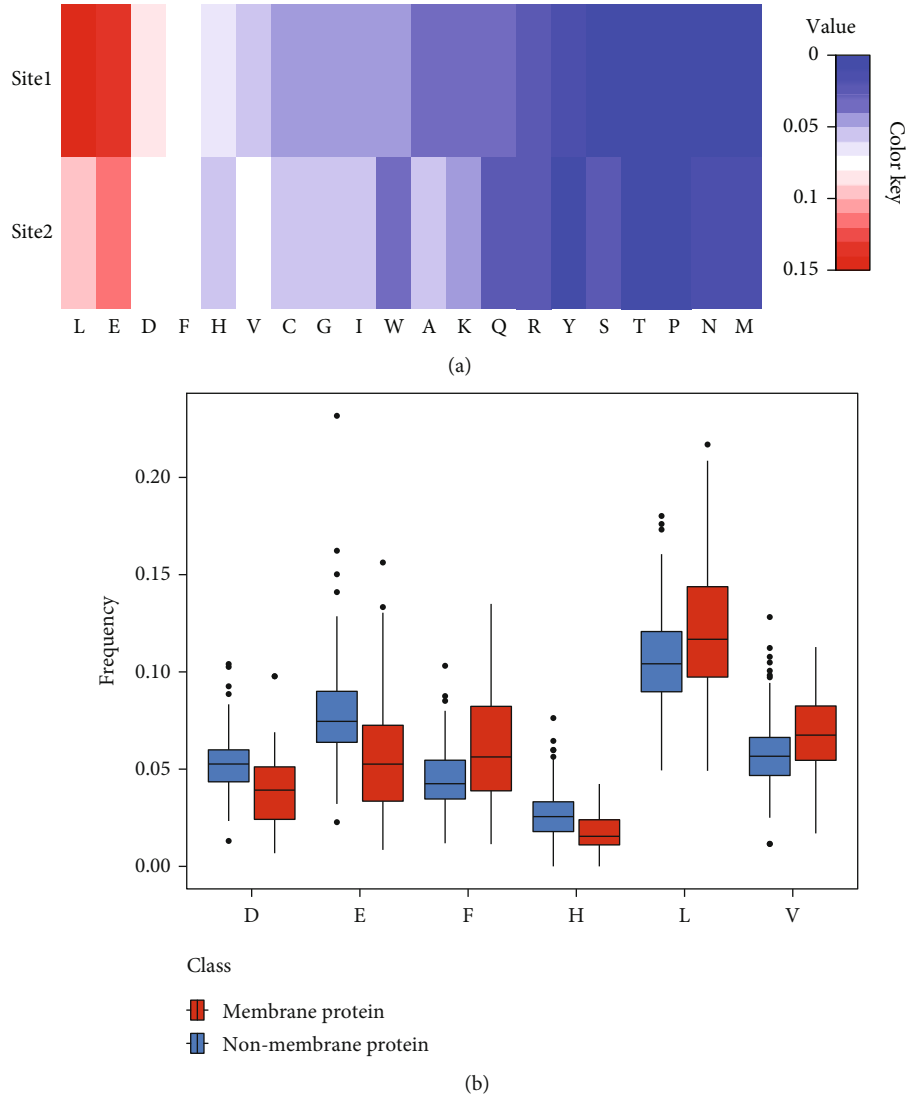

Figure 4: (a) The heat map of AAC of the model features. (b) The frequency of the six amino acids in the two classes.

$$Sn = \frac{TP}{TP + FN}, \tag{8}$$

$$Sp = \frac{TN}{TN + FP}, \tag{9}$$

$$Pre = \frac{TP}{TP + FP}, \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}, \tag{11}$$

where TP (true positive) and TN (true negative) present the numbers of correctly identified membrane proteins and nonmembrane proteins, respectively. FP (false positive) and FN (false negative) denote the number of nonmembrane proteins incorrectly classified as membrane proteins and the number of membrane proteins incorrectly classified as nonmembrane proteins, respectively. Receiver operating characteristics (ROC) analysis was used to measure the performance of the model with the varying decision thresholds [61–63]. Due to the small sample size, the result of the 5-fold cross-validation was used to evaluate the model performance.

