frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 06 January 2021
doi: 10.3389/fbioe.2020.627335

Predicting Cell Wall Lytic Enzymes Using Combined Features

*Xiao-Yang Jing and Feng-Min Li**

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

Due to the overuse of antibiotics, people are worried that existing antibiotics will become ineffective against pathogens with the rapid rise of antibiotic-resistant strains. The use of cell wall lytic enzymes to destroy bacteria has become a viable alternative to avoid the crisis of antimicrobial resistance. In this paper, an improved method for cell wall lytic enzymes prediction was proposed and the amino acid composition (AAC), the dipeptide composition (DC), the position-specific score matrix auto-covariance (PSSM-AC), and the auto-covariance average chemical shift (acACS) were selected to predict the cell wall lytic enzymes with support vector machine (SVM). In order to overcome the imbalanced data classification problems and remove redundant or irrelevant features, the synthetic minority over-sampling technique (SMOTE) was used to balance the dataset. The F-score was used to select features. The S_n, S_p, MCC, and Acc were 99.35%, 99.02%, 0.98, and 99.19% with jackknife test using the optimized combination feature AAC+DC+acACS+PSSM-AC. The S_n, S_p, MCC, and Acc of cell wall lytic enzymes in our predictive model were higher than those in existing methods. This improved method may be helpful for protein function prediction.

Keywords: cell wall lytic enzymes, optimized combination feature, synthetic minority over-sampling technique, F-score, support vector machine, jackknife test

OPEN ACCESS

Edited by:
Zhibin Lv,
University of Electronic Science and Technology of China, China

Reviewed by:
Fei Guo,
Tianjin University, China
Hua Tang,
Southwest Medical University, China

****Correspondence:***
Feng-Min Li
fmli@imau.edu.cn

Specialty section:
This article was submitted to Synthetic Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *09 November 2020*
Accepted: *04 December 2020*
Published: *06 January 2021*

Citation:
Jing X-Y and Li F-M (2021) Predicting Cell Wall Lytic Enzymes Using Combined Features. Front. Bioeng. Biotechnol. 8:627335. doi: 10.3389/fbioe.2020.627335

INTRODUCTION

Bacteria are constantly around us, and bacterial infections have become a major public health problem. The overuse of antibiotics leads to the rapid rise of antibiotic-resistant strains, and people are worried that existing antibiotics will become ineffective against pathogens. Using cell wall lytic enzymes to destroy bacteria has become a viable alternative method to avoid the crisis of antimicrobial resistance (Sommer et al., 2017; Wu et al., 2017; Bhagwat et al., 2019; Cheng et al., 2020). Cell wall lytic enzymes are divided into two enzymes: endolysin and autolysin. Endolysins are phage-encoded enzymes that have evolved to degrade the bacterial cell wall (Shavrina et al., 2016). Many studies have shown that endolysin has an excellent bactericidal effect on *Staphylococcus aureus* (Ajuebor et al., 2016), *Escherichia coli* (Yan et al., 2019), *Streptococcus suis* (Der Ploeg, 2008), and other pathogens. Compared with conventional antibiotics, endolysin has many advantages, such as rapid host killing, host specificity, low chances of developing drug resistance, and efficacy against multidrug-resistant bacteria (Gondil et al., 2020). Autolysin is the other cell wall lytic enzyme that degrades some bonds in the peptidoglycan backbone of the bacterial cell wall (Usobiaga et al., 1996), and it is closely related to the life of cells and participates in the control of cell growth, cell lysis, daughter-cell separation, and biofilm formation (Kalali et al., 2019). Cell wall lytic enzymes have become a valuable tool for biological researchers in the medical and food industry and in agricultural applications (Yu, 1997).

Experimental determination of the cell wall lytic enzymes is time-consuming and laborious, so it is necessary to use an effective method to predict cell wall lytic enzymes. Recently some computational methods for predicting cell wall lytic enzymes have been proposed. Ding et al. (2009) used Chou's amphiphilic pseudo to predict cell wall lytic enzymes; the predictive accuracy was 80.40% with jackknife test. Chen et al. (2016) developed a predictor called "Lypred" that used pseudo amino acid composition (PseAAC) as a feature vector; the predictive accuracy was 91.3% with fivefold cross-validation. Meng et al. (2020) developed a predictor called "CWLy-SVM" that employed the 473-dimensional sequence-based feature descriptor to predict cell wall lytic enzymes; the result was 95.50% with jackknife test. In this paper, the amino acid composition (AAC), the dipeptide composition (DC), the position-specific score matrix auto-covariance (PSSM-AC), and the Auto-covariance average chemical shift (acACS) were used to predict the cell wall lytic enzymes with the same datasets as investigated by Chen et al. (2016).

Data imbalance is always considered a problem in developing efficient and reliable prediction systems; in imbalanced datasets, the classifier would tend to the majority class. Here, the synthetic minority over-sampling technique (SMOTE) was used to solve the problem of imbalance. To remove redundant or irrelevant features, we selected features using the F-score algorithm. The accuracy (Acc) was 99.19% with a balanced dataset in jackknife test by using the optimized combination feature AAC+DC+PSSM-AC+acACS.

MATERIALS AND METHODS

Benchmark Dataset

The benchmark dataset was generated by Chen et al. (2016), The dataset was taken from the Universal Protein Resource (UniProt), using the following steps to collect the sequence: (1) sequences annotated with "Inferred from homology" or "Predicted" were removed. (2) Sequences which were the fragments of other proteins were not included. (3) Sequences containing ambiguous letters such as "B," "J," "O," "U," "X," and "Z" were excluded. To reduce homologous bias and redundancy, the program CD–HIT (Li and Godzik, 2006) was used to remove those sequences that have ≥ 40% pairwise sequence identity. Finally, 375 sequences were obtained; they contained 68 lyases and 307 non-lyases, and the dataset can be expressed as:

$$S = S_{lysases} \cup S_{nonlysases} \tag{1}$$

The dataset can be freely downloaded from http://lin-group.cn/server/Lypred/data.html.

