Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 1654623, 8 pages
http://dx.doi.org/10.1155/2016/1654623

Research Article

Identification of Bacterial Cell Wall Lyases via Pseudo Amino Acid Composition

Xin-Xin Chen,[1] Hua Tang,[2] Wen-Chao Li,[1] Hao Wu,[3] Wei Chen,[1,4] Hui Ding,[1] and Hao Lin[1]

[1] *Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics and Center for Information in Biomedicine, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China*
[2] *Department of Pathophysiology, Southwest Medical University, Luzhou 646000, China*
[3] *School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China*
[4] *Department of Physics, School of Sciences, and Center for Genomics and Computational Biology, North China University of Science and Technology, Tangshan 063000, China*

Correspondence should be addressed to Wei Chen; greatchen@heuu.edu.cn, Hui Ding; hding@uestc.edu.cn, and Hao Lin; hlin@uestc.edu.cn

Received 24 April 2016; Accepted 30 May 2016

Academic Editor: Qin Ma

Owing to the abuse of antibiotics, drug resistance of pathogenic bacteria becomes more and more serious. Therefore, it is interesting to develop a more reasonable way to solve this issue. Because they can destroy the bacterial cell structure and then kill the infectious bacterium, the bacterial cell wall lyases are suitable candidates of antibacteria sources. Thus, it is urgent to develop an accurate and efficient computational method to predict the lyases. Based on the consideration, in this paper, a set of objective and rigorous data was collected by searching through the Universal Protein Resource (the UniProt database), whereafter a feature selection technique based on the analysis of variance (ANOVA) was used to acquire optimal feature subset. Finally, the support vector machine (SVM) was used to perform prediction. The jackknife cross-validated results showed that the optimal average accuracy of 84.82% was achieved with the sensitivity of 76.47% and the specificity of 93.16%. For the convenience of other scholars, we built a free online server called *Lypred*. We believe that *Lypred* will become a practical tool for the research of cell wall lyases and development of antimicrobial agents.

1. Introduction

Bacteria are widely distributed on the earth, a significant proportion of which can cause disease. The antibiotic can efficiently treat infectious diseases caused by pathogens. However, antibiotics abuse may cause bacterial drug resistance. Thus, there is an ever-increasing need to find new ways to address this important issue [1, 2]. In the search for more effective therapeutic strategies, great effort has been placed on the study and development of lyases, which benefits from high potency activity toward drug-resistant strains and a low inherent susceptibility to emergence of new resistance phenotypes [3–7].

In 1896, the British bacteriologist Hankin found that the bacteriophage has antibacterial activity [3]. Subsequently, in 1921, Brunoghe and Maisin used bacteriophage to treat staphylococcal skin disease in France, which was the first reported application of bacteriophage to treat infectious diseases [8]. Maxted [9], Krause [10], and Fischetti et al. [11] found that the lysates of Group C streptococci infected with C1 bacteriophage contain an enzyme which has the ability to lyse streptococci and their isolated cell walls. The enzyme is called endolysin which is encoded by bacteriophage gene. It can cause bacteria death by degrading cell wall. It has been reported that 10 ng endolysins can lead to 10^7 bacteria's lysis within 30 seconds [4, 12].

Autolysins are another kind of lyases that are functionally similar to endolysins except they are bacteria-encoded enzymes [13]. It has been reported that autolysins play important roles in several fundamental biological phenomena, such as cell wall enlargement, genetic transformation, flagella extrusion, cell division, and lysis induced by fl-lactam antibiotics, as well as in the "suicidal tendencies" of pneumococci [14–16].

Due to their special biological activity, lyases have been applied in antibacteria drug development. Thus, it is necessary to perform intensive research on lyases to understand the antibacterial mechanism. Although wet experiments are an objective approach for accurately recognizing the lyases, they are often time-consuming and costly. Due to the convenience and high efficiency, computational methods have attracted more and more attention. Many algorithms such as common support vector machine (SVM) [17–19], structured SVM [20], artificial neural network (ANN) [21], Random Forest (RF) [22], K-nearest neighbor (KNN) [23–25], Bayesian classifier [26, 27], Mahalanobis discriminant [28, 29], LibD3C [30], genetic algorithm [31], imbalanced classifier [32], learning to rank [33], and ensemble learning [34, 35] have been developed for protein function prediction. Various sequence features descriptors such as amino acid composition [36, 37], pseudo amino acid composition (PseAAC) [38], physicochemical properties [39], secondary structure features [40], and N-peptide composition [41] were proposed to represent protein sequences [42].