3. Results and Discussion

3.1. Feature Optimization. As shown in equations (3), (4), and (5), the description of the protein sequences depends on parameters k, g, ω, and λ. For k-mer feature encoding, $k = 2, 3, 4$ was tried in this study. The model achieved the best accuracy of 90.09% with the top 150 binomial distribution-ranked 2-mer features (Figure 2(a)). For gapped k-mer feature encoding, we set $k = 2$ and traverse g from 1 to 20, when $g = 15$, and the model achieved the best accuracy of 90.39% with the top 89 ANOVA-ranked features (Figure 2(b)). For PseAAC, we set the weight factor $\omega = 0.5$ and parameter λ from 1 to 70 with step size 5, and the best performance achieved was 88.59% when the λ is 20 and feature number is 10 (Figure 2(c)). To represent the sequence

information comprehensively, all best feature subsets were merged and ranked by ANOVA. IFS was performed again to filter out the redundant features. As we can see in Figure 2(d), the model achieved the best accuracy of 91.29% when the top 109 ANOVA-ranked features were used to train the model.

3.2. Model Construction and Evaluation. Finally, 109 features were used to construct the SVM-based model for the prediction of membrane proteins. And the soft margin SVM penalty coefficient c and Gaussian kernel function width parameter γ are 0.5.

To show the prediction capability of the final model, six evaluation metrics were calculated based on the result of the 5-fold cross-validation. The model achieved the *ACC* of 91.29%, *Sn* of 82.46%, *Sp* of 95.9%, *Pre* of 91.26%, and *MCC* of 0.804. We also drew the ROC curve in Figure 3. It shows that the AUC reaches the value of 0.931, suggesting that the proposed model has an excellent prediction capability on membrane protein classification.

3.3. Amino Acid Composition (AAC) of Optimal Features. The AAC of the model features was used to analyse the preference of membrane proteins for specific amino acids. Among the optimal feature set, there are 83 2-mer features, 16 gapped 2-mer features, and 10 PseAAC features. Focusing on the 2-mer and gapped 2-mer features, we found that the occurrence of leucine (L), glutamic acid (E), aspartic acid (D), phenylalanine (F), valine (V), and histidine (H) exceeds 50% of the total (Figure 4(a)). And the frequencies of F, L, and V in membrane protein sequences are significantly higher than those in nonmembrane protein sequences ($p < 0.001$). In contrast, the frequencies of D, E, and H in nonmembrane protein sequences are significantly higher than those in membrane proteins ($p < 0.001$) (Figure 4(b)).

4. Conclusions

H. pylori membrane proteins are an important class of molecules that play key roles in host-pathogen interactions. However, it is a new area in the prediction of *H. pylori* membrane proteins with machine learning methods. Hence, we developed an *H. pylori* membrane proteins predictor on the basis of sequence-based information. The model will powerfully support the discovery of *H. pylori* membrane proteins and the research of *H. pylori* infection. It has the potential to be significant in novel vaccine candidate antigens and drug development [64, 65]. In the future, we will stay focused on the *H. pylori* membrane protein prediction issues and screen the possible vaccine candidates and drug targets. Moreover, we will collect more data to train a deep learning model [66–71] to improve prediction performance.

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 regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62102067).

References

[1] Z. Li, T. Zhang, H. Lei et al., "Research on gastric cancer's drug-resistant gene regulatory network model," *Current Bioinformatics*, vol. 15, no. 3, pp. 225–234, 2020.

[2] 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 Research*, vol. 48, no. D1, pp. D554–D560, 2020.

[3] L. Cheng, C. Qi, H. Yang et al., "gutMGene: a comprehensive database for target genes of gut microbes and microbial metabolites," *Nucleic Acids Research*, 2021.

[4] Y. Matsuo, Y. Kido, and Y. Yamaoka, "Helicobacter pylori outer membrane protein-related pathogenesis," *Toxins (Basel)*, vol. 9, no. 3, p. 101, 2017.

[5] S. Ansari, E. T. Kabamba, P. K. Shrestha et al., "Helicobacter pylori Bab characterization in clinical isolates from Bhutan, Myanmar, Nepal and Bangladesh," *PLoS One*, vol. 12, no. 11, article e0187225, 2017.

[6] M. Suganuma, M. Kurusu, S. Okabe et al., "Helicobacter pylori membrane protein 1: a new carcinogenic factor of Helicobacter pylori," *Cancer Research*, vol. 61, no. 17, pp. 6356–6359, 2001.

[7] Y. Yamaoka, O. Ojo, S. Fujimoto et al., "Helicobacter pylori outer membrane proteins and gastroduodenal disease," *Gut*, vol. 55, no. 6, pp. 775–781, 2006.