Feature Extraction Techniques

Feature extraction is a crucial step in developing a powerful predictor; a set of reasonable features contains more protein sequence information (Zhu et al., 2018; Yang et al., 2019; Zhang and Liu, 2019). Generally, the feature combination can boost the prediction performance. In this paper, the AAC, the DC, the PSSM-AC, and the acACS were used to predict the cell wall lytic enzymes.

Amino Acid Composition

The amino acid composition of proteins is the most basic feature information in all features. 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). AAC calculates the occurrence frequency of the 20 native amino acids so that the protein sequence can be expressed as 20 features in a feature vector. It can be defined as:

$$P = [x_1, x_2, x_3, \cdots, x_i, \cdots, x_{20}] \tag{2}$$

$$x_i = \frac{n_i}{L} \tag{3}$$

Where n_i is the occurrence number of the 20 native amino acid in protein sequence and L is the length of the protein sequence.

Dipeptide Composition

Dipeptide composition (DC) is calculated as the occurrence frequency of each two adjacent amino acid residues. There are 20*20 = 400 combinations of amino acid pairs. Compared with AAC, DC is a feature that considers some sequence-order information. It can be calculated as:

$$P = \left[f_1, f_2, f_3, \ldots, f_i, \ldots, f_{400}\right] \tag{4}$$

$$f_i = \frac{m_i}{L-1} \tag{5}$$

Where m_i is the occurrence number of i-th dipeptide in protein sequence and L is the length of the protein sequence.

Position-Specific Score Matrix Auto-Covariance

Position-Specific Score Matrix Auto-Covariance (PSSM-AC) is a feature that extracts the evolutionary information of a protein sequence. PSSM-AC was first proposed to predict the protein fold recognition by Dong et al. (2009). Recently, the PSSM-AC was used successfully in many works for the prediction of protein function (Zou et al., 2013; Huang and Li, 2018; Wang et al., 2019b, 2020a). In PSSM-AC, the PSI-BLAST (Position-Specific Iterative Basic Local Alignment Tool) was used to generate PSSM; the threshold of *e*-value is 0.001 and the maximum number of iterations is 3. PSSM-AC is calculated as the correlation between two residues within PSSM. This method can be represented as:

$$P_{PSSM} = \begin{bmatrix} R_{1,1} & R_{1,2} & \ldots & R_{1,j} & \ldots & R_{1,20} \\ R_{2,1} & R_{2,2} & \ldots & R_{2,j} & \ldots & R_{2,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ R_{i,1} & R_{i,2} & \ldots & R_{i,j} & \ldots & R_{i,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ R_{L,1} & R_{L,2} & \ldots & R_{L,j} & \ldots & R_{L,20} \end{bmatrix} \tag{6}$$

$$P_{PSSM} - AC\left(j, \mathrm{lg}\right) = \frac{1}{L - \mathrm{lg}} \sum_{i=1}^{L-\mathrm{lg}} \left(R_{i,j} - \overline{R}_j\right)\left(R_{i+\mathrm{lg},j} - \overline{R}_j\right) \tag{7}$$

$$\overline{R}_j = \frac{1}{L}\sum_{i=1}^{L} R_{i,j} \left(j = 1, \ldots, 20\right) \tag{8}$$

Where $R_{i,j}$ is the score of the residue of the i-th position mutated to the j-th amino acids residue in the protein sequence; a high score means a highly conserved position. L is the length of the protein sequence, lg is the distance along the sequence, and $0 < lg < L$. As a result, the protein sequence generates a $20 \times lg$ dimensional feature vector with PSSM-AC.

Auto-Covariance Average Chemical Shift

As important parameters are measured by nuclear magnetic resonance (NMR) spectroscopy, the chemical shift has been used as a powerful indicator of the protein structure. Several researchers revealed that the average chemical shift (ACS) of a particular nucleus in the protein backbone empirically correlates to its secondary structure (Sibley et al., 2003). acACS was proposed by Fan et al. (2014), In acACS, the secondary structure was converted into the average chemical shift, and then the auto-covariance function was used to construct the vector representing the protein sequence by selecting different. In this work, the secondary structure was obtained by submitting the protein sequence to PSIPRED[1], and then the protein sequence and the corresponding secondary structure were submitted to the acACS web server[2]. It can be calculated as:

For a protein P, where each amino acid in the sequence is substituted by its averaged chemical shift, P can be expressed as:

$$P = \left[A_1^i, A_2^i, A_3^i, \ldots, A_L^i\right] \left(i = {}^{15}N, {}^{13}C_\alpha, {}^{1}H_\alpha, {}^{1}H_N\right) \tag{9}$$

Where ${}^{15}N$ stands for Nitrogen, ${}^{13}C_\alpha$ for alpha Carbon, ${}^{1}H_\alpha$ for alpha Hydrogen, and ${}^{1}H_N$ for Hydrogen linked with Nitrogen.