To deal with the problem about lyases prediction, recently, a method was developed to identify cell wall enzymes by using PseAAC and Fisher discriminant [43]. A maximum overall accuracy of 80.4% was obtained with the sensitivity of 66.7% and the specificity of 88.6% [43]. However, further work is needed due to the following reasons. (i) The prediction quality can be further improved. (ii) No web server for the prediction method in [43] was provided, and hence its usage is quite limited, especially for the majority of experimental scientists.

The present study was devoted to development of a new predictor for identifying lyases. For this purpose, an objective and strict benchmark dataset was constructed for training and testing the proposed model in which protein sequences were formulated by using an improved PseAAC. For the convenience of other scholars, a free online server called *Lypred* (at http://lin.uestc.edu.cn/server/Lypred/) was established.

2. Material and Method

2.1. Benchmark Dataset. A high quality dataset is the key to building a robust and accurate predictor. The lyases in bacteria or bacteriophage were regarded as positive samples which were derived from the UniProt [44]. Negative samples, namely, the nonlyases, were also derived from bacteriophage and downloaded from the UniProt. In order to guarantee the reliability of the benchmark dataset, we optimized the data according to the following standards: firstly, the sequences whose protein was with annotations of "Inferred from homology" or "Predicted" were excluded; secondly, we removed the sequences which are the fragments of other proteins; thirdly, the protein sequences containing unknown residues, such as "B," "J," "O," "U," "X," and "Z," were eliminated; fourthly, to avoid overestimation of prediction model that resulted from the high sequence identity, the CD-HIT program [45] was adopted to eliminate redundant sequence by setting the cutoff of sequence identity to 40%. As a result, a total of 68 lyases and 307 nonlyases were obtained to form the final benchmark dataset.

2.2. Features Extraction. A sequence can be represented by two different forms: one is the sequential form and the other is the discrete form [46]. The most common and straightforward way to characterize a protein is to use all the residues in its sequence written as follows:

$$P = R_1 R_2 R_3 R_4, \ldots, R_{L-1} R_L, \tag{1}$$

where R_1, R_2, and R_L are the 1st, 2nd, and Lth amino acid residue of protein P, respectively. Based on the information, a query protein can be predicted by the BLAST or FASTA program. The results are always good for the query sequence which has high similar sequences in benchmark dataset; however, it fails to work when the similar sequences for the query sequence are not found in the training dataset [47]. Therefore, the similarity-based method is not suitable for the case that no homologue was found in the benchmark dataset. The discrete form can overcome the shortcoming and is easy to be treated in statistical prediction. Thus, it has been widely used in protein and DNA formulation [48, 49]. The PseAAC is a typical discrete form that has been widely used for protein function prediction [46, 50, 51].

It is well known that the polypeptide chains fold to tertiary structures based on the physicochemical properties of residues. Thus, it is not enough to analyze the residue compositions of protein molecules. Hence, we proposed to represent protein samples by using an improved PseAAC which includes not only g-gap dipeptide composition, but also correlation of physicochemical property between two residues.

According to the concept of PseAAC, a protein P with the length of L can be formulated in a $(400+n\delta)$ dimension space as given by

$$D = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{400} \\ f_{401} \\ \vdots \\ f_{400+n\delta} \end{bmatrix}, \tag{2}$$

where

$$f_i = \begin{cases} \varphi_i, & 1 \le i \le 400 \\ \varepsilon_i, & 400 < x \le 400 + n\delta, \end{cases} \tag{3}$$

where φ_i denotes the normalized occurrence frequency of the ith kind of g-gap dipeptide in protein P formulated as

$$\varphi_i = \frac{n_i^g}{\sum_{i=1}^{400} n_i^g} = \frac{n_i^g}{L-g-1}, \tag{4}$$

where n_i^g $(i = 1, 2, \ldots, 400)$ denotes the number of the ith g-gap dipeptide in P.