[8] L. Yu, M. Xia, and Q. An, "A network embedding framework based on integrating multiplex network for drug combination prediction," *Briefings in Bioinformatics*, 2021.

[9] M. Kabir, M. Arif, F. Ali, S. Ahmad, Z. N. K. Swati, and D. J. Yu, "Prediction of membrane protein types by exploring local discriminative information from evolutionary profiles," *Analytical Biochemistry*, vol. 564-565, pp. 123–132, 2019.

[10] Y. C. Zuo, W. X. Su, S. H. Zhang et al., "Discrimination of membrane transporter protein types using K-nearest neighbor method derived from the similarity distance of total diversity measure," *Molecular BioSystems*, vol. 11, no. 3, pp. 950–957, 2015.

[11] C. Ding, L. F. Yuan, S. H. Guo, H. Lin, and W. Chen, "Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions," *Journal of Proteomics*, vol. 77, pp. 321–328, 2012.

[12] E. Heinz, P. Tischler, T. Rattei, G. Myers, M. Wagner, and M. Horn, "Comprehensive in silico prediction and analysis of chlamydial outer membrane proteins reflects evolution and life style of the Chlamydiae," *BMC Genomics*, vol. 10, p. 634, 2009.

[13] D. Zhang, H.-D. Chen, H. Zulfiqar et al., "iBLP: an XGBoost-based predictor for identifying bioluminescent proteins," *Computational and Mathematical Methods in Medicine*, vol. 2021, 2021.

[14] W. Su, M. L. Liu, Y. H. Yang et al., "PPD: a manually curated database for experimentally verified prokaryotic promoters," *Journal of Molecular Biology*, vol. 433, no. 11, article ???, 2021.

[15] 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, 2018.

[16] L. Wei, W. He, A. Malik, R. Su, L. Cui, and B. Manavalan, "Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework," *Briefings in Bioinformatics*, vol. 22, no. 4, 2021.

[17] M. M. Hasan, M. A. Alam, W. Shoombuatong, H. W. Deng, B. Manavalan, and H. Kurata, "NeuroPred-FRL: an interpretable prediction model for identifying neuropeptide using feature representation learning," *Briefings in Bioinformatics*, vol. 22, no. 6, 2021.

[18] P. Charoenkwan, C. Nantasenamat, M. M. Hasan, B. Manavalan, and W. Shoombuatong, "Bert4bitter: a bidirectional encoder representations from transformers (Bert)-based model for improving the prediction of bitter peptides," *Bioinformatics*, vol. 37, no. 17, pp. 2556–2562, 2021.

[19] C. UniProt, "UniProt: the universal protein knowledgebase in 2021," *Nucleic Acids Research*, vol. 49, no. D1, pp. D480–D489, 2021.

[20] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: an empirical study," *Briefings in Bioinformatics*, vol. 21, no. 1, pp. 1–10, 2020.

[21] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "Cd-Hit: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[22] H. Zulfiqar, Z. J. Sun, Q. L. Huang et al., "Deep-4mCW2V: a sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli," *Methods*, 2021.

[23] D. Zhang, Z. C. Xu, W. Su et al., "PROBselect: accurate prediction of protein-binding residues from proteins sequences via dynamic predictor selection," *Bioinformatics*, vol. 36, Supplement_2, pp. i735–i744, 2020.

[24] H. Yang, Y. Luo, X. Ren et al., "Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators," *Information Fusion*, vol. 75, pp. 140–149, 2021.

[25] J. Long, H. Yang, Z. Yang et al., "Integrated biomarker profiling of the metabolome associated with impaired fasting glucose and type 2 diabetes mellitus in large-scale Chinese patients," *Clinical and Translational Medicine*, vol. 11, no. 6, article e432, 2021.

[26] H. Lv, F. Y. Dao, Z. X. Guan, H. Yang, Y. W. Li, and H. Lin, "Landscape of cancer diagnostic biomarkers from specifically expressed genes," *Briefings in Bioinformatics*, vol. 21, no. 6, pp. 2175–2184, 2020.

[27] L. Yu, M. Wang, Y. Yang et al., "Predicting therapeutic drugs for hepatocellular carcinoma based on tissue-specific pathways," *PLoS Computational Biology*, vol. 17, no. 2, article e1008696, 2021.