After we select λ = 17 and $i = {}^{15}N, {}^{13}C_\alpha, {}^{1}H_\alpha, {}^{1}H$, the acACS could be expressed as:

$$\varphi_i^\lambda = \frac{1}{L-\lambda}\sum_{k=1}^{L-\lambda}\left[A_k^i - A_{k+\lambda}^i\right] \left(i = {}^{15}N, {}^{13}C_\alpha, {}^{1}H_\alpha, {}^{1}H_N;\ \lambda < L\right) \tag{10}$$

$$P = \left[\varphi_i^0, \varphi_i^1, \varphi_i^2, \ldots, \varphi_i^\lambda\right] \left(i = {}^{15}N, {}^{13}C_\alpha, {}^{1}H_\alpha, {}^{1}H_N\right) \tag{11}$$

Synthetic Minority Over-Sampling Technique

The numbers of non-lyases are about 4.5 times that of lyases, and this leads to imbalanced data classification problems. In order to overcome this problem, we used SMOTE to solve the problem of imbalance. SMOTE is an over-sampling approach for imbalanced data classification (Wang et al., 2018a; Zhou et al., 2019). The algorithm of SMOTE is described as follows: (1) randomly choose the samples x_i from the minority class, and calculate the Euclidean distance to all other samples in this class, then K nearest neighbors of this sample were selected, (2) select x_i samples from the k nearest neighbors, and (3) generate a new sample x_{new} by: $x_{new} = x_i + \alpha\,(x - x_i)$, α is a random number in (0, 1). In this paper, the protein numbers of lyases and non-lyases are in equilibrium with SMOTE.

Feature Selection

Redundant or irrelevant features will decrease the accuracy of prediction and increase computational time. In order to remove redundant or irrelevant features, a variety of feature selection techniques have been proposed: the analysis of variance (ANOVA) (Tan et al., 2018; Li et al., 2019; Zhang et al., 2020a), Max-Relevance-Max-Distance algorithms (MRMD) (Zou et al., 2016; Wan et al., 2017; Ru et al., 2019; Kwon et al., 2020), and Minimal-Redundancy-Maximal-Relevance (MRMR) (Jiao and Du, 2016; Xu et al., 2016; Wang et al., 2018b; Kabir et al., 2020) are the representative feature selection algorithms. In this study, we selected features using the F-score algorithm; the F-score algorithm was proposed by Yi-Wei (Chen and Lin, 2006). All features are ranked according to F-score values; a higher score indicates a higher likelihood that this feature is more discriminative (Zhang et al., 2020b). It can be calculated as:

$$F_i = \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\sum_{k=1}^{n^+}\left(\bar{x}_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n^- - 1}\sum_{k=1}^{n^-}\left(\bar{x}_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2} \tag{12}$$

Where $\bar{x}_i$ is the average of the i-th feature of the whole sample, $\bar{x}_i^{(+)}$ is the average of the i-th feature of the positive samples, $\bar{x}_i^{(-)}$ is the average of the i-th feature of the negative samples; n^+ is the total number of positive samples, n^- is the total number of negative samples; $\bar{x}_{k,i}^{(+)}$ is the average of the i-th feature of the k-th sample in the positive samples, and $\bar{x}_{k,i}^{(-)}$ is the average of the i-th feature of the k-th sample in the negative samples.

To determine the optimal features, the incremental feature selection (IFS) (Ju and He, 2017; Tang et al., 2018) was employed based on the features ranked. The IFS procedure starts with one feature with the highest score, then adds features to the start feature based on their scores until all the features are added.

Support Vector Machine

The support vector machine was proposed by Vapnik; the basic idea of SVM is to transform the input data into a high-dimensional Hilbert space and then determine the optional separating hyperplane. SVM has been successfully applied in the field of computational biology and bioinformatics (Fan et al., 2013; Li and Wang, 2016; Arif et al., 2018; Chen et al., 2019; Tian et al., 2019; Wang et al., 2019a; Du et al., 2020; Jing and Li, 2020; Yang et al., 2020). Therefore, we used this classifier to build our model. The radial basis function (RBF) kernel was adopted to perform prediction. The regulation parameter c and kernel width parameter γ were tuned via the grid search method. In this paper, the LibSVM package was used to predict cell wall lytic enzymes, which can be downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm.

[1] http://bioinf.cs.ucl.ac.uk/psipred/

[2] http://202.207.14.87:8032/bioinformation/acACS/index.asp

Performance Evaluation

In statistical prediction, three cross-validation methods are commonly used to examine a predictor for its effectiveness in practical applications: k-fold cross-validation, independent dataset test, and jackknife test (Li and Li, 2008; Tan et al., 2019; Dao et al., 2020a,b). Among the three methods, the jackknife test is deemed the most objective and rigorous. Hence, the jackknife test was used to evaluate the performance of this paper.

In order to evaluate the predictive capability and reliability of our model, the sensitivity (Sn), specificity (Sp), Matthew's correlation coefficient (MCC), and accuracy (Acc) (Bustamam et al., 2019; Cheng, 2019; Cheng et al., 2019; Feng et al., 2019; Malebary et al., 2019; Chen et al., 2020; Li and Gao, 2020; Wang et al., 2020b) were measured and defined by:

$$s_n = \frac{TP}{TP + FN} \tag{13}$$

$$s_p = \frac{TN}{TN + FN} \tag{14}$$

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \tag{15}$$

$$Acc = \frac{TP+TN}{TP + TN + FP + FN} \tag{16}$$

Where TP represents the true positive, TN represents the true negative, FP represents the false positive, and FN represents the false negative.

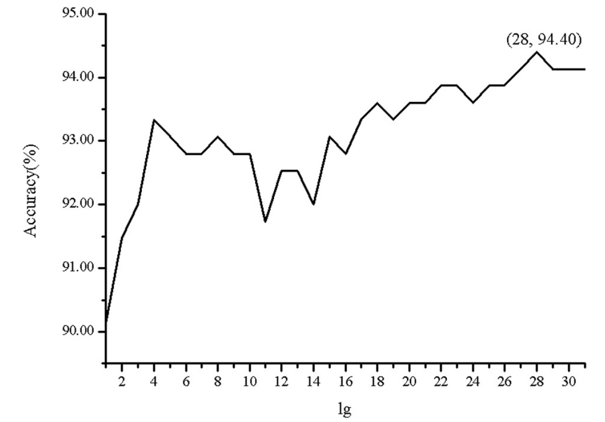

FIGURE 1 | The Acc of position-specific score matrix auto-covariance (PSSM-AC) with different *lg*.