ε_i in (3) is the i-tier sequence correlation factor calculated by the following formulas:

$$\begin{aligned}
\varepsilon_{400+1} &= \frac{1}{L-1}\sum_{t=1}^{L-1}\theta_{t,t+1}^1 \\
\varepsilon_{400+2} &= \frac{1}{L-1}\sum_{t=1}^{L-1}\theta_{t,t+1}^2 \\
&\vdots \\
\varepsilon_{400+n} &= \frac{1}{L-1}\sum_{t=1}^{L-1}\theta_{t,t+1}^n \\
\varepsilon_{400+n+1} &= \frac{1}{L-2}\sum_{t=1}^{L-2}\theta_{t,t+2}^1 \\
\varepsilon_{400+n+2} &= \frac{1}{L-2}\sum_{t=1}^{L-2}\theta_{t,t+2}^2 \\
&\vdots \\
\varepsilon_{400+n+n} &= \frac{1}{L-2}\sum_{t=1}^{L-2}\theta_{t,t+2}^n \\
&\vdots \\
\varepsilon_{400+n\delta} &= \frac{1}{L-\delta}\sum_{t=1}^{L-\delta}\theta_{t,t+\delta}^n \\
&\qquad (\delta < L).
\end{aligned} \tag{5}$$

The correlation $\theta_{x,y}^n$ of physicochemical property between two residues is given by

$$\theta_{x,y}^n = \rho^n(R_x)\,\rho^n\left(R_y\right), \tag{6}$$

where $\rho^n(R_x)$ denotes the nth physicochemical value of amino acid residue R_x. The value is obtained by

$$\rho^n(R_x) = \frac{\rho_0^n(R_x) - \sum_{k=1}^{20}\rho_0^n(R_k)/20}{\sqrt{\sum_{t=1}^{20}\left(\rho_0^n(R_t) - \sum_{k=1}^{20}\rho_0^n(R_k)/20\right)^2/20}}, \tag{7}$$

where $\rho_0^n(R_x)$ is the nth physicochemical original value of amino acid R_x.

Thus, each protein sample can be expressed by $400 + n\delta$ kinds of features according to (2)–(7).

2.3. Feature Selection. Some features are noise or redundant information which will reduce the predictive performance of classification models. Thus, it is very important to develop a method to evaluate the contribution of every feature to the classification. Here, we used ANOVA [52] to rank features defined as

$$\begin{aligned}
&F(i) \\
&= \frac{\sum_{j=1}^{2} m_j\left(\sum_{s=1}^{m_j} f_i(s,j)/m_j - \sum_{j=1}^{2}\sum_{s=1}^{m_j} f_i(s,j)/\sum_{j=1}^{2} m_j\right)^2}{\sum_{j=1}^{2}\sum_{s=1}^{m_j}\left(f_i(s,j) - \sum_{s=1}^{m_j} f_i(s,j)/m_j\right)^2/\left(\sum_{j=1}^{2} m_j - 2\right)},
\end{aligned} \tag{8}$$

where $F(i)$ represents the F-score of the ith feature type, $f_i(s, j)$ is the feature value of the ith feature type of the sth sample in the jth protein type, and m_j is the number of samples in the jth protein type. It is obvious that the larger the $F(i)$ value, the better the discriminative capability the ith feature has.

In order to eliminate the redundant features, we firstly ranked all features according to their F-score from high to low. The first feature subset only contained the feature with the largest F-score; then, a new feature subset was generated when the feature with the second largest F-score was added. The process was repeated until all features were added. The SVM was used to evaluate the performance for each feature subset. The feature subset with the best performance is deemed the optimal feature subset which does not contain redundant features.

2.4. Support Vector Machine. The SVM is a linear-classifier-based supervised machine learning method, which has been successfully used in many bioinformatics fields [48–51, 53–57]. To attain the goal of classification, SVM utilizes the kernel function to deal with the nonlinear transformation, and thus linear inseparable can be converted to a linear problem in high-dimension Hilbert space. In this work, the software LIBSVM [58] was used to execute SVM.