[28] X. G. Chen, W. W. Shi, and L. Deng, "Prediction of disease comorbidity using HeteSim scores based on multiple heterogeneous networks," *Current Gene Therapy*, vol. 19, no. 4, pp. 232–241, 2019.

[29] M. L. Liu, W. Su, J. S. Wang, Y. H. Yang, H. Yang, and H. Lin, "Predicting preference of transcription factors for methylated DNA using sequence information," *Mol Ther Nucleic Acids*, vol. 22, pp. 1043–1050, 2020.

[30] 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.

[31] Y. Zuo, Y. Li, Y. Chen, G. Li, Z. Yan, and L. Yang, "PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition," *Bioinformatics*, vol. 33, no. 1, pp. 122–124, 2017.

[32] 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.

[33] L. Zheng, D. Liu, W. Yang, L. Yang, and Y. Zuo, "Location deviations of DNA functional elements affected SNP mapping in the published databases and references," *Briefings in Bioinformatics*, vol. 21, no. 4, pp. 1293–1301, 2020.

[34] L. Zheng, S. Huang, N. Mu et al., "RAACBook: a web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule," *Database: The Journal of Biological Databases and Curation*, vol. 2019, 2019.

[35] Y. Y. Cao, C. L. Yu, S. H. Huang, S. Y. Wang, Y. C. Zuo, and L. Yang, "Characterization and prediction of presynaptic and postsynaptic neurotoxins based on reduced amino acids and biological properties," *Current Bioinformatics*, vol. 16, no. 3, pp. 364–370, 2021.

[36] H. B. Shen and K. C. Chou, "PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition," *Analytical Biochemistry*, vol. 373, no. 2, pp. 386–388, 2008.

[37] S. Naseer, W. Hussain, Y. D. Khan, and N. Rasool, "Sequence-based identification of arginine amidation sites in proteins using deep representations of proteins and PseAAC," *Current Bioinformatics*, vol. 15, no. 8, pp. 937–948, 2021.

[38] M. A. M. Hasan, M. K. Ben Islam, J. Rahman, and S. Ahmad, "Citrullination site prediction by incorporating sequence coupled effects into PseAAC and resolving data imbalance issue," *Current Bioinformatics*, vol. 15, no. 3, pp. 235–245, 2020.

[39] S. Amanat, A. Ashraf, W. Hussain, N. Rasool, and Y. D. Khan, "Identification of lysine carboxylation sites in proteins by integrating statistical moments and position relative features via general PseAAC," *Current Bioinformatics*, vol. 15, no. 5, pp. 396–407, 2020.

[40] X. Han, Q. Kong, C. Liu, L. Cheng, and J. Han, "Subtypedrug: a software package for prioritization of candidate cancer subtype-specific drugs," *Bioinformatics*, vol. 37, no. 16, pp. 2491–2493, 2021.

[41] Y. Sheng, Y. Jiang, Y. Yang et al., "Selecting gene features for unsupervised analysis of single-cell gene expression data," *Briefings in Bioinformatics*, vol. 22, no. 6, 2021.

[42] 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, pp. 234–240, 2019.

[43] S. He, F. Guo, Q. Zou, and H. Ding, "MRMD2.0: a Python tool for machine learning with feature ranking and reduction," *Current Bioinformatics*, vol. 15, no. 10, pp. 1213–1221, 2021.

[44] X. Wu and L. Yu, *EPSOL: sequence-based protein solubility prediction using multidimensional embedding*, Bioinformatics, Oxford, England, 2021.

[45] J. W. Li, X. Y. Wang, N. Li et al., "Feasibility of mesenchymal stem cell therapy for Covid-19: a mini review," *Current Gene Therapy*, vol. 20, no. 4, pp. 285–288, 2020.

[46] Z. Y. Zhang, Y. H. Yang, H. Ding, D. Wang, W. Chen, and H. Lin, "Design powerful predictor for mRNA subcellular location prediction in Homo sapiens," *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 526–535, 2021.

[47] C. Q. Feng, Z. Y. Zhang, X. J. Zhu et al., "iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators," *Bioinformatics*, vol. 35, no. 9, pp. 1469–1477, 2019.

[48] H. Wang, P. Liang, L. Zheng, C. Long, H. Li, and Y. Zuo, "Correction to: ncDLRES: a novel method for non-coding RNAs family prediction based on dynamic LSTM and ResNet," *Bioinformatics*, vol. 22, no. 1, 2021.