RESULTS AND DISCUSSION

The Choice of Our Model Parameters lg, and Combination Schemes of Chemical Shifts

In order to investigate the effectiveness of the predictive model, the AAC, the DC, PSSM-AC, and the auto-covariance, average chemical shift was selected to predict the cell wall lytic enzymes. Furthermore, for the sake of the best performance of predicting

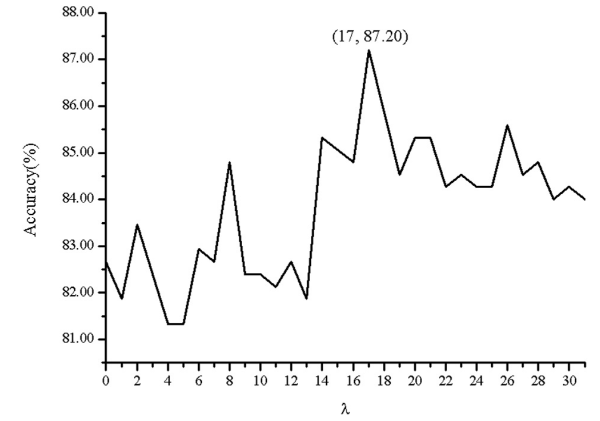

FIGURE 2 | The Acc with respect to the correlation factor λ of the combination mode of chemically shifted atoms ^{15}N, $^{13}C_\alpha$, $^1H_\alpha$, ^1H.

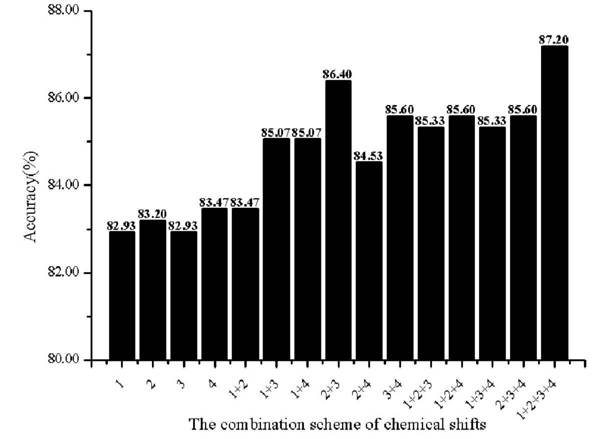

FIGURE 3 | The Acc of different combination schemes of chemical shifts. Numbers denote the chemical shifts of atoms: 1 denotes ^{15}N, 2 denotes $^{13}C_\alpha$, 3 denotes $^1H_\alpha$, 4 denotes ^1H_N.

TABLE 1 | The predictive results of individual features with jackknife test by using SVM.

Features	Sn (%)	Sp (%)	MCC	Acc (%)
AAC	47.06	95.77	0.51	86.93
DC	38.24	97.39	0.48	86.67
PSSM-AC	72.06	99.67	0.81	94.40
acACS	57.35	93.81	0.55	87.20

cell wall lytic enzyme, the lg of the distance was selected, with results in **Figure 1**, and the best lg was 28 when the accuracy was the highest. In addition, the combination mode of chemically shifted atoms and the best parameter λ were selected. **Figure 2** shows that the best parameter λ was 17. The results of combination mode of chemically shifted atoms were shown in **Figure 3**; the best combination mode of chemically shifted atoms was ^{15}N, $^{13}C_{\alpha}$, $^{1}H_{\alpha}$, ^{1}H when the accuracy was the highest.

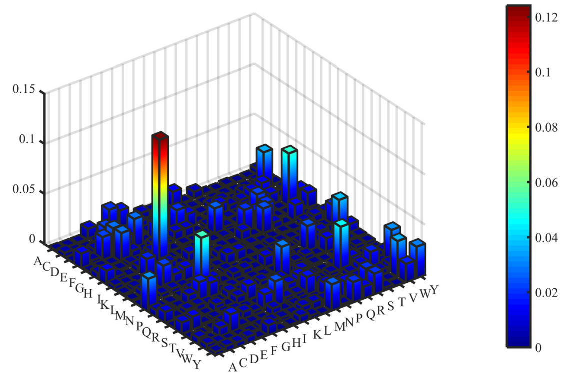

FIGURE 4 | Three-dimensional heat map of DC's F-score value.

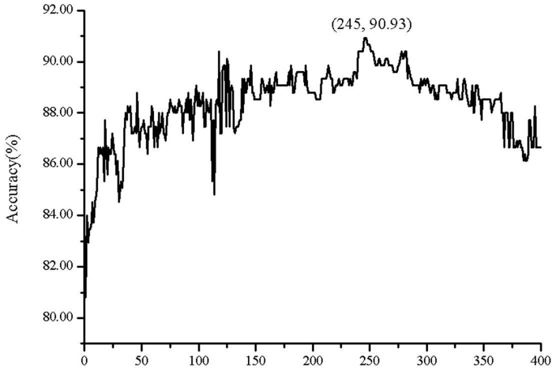

FIGURE 5 | The Acc of dipeptide composition (DC) with the incremental feature selection.