2.5. Performance Standard. To provide a more intuitive and easier-to-understand method to evaluate the prediction performance, we used the following criteria: the sensitivity (Sn), the specificity (Sp), Mathew's correlation coefficient (MCC), the overall accuracy (OA), and the average accuracy (AA), which were defined as

$$\begin{aligned}
\mathrm{Sn} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\
\mathrm{Sp} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\
&\mathrm{MCC} \\
&= \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})}} \\
\mathrm{OA} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}} \\
\mathrm{AA} &= \frac{1}{2}\times\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right),
\end{aligned} \tag{9}$$

Table 1: The original values of nine physicochemical properties used in this study.

Amino acids	Hydrophobicity	Hydrophilicity	Rigidity	Flexibility	Irreplaceability	Mass	pI	pK(α-COOH)	pK(α-NH_3^+)
A	0.62	−0.5	−1.338	−3.102	0.52	15	6.11	2.35	9.87
C	0.29	−1	−1.511	0.957	1.12	47	5.02	1.71	10.78
D	−0.9	3	−0.204	0.424	0.77	59	2.98	1.88	9.6
E	−0.74	3	−0.365	2.009	0.76	73	3.08	2.19	9.67
F	1.19	−2.5	2.877	−0.466	0.86	91	5.91	2.58	9.24
G	0.48	0	−1.097	−2.746	0.56	1	6.06	2.34	9.6
H	−0.4	−0.5	2.269	−0.223	0.94	82	7.64	1.78	8.97
I	1.38	−1.8	−1.741	0.424	0.65	57	6.04	2.32	9.76
K	−1.5	3	−1.822	3.950	0.81	73	9.47	2.2	8.9
L	1.06	−1.8	−1.741	0.424	0.58	57	6.04	2.36	9.6
M	0.64	−1.3	−1.741	2.484	1.25	75	5.74	2.28	9.21
N	−0.78	0.2	−0.204	0.424	0.79	58	10.76	2.18	9.09
P	0.12	0	1.979	−2.404	0.61	42	6.3	1.99	10.6
Q	−0.85	0.2	−0.365	2.009	0.86	72	5.65	2.17	9.13
R	−2.53	3	1.169	3.060	0.60	101	10.76	2.18	9.09
S	−0.18	0.3	−1.511	0.957	0.64	31	5.68	2.21	9.15
T	−0.05	−0.4	−1.641	−1.339	0.56	45	5.6	2.15	9.12
V	1.08	−1.5	−1.641	−1.339	0.54	43	6.02	2.29	9.74
W	0.81	−3.4	5.913	−1.000	1.82	130	5.88	2.38	9.39
Y	0.26	−2.3	2.714	−0.672	0.98	107	5.63	2.2	9.11

where TP is the number of lyases that were correctly predicted, FN denotes the number of lyases that were predicted as the nonlyases, TN is the number of nonlyases that were correctly predicted, and FP denotes the number of nonlyases that were predicted as the lyases.

In addition, we also chose the receiver operating characteristic curve (ROC curve) to measure the performance of the proposed model. ROC curve is a kind of comprehensive index that is drawn by using (1 − Sp) as the abscissa and Sn as the ordinate. Thus, it reveals the continuous variable of Sn and Sp. Generally, we only need to calculate the area under the ROC curve (auROC). The greater the auROC is, the better the discriminate capability the prediction model has is.

3. Results and Discussion

3.1. Forecasting Accuracy. In this work, 9 kinds of physicochemical properties were selected in improved PseAAC [47]. The nine physicochemical properties are hydrophobicity, hydrophilicity, rigidity, flexibility, irreplaceability, side chain mass, pI at 25°C, pK of the α-COOH group, and pK of the α-NH_3^+ group [47], respectively. The original values of the physicochemical properties for 20 amino acids were all listed in Table 1. According to (2)–(7), each protein sample can be formulated by a $(400 + 9\delta)$ dimension vector including 400 g-gap dipeptide compositions and 9δ correlation factors based on physicochemical properties between two residues. From (3)–(5), the prediction performance of our method was influenced by two parameters, namely, g and δ, where g describes the local sequence-order effect and δ represents the global sequence-order effect. The current study searched for the optimal values for the two parameters according to the following standard:

$$\begin{aligned} &0 \le g \le 9, \quad \text{with step } \Delta = 1 \\ &1 \le \delta \le 10, \quad \text{with step } \Delta = 1. \end{aligned} \tag{10}$$

In cross-validation test, n-fold cross-validation, jackknife cross-validation, and independent dataset test are often used for measuring the performance of prediction model. Although the jackknife cross-validation is deemed the most objective because it can always yield a unique result for benchmark dataset given [59, 60] and it has been more and more widely used, it also has obvious drawbacks, such as the large calculation and being time-consuming. Hence, the 5-fold cross-validation was adopted in this work for searching the optimal parameters and the optimal feature subset. Once the optimal feature subset was determined, we used jackknife cross-validation for verification ulteriorly.