[49] F. Y. Dao, H. Lv, F. Wang et al., "Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique," *Bioinformatics*, vol. 35, no. 12, pp. 2075–2083, 2019.

[50] C. Ao, L. Yu, and Q. Zou, "Prediction of bio-sequence modifications and the associations with diseases," *Briefings in Functional Genomics*, vol. 20, no. 1, pp. 1–18, 2021.

[51] S. Basith, G. Lee, and B. Manavalan, "Stallion: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction," *Briefings in Bioinformatics*, 2021.

[52] S. Basith, M. M. Hasan, G. Lee, L. Wei, and B. Manavalan, "Integrative machine learning framework for the identification of cell-specific enhancers from the human genome," *Briefings in Bioinformatics*, vol. 22, no. 6, 2021.

[53] 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.

[54] C. C. Chang and C. J. Lin, "LIBSVM," *ACM transactions on intelligent systems and technology*, vol. 2, no. 3, pp. 1–27, 2011.

[55] 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.

[56] 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, p. 1750050, 2017.

[57] L. Cheng, H. Shi, Z. Wang et al., "IntNetLncSim: an integrative network analysis method to infer human lncRNA functional similarity," *Oncotarget*, vol. 7, no. 30, pp. 47864–47874, 2016.

[58] F. Mo, Y. Luo, D. A. Fan et al., "Integrated analysis of mRNA-seq and miRNA-seq to identify c-MYC, YAP1 and miR-3960 as major players in the anticancer effects of caffeic acid phenethyl ester in human small cell lung cancer cell line," *Current Gene Therapy*, vol. 20, no. 1, pp. 15–24, 2020.

[59] R. G. Govindaraj, S. Subramaniyam, and B. Manavalan, "Extremely-randomized-tree-based prediction of N(6)-methyladenosine sites in saccharomyces cerevisiae," *Current Genomics*, vol. 21, no. 1, pp. 26–33, 2020.

[60] 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.

[61] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. 8, no. 4, pp. 283–298, 1978.

[62] H. Lv, F. Y. Dao, H. Zulfiqar, and H. Lin, "DeepIPs: comprehensive assessment and computational identification of phosphorylation sites of Sars-Cov-2 infection using a deep learning-based approach," *Briefings in Bioinformatics*, vol. 22, no. 6, 2021.

[63] Q. An and L. Yu, "A heterogeneous network embedding framework for predicting similarity-based drug-target interactions," *Briefings in Bioinformatics*, vol. 22, no. 6, 2021.

[64] D. Liu, G. Li, and Y. Zuo, "Function determinants of Tet proteins: the arrangements of sequence motifs with specific codes," *Briefings in Bioinformatics*, vol. 20, no. 5, pp. 1826–1835, 2019.

[65] B. F. Xu, D. Y. Liu, Z. R. Wang, R. X. Tian, and Y. C. Zuo, "Multi-substrate selectivity based on key loops and non-homologous domains: new insight into ALKBH family," *Cellular and Molecular Life Sciences*, vol. 78, no. 1, pp. 129–141, 2021.

[66] D. Wang, Z. Zhang, Y. Jiang et al., "DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism," *Nucleic Acids Research*, vol. 49, no. 8, article e46, 2021.

[67] F. Y. Dao, H. Lv, W. Su, Z. J. Sun, Q. L. Huang, and H. Lin, "iDHS-Deep: an integrated tool for predicting DNase I hypersensitive sites by deep neural network," *Briefings in Bioinformatics*, vol. 22, no. 5, 2021.

[68] Y. Zhang, J. Yan, S. Chen et al., "Review of the applications of deep learning in bioinformatics," *Current Bioinformatics*, vol. 15, no. 8, pp. 898–911, 2020.

[69] F. Cui, Z. Zhang, and Q. Zou, "Sequence representation approaches for sequence-based protein prediction tasks that use deep learning," *Briefings in Functional Genomics*, vol. 20, no. 1, pp. 61–73, 2021.

[70] X. Peng, L. Chen, and J.-P. Zhou, "Identification of carcinogenic chemicals with network embedding and deep learning methods," *Current Bioinformatics*, vol. 15, no. 9, pp. 1017–1026, 2021.

[71] Z. B. Lv, C. Y. Ao, and Q. Zou, "Protein function prediction: from traditional classifier to deep learning," *Proteomics*, vol. 19, no. 14, p. 2, 2019.