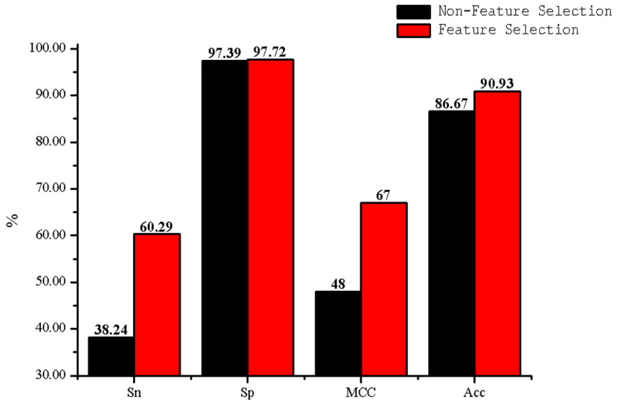

FIGURE 6 | The Acc of DC with feature selection and non-feature selection.

The Predictive Performance of Cell Wall Lytic Enzymes

The predictive performance of cell wall lytic enzymes by using the SVM classification algorithm with SMOTE was listed in **Table 1**. The highest sensitivity (Sn), specificity (Sp), Matthew's correlation coefficient (MCC), and accuracy (Acc) of individual parameters were 72.06%, 99.67%, 0.81, and 94.40% with jackknife test by using PSSM-AC. By comparison, the result of acACS was better than AAC and DC; this is probably due to the fact that

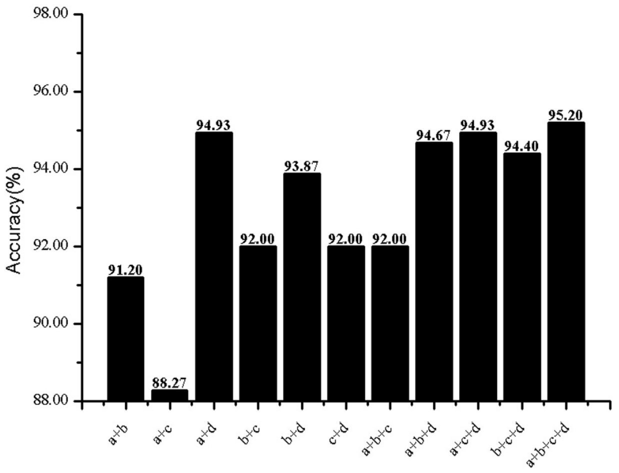

FIGURE 7 | Prediction results of different combined features. Letters denote features: a for AAC, b for DC, c for acACS, d for PSSM-AC.

TABLE 2 | The predictive results of combined feature AAC+DC+acACS+PSSM-AC by using different algorithms with and without SMOTE.

Algorithms	SMOTE (N/Y)	Sn (%)	Sp (%)	MCC	Acc (%)
SVM	N	75.00	99.67	0.83	95.20
RF		41.18	85.99	0.27	77.87
KNN		66.18	80.13	0.40	77.60
NB		86.76	66.78	0.42	70.40
SVM	Y	99.35	99.02	0.98	99.19
RF		85.99	77.52	0.64	81.76
KNN		100.00	73.94	0.77	86.97
NB		92.18	69.38	0.63	80.78

TABLE 3 | The comparison of the predictive results between this paper and existing methods.

Method	Sn (%)	Sp (%)	MCC	Acc (%)
Ding et al.	66.70	88.60	0.573	80.40
Lypred	76.47	93.16	0.678	91.30
CWLy-SVM	85.30	97.70	0.845	95.50
Our predictive model	99.35	99.02	0.98	99.19

acACS considers the protein secondary structure information. The sensitivity (Sn), Matthew's correlation coefficient (MCC), and accuracy (Acc) of AAC were all higher than DC, because DC displays redundant or irrelevant features, so we used "F-score" to select the feature. As shown in **Figure 4**, the closer the color is to red, the higher the F-score of adjacent amino acid residue and the easier it is to distinguish. On the contrary, the closer the color is to blue, the harder it is to distinguish. It can be seen that DC has some redundant information; this redundant information will reduce the prediction success rate. **Figure 5** showed the Acc of DC based on the incremental feature selection (IFS). The peak (the maximum accuracy) can be found in this curve, and it was 90.93% with 245D features. **Figure 6** showed the comparison of DC with feature selection and non-feature selection; we can see that feature selection was successfully applied to remove the irrelevant and redundant features. The Sn, MCC, and Acc were improved remarkably; Acc increased from 86.67 to 90.93%, Sn increased from 38.24 to 60.29%, and the results indicate that feature selection was helpful to enhance the predictive performance. The predictive results of different combined features with SVM without SMOTE were displayed in **Figure 7**. From **Figure 7** we can see the combined feature AAC+DC+acACS+PSSM-AC was better than other parameters. The accuracy (Acc) of combined feature AAC+DC+acACS+PSSM-AC was 95.20% with the jackknife test. This result indicates that the combined feature was powerful in the prediction of cell wall lytic enzymes.

Comparison With Different Classifiers

In order to display the power of our predictive model, our predictive model [Support Vector Machine (SVM)], Random Forest (RF), K-Nearest Neighbors (KNN), and Naive Bayes (NB) were used to predict cell wall lytic enzymes. The predictive performance of SVM, RF, KNN, and NB were listed in **Table 2**. From **Table 2**, we can see the predictive performance of SVM, RF, KNN, and NB with SMOTE were superior to those without SMOTE. The Acc of SVM, RF, KNN, and NB increased by 3.99, 3.89, 9.37, and 10.38% when using SMOTE; the MCC of SVM, RF, KNN, and NB increased by 0.15, 0.37, 0.37, and 0.21 when using SMOTE. In addition, the Sn, Sp, MCC, and Acc of SVM reached 99.35%, 99.02%, 0.98, and 99.19% by using SMOTE. The experimental results show that SVM was useful for improving the predictive performance of cell wall lytic enzymes.

Comparison With Existing Methods

To further investigate the effectiveness of our predictive model, we compared it with existing methods with the same dataset. The comparison results were listed in **Table 3**. From **Table 3**, we can see that the predictive results of cell wall lytic enzymes in our predictive model were better than those of the other methods. Furthermore, the Sn, Sp, MCC, and Acc in our predictive model reached 99.35%, 99.02%, 0.98, and 99.19%, which were 32.65%, 10.42%, 0.407, and 18.79% higher than the Ding et al. (2009) method, 22.88%, 5.86%, 0.302, and 7.89% higher than Lypred, and 14.05%, 1.32%, 0.135, and 3.69% higher than CWLy-SVM. These results indicate that our predictive model was superior to existing methods.

CONCLUSION

With the rapid rise of antibiotic-resistant strains, cell wall lytic enzymes used to destroy bacteria is a viable alternative method to avoid the crisis of antimicrobial resistance. In this work, a reliable and effective computational method was developed to identify the cell wall lytic enzymes. This model was derived from the SVM machine learning algorithm; SMOTE was used to counter the imbalanced data classification problems, and the F-score algorithm was used to remove redundant or irrelevant features. A series of experiments demonstrated that the proposed method is powerful. This method has good capability for distinguishing lyases.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://lin-group.cn/server/Lypred/data.html.

AUTHOR CONTRIBUTIONS

F-ML conceived the selection of feature parameters and performed the results analysis. X-YJ carried out the computation and wrote the manuscript. Both authors reviewed the manuscript.

FUNDING

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

REFERENCES

Ajuebor, J., McAuliffe, O., O'Mahony, J., Ross, R. P., Hill, C., and Coffey, A. (2016). Bacteriophage endolysins and their applications. *Sci. Prog.* 99, 183–199. doi: 10.3184/003685016x14627913637705

Arif, M., Hayat, M., and Jan, Z. (2018). iMem-2LSAAC: a two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into chou's pseudo amino acid composition. *J. Theor. Biol.* 442, 11–21. doi: 10.1016/j.jtbi.2018.01.008

Bhagwat, A., Collins, C. H., and Dordick, J. S. (2019). Selective antimicrobial activity of cell lytic enzymes in a bacterial consortium. *Appl. Microbiol. Biotechnol.* 103, 7041–7054. doi: 10.1007/s00253-019-09955-0

Bustamam, A., Musti, M. I. S., Hartomo, S., Aprilia, S., Tampubolon, P. P., and Lestari, D. (2019). Performance of rotation forest ensemble classifier and feature extractor in predicting protein interactions using amino acid sequences. *BMC Genomics* 20:950. doi: 10.1186/s12864-019-6304-y

Chen, J., Zhao, J., Yang, S., Chen, Z., and Zhang, Z. (2019). Prediction of protein ubiquitination sites in *Arabidopsis thaliana*. *Curr. Bioinform.* 14, 614–620. doi: 10.2174/1574893614666190311141647

Chen, W., Feng, P., and Nie, F. (2020). iATP: a sequence based method for identifying anti-tubercular peptides. *Med. Chem.* 16, 620–625. doi: 10.2174/1573406415666191002152441

Chen, X., Tang, H., Li, W., Wu, H., Chen, W., Ding, H., et al. (2016). Identification of bacterial cell wall lyases via pseudo amino acid composition. *Biomed. Res. Int.* 2016, 1654623–1654623. doi: 10.1155/2016/1654623

Chen, Y. W., and Lin, C. J. (2006). "Combining SVMs with various feature selection strategies," in *Feature Extraction. Studies in Fuzziness and Soft Computing*, Vol. 207, eds I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh (Berlin: Springer).

Cheng, L. (2019). Computational and biological methods for gene therapy. *Curr. Gene Ther.* 19:210. doi: 10.2174/156652321904191022113307

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* 48, D554–D560. doi: 10.1093/nar/gkz843

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019). Computational Methods for Identifying Similar Diseases. *Mol. Ther. Nucleic Acids* 18, 590–604. doi: 10.1016/j.omtn.2019.09.019

Dao, F. Y., Lv, H., Yang, Y. H., Zulfiqar, H., Gao, H., and Lin, H. (2020a). Computational identification of N6-methyladenosine sites in multiple tissues of mammals. *Comput. Struct. Biotechnol. J.* 18, 1084–1091. doi: 10.1016/j.csbj.2020.04.015

Dao, F. Y., Lv, H., Zulfiqar, H., Yang, H., Su, W., Gao, H., et al. (2020b). A computational platform to identify origins of replication sites in eukaryotes. *Brief. Bioinform.* bbaa017. doi: 10.1093/bib/bbaa017

Der Ploeg, J. R. V. (2008). Characterization of *Streptococcus gordonii* prophage PH15: complete genome sequence and functional analysis of phage-encoded integrase and endolysin. *Microbiology* 154, 2970–2978. doi: 10.1099/mic.0.2008/018739-0

Ding, H., Luo, L., and Lin, H. (2009). Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. *Protein Pept. Lett.* 16, 351–355. doi: 10.2174/092986609787848045

Dong, Q., Zhou, S., and Guan, J. (2009). A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics* 25, 2655–2662. doi: 10.1093/bioinformatics/btp500

Du, L., Meng, Q., Chen, Y., and Wu, P. (2020). Subcellular location prediction of apoptosis proteins using two novel feature extraction methods based on evolutionary information and LDA. *BMC Bioinform.* 21:212. doi: 10.1186/s12859-020-3539-1

Fan, G. L., Li, Q. Z., and Zuo, Y. C. (2013). Predicting acidic and alkaline enzymes by incorporating the average chemical shift and gene ontology informations into the general form of Chou's PseAAC. *Process Biochem.* 48, 1048–1053. doi: 10.1016/j.procbio.2013.05.012

Fan, G. L., Liu, Y. L., Zuo, Y. C., Mei, H. X., Rang, Y., Hou, B. Y., et al. (2014). acACS: improving the prediction accuracy of protein subcellular locations and protein classification by incorporating the average chemical shifts composition. *Sci. World J.* 2014:864135. doi: 10.1155/2014/864135

Feng, P., Xu, Z., Yang, H., Lv, H., Ding, H., and Liu, L. (2019). Identification of D modification sites by integrating heterogeneous features in *Saccharomyces cerevisiae*. *Molecules* 24:380. doi: 10.3390/molecules24030380

Gondil, V. S., Harjai, K., and Chhibber, S. (2020). Endolysins as emerging alternative therapeutic agents to counter drug-resistant infections. *Int. J. Antimicrob. Agents* 55:105844. doi: 10.1016/j.ijantimicag.2019.11.001

Huang, G., and Li, J. (2018). Feature extractions for computationally predicting protein post- translational modifications. *Curr. Bioinform.* 12, 387–395. doi: 10.2174/1574893612666170707094916

Jiao, Y. S., and Du, P. F. (2016). Prediction of golgi-resident protein types using general form of chou's pseudo-amino acid compositions: approaches with minimal redundancy maximal relevance feature selection. *J. Theor. Biol.* 402, 38–44. doi: 10.1016/j.jtbi.2016.04.032

Jing, X. Y., and Li, F. M. (2020). Identifying heat shock protein families from imbalanced data by using combined features. *Comput. Math. Methods Med.* 2020:8894478. doi: 10.1155/2020/8894478

Ju, Z., and He, J. J. (2017). Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC. *J. Mol. Graph. Model.* 76, 356–363. doi: 10.1016/j.jmgm.2017.07.022

Kabir, M., Ahmad, S., Iqbal, M., and Hayat, M. (2020). iNR-2L: a two-level sequence-based predictor developed via Chou's 5-steps rule and general PseAAC for identifying nuclear receptors and their families. *Genomics* 112, 276–285. doi: 10.1016/j.ygeno.2019.02.006

Kalali, Y., Haghighat, S., and Mahdavi, M. (2019). Passive immunotherapy with specific IgG fraction against autolysin: analogous protectivity in the MRSA infection with antibiotic therapy. *Immunol. Lett.* 212, 125–131. doi: 10.1016/j.imlet.2018.11.010

Kwon, E., Cho, M., Kim, H., and Son, H. S. (2020). A study on host tropism determinants of influenza virus using machine learning. *Curr. Bioinform.* 15, 121–134. doi: 10.2174/1574893614666191104160927

Li, F. M., and Gao, X. W. (2020). Predicting gram-positive bacterial protein subcellular location by using combined features. *Biomed. Res. Int.* 2020:9701734. doi: 10.1155/2020/9701734

Li, F. M., and Li, Q. Z. (2008). Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. *Protein Pept. Lett.* 15, 612–616. doi: 10.2174/092986608784966930

Li, F. M., and Wang, X. Q. (2016). Identifying anticancer peptides by using improved hybrid compositions. *Sci. Rep.* 6:33910. doi: 10.1038/srep33910

Li, S., Zhang, J., Zhao, Y., Dao, F., Ding, H., Chen, W., et al. (2019). iPhoPred: a predictor for identifying phosphorylation sites in human protein. *IEEE Access* 7, 177517–177528. doi: 10.1109/ACCESS.2019.2953951

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Malebary, S. J., Rehman, M. S. U., and Khan, Y. D. (2019). iCrotoK-PseAAC: identify lysine crotonylation sites by blending position relative statistical features according to the Chou's 5-step rule. *PloS One* 14:e0223993. doi: 10.1371/journal.pone.0223993

Meng, C., Guo, F., and Zou, Q. (2020). CWLy-SVM: a support vector machine-based tool for identifying cell wall lytic enzymes. *Comput. Biol. Chem.* 87:107304. doi: 10.1016/j.compbiolchem.2020.107304

Ru, X., Li, L., and Wang, C. (2019). Identification of phage viral proteins with hybrid sequence features. *Front. Microbiol.* 10:507. doi: 10.3389/fmicb.2019.00507

Shavrina, M. S., Zimin, A. A., Molochkov, N. V., Chernyshov, S. V., Machulin, A. V., and Mikoulinskaia, G. V. (2016). In vitro study of the antibacterial effect of the bacteriophage T5 thermostable endolysin on *Escherichia coli* cells. *J. Appl. Microbiol.* 121, 1282–1290. doi: 10.1111/jam.13251

Sibley, A. B., Cosman, M., and Krishnan, V. V. (2003). An empirical correlation between secondary structure content and averaged chemical shifts in proteins. *Biophys. J.* 84, 1223–1227. doi: 10.1016/s0006-3495(03)74937-6

Sommer, M. O. A., Munck, C., Toftkehler, R. V., and Andersson, D. I. (2017). Prediction of antibiotic resistance: time for a new preclinical paradigm? *Nat. Rev. Microbiol.* 15, 689–696. doi: 10.1038/nrmicro.2017.75

Tan, J. X., Dao, F. Y., Lv, H., Feng, P. M., and Ding, H. (2018). Identifying phage virion proteins by using two-step feature selection methods. *Molecules* 23:2000. doi: 10.3390/molecules23082000

Tan, J. X., Li, S. H., Zhang, Z. M., Chen, C. X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* 16, 2466–2480. doi: 10.3934/mbe.2019123

Tang, H., Zhao, Y. W., Zou, P., Zhang, C. M., Chen, R., Huang, P., et al. (2018). HBPred: a tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.* 14, 957–964. doi: 10.7150/ijbs.24174

Tian, B., Wu, X., Chen, C., Qiu, W., Ma, Q., and Yu, B. (2019). Predicting protein-protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach. *J. Theor. Biol.* 462, 329–346. doi: 10.1016/j.jtbi.2018.11.011

Usobiaga, P., Medrano, F. J., Gasset, M., Garcia, J. L., Saiz, J. L., Rivas, G., et al. (1996). Structural organization of the major autolysin from *Streptococcus pneumoniae*. *J. Biol. Chem.* 271, 6832–6838. doi: 10.1074/jbc.271.12.6832

Wan, S., Duan, Y., and Zou, Q. (2017). HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. *Proteomics* 17, 17–18. doi: 10.1002/pmic.201700262

Wang, C., Li, J., Liu, X., and Guo, M. (2019a). Predicting sub-Golgi apparatus resident protein with primary sequence hybrid features. *IEEE Access* 8, 4442–4450. doi: 10.1109/ACCESS.2019.2962821

Wang, J., Dai, W., Li, J., Xie, R., Dunstan, R. A., Stubenrauch, C., et al. (2020a). PaCRISPR: a server for predicting and visualizing anti-CRISPR proteins. *Nucleic Acids Res.* 48, W348–W357. doi: 10.1093/nar/gkaa432

Wang, J., Yang, B., An, Y., Marquez-Lago, T., Leier, A., Wilksch, J., et al. (2019b). Systematic analysis and prediction of type IV secreted effector proteins by machine learning approaches. *Brief. Bioinform.* 20, 931–951. doi: 10.1093/bib/bbx164

Wang, S., Wang, D., Li, J., Huang, T., and Cai, Y. D. (2018a). Identification and analysis of the cleavage site in a signal peptide using SMOTE, dagging, and feature selection methods. *Mol. Omics* 14, 64–73. doi: 10.1039/c7mo00030h

Wang, S., Zhang, Q., Lu, J., and Cai, Y. (2018b). Analysis and prediction of nitrated tyrosine sites with the mRMR method and support vector machine algorithm. *Curr. Bioinform.* 13, 3–13. doi: 10.2174/1574893611666160608075753

Wang, X. F., Gao, P., Liu, Y. F., Li, H. F., and Lu, F. (2020b). Predicting thermophilic proteins by machine learning. *Curr. Bioinform.* 15, 493–502. doi: 10.2174/1574893615666200207094357

Wu, X., Kwon, S. J., Kim, J., Kane, R. S., and Dordick, J. S. (2017). Biocatalytic Nanocomposites for Combating Bacterial Pathogens. *Annu. Rev. Chem. Biomol. Eng.* 8, 87–113. doi: 10.1146/annurev-chembioeng-060816-101612

Xu, Y., Ding, Y. X., Ding, J., Wu, L. Y., and Xue, Y. (2016). Mal-Lys: prediction of lysine malonylation sites in proteins integrated sequence-based features with mRMR feature selection. *Sci. Rep.* 6:38318. doi: 10.1038/srep38318

Yan, G., Yang, R., Fan, K., Dong, H., Gao, C., Wang, S., et al. (2019). External lysis of *Escherichia coli* by a bacteriophage endolysin modified with hydrophobic amino acids. *AMB Express* 9:106. doi: 10.1186/s13568-019-0838-x

Yang, H., Yang, W., Dao, F. Y., Lv, H., Ding, H., Chen, W., et al. (2020). A comparison and assessment of computational method for identifying recombination hotspots in *Saccharomyces cerevisiae*. *Brief. Bioinform.* 21, 1568–1580. doi: 10.1093/bib/bbz123

Yang, W., Zhu, X. J., Huang, J., Ding, H., and Lin, H. (2019). A brief survey of machine learning methods in protein sub-golgi localization. *Curr. Bioinform.* 14, 234–240. doi: 10.2174/1574893613666181113131415

Yu, J. (1997). Microbial cell wall lytic enzymes which can be used for industrial and pharmaceutical uses. *Food Sci. Biotechnol.* 6, 65–66.

Zhang, H., Xi, Q., Huang, S., Zheng, L., Yang, W., and Zuo, Y. (2020a). iSP-RAAC: identify secretory proteins of malaria parasite using reduced amino acid composition. *Comb. Chem. High Throughput Screen.* 23, 536–545. doi: 10.2174/1386207323666200402084518

Zhang, J., and Liu, B. (2019). A review on the recent developments of sequence-based protein feature extraction methods. *Curr. Bioinform.* 14, 190–199. doi: 10.2174/1574893614666181212102749

Zhang, Z., Yang, Y.-H., Ding, H., Wang, D., Chen, W., and Lin, H. (2020b). Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief. Bioinform* bbz177. doi: 10.1093/bib/bbz177

Zhou, H., Chen, C., Wang, M., Ma, Q., and Yu, B. (2019). Predicting golgi-resident protein types using conditional covariance minimization with XGBoost based on multiple features fusion. *IEEE Access* 7, 144154–144164. doi: 10.1109/ACCESS.2019.2938081

Zhu, X., Feng, C., Lai, H., Chen, W., and Hao, L. (2018). Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Based Syst* 163, 787–793. doi: 10.1016/j.knosys.2018.10.007

Zou, L., Nan, C., and Hu, F. (2013). Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics* 29, 3135–3142. doi: 10.1093/bioinformatics/btt554

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

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