Based on (10), a total of $10 \times 10 = 100$ groups of parameters (g, δ) were investigated. For each parameter group (g, δ), there are $400 + 9\delta$ feature subsets. Then, we used feature selection technique defined in (8) to find out the best one in each parameter group. Thus, we obtained the 100 highest OAs for 100 groups of parameters (g, δ). To provide an overall and intuitive analysis, the best OAs were drawn into a heat map, where the column and row of the heat map represent the parameters g and δ, respectively. Each element in the heat map represents one of the 100 groups of parameters (g, δ) and was colorized according to its highest overall accuracy in feature selection process. From Figure 1, we noticed that several elements are red indicating the maximum overall

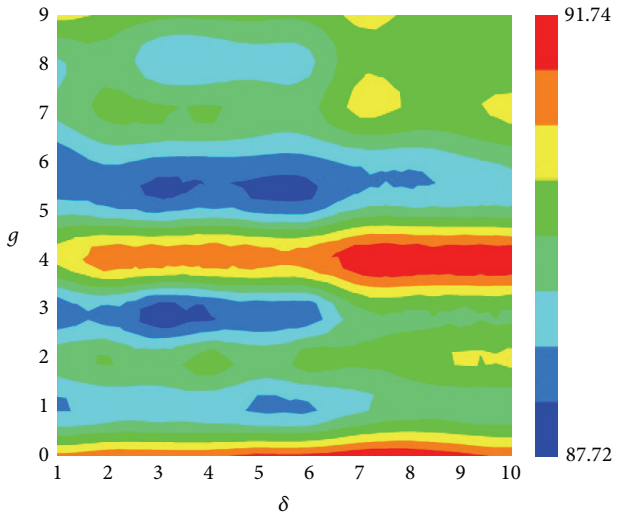

Figure 1: A heat map to show the overall accuracy in 5-fold cross-validation with different parameter groups (g, δ).

accuracy of 91.73% when g equals 0 or 4 and δ equals 7, 8, 9, and 10. Generally, a model with a small number of features can reduce the risk of overfitting. After checking the feature selected results, we found that when using feature selection technique to optimize parameter group ($g = 4$ and $\delta = 7$), the optimal feature subset contains 63 features, which is less than the optimal feature subset in other groups. Thus, the final model was established based on the 63 features from parameter group ($g = 4$ and $\delta = 7$).

Because there is imbalance in our benchmark dataset, the average accuracy and ROC curve were employed to evaluate the model. Thus, we set a series of different classification thresholds to seek the maximum of average accuracy. The maximum AA and corresponding Sn, Sp, MCC, and OA were listed in Table 2. The ROC curve can demonstrate the predictive capability of the proposed method across the entire range of SVM decision values. Thus, we plotted the ROC curve in Figure 2. It shows that auROC is 0.926, demonstrating that our model has capability to predict cell wall lyases.

To investigate whether other algorithms have the same or higher discriminate capability in the same feature space, the performances of Random Forest, Naïve Bayes, and LibD3C were examined by using jackknife cross-validation. Random Forest and Naïve Bayes were executed by using free package WEKA [61]. The LibD3C, a new selective ensemble algorithm, is a hybrid model of ensemble pruning that is based on k-means clustering and the framework of dynamic selection and circulating in combination with a sequential search method [30]. We used default parameters in LibD3C to perform classification.

The jackknife cross-validated results were also recorded in Table 2 for clear comparison. Note that the result for each algorithm in Table 2 was calculated with the maximum AA. As can be seen from the table, although Sn's of Random Forest and Naïve Bayes are no lower than SVM, other indicators (Sp, MCC, OA, AA, and auROC) of SVM are the best.

Table 2: Comparison among the performances of different algorithms.

Algorithm	Sn (%)	Sp (%)	MCC	OA (%)	AA (%)	auROC
SVM	76.47	93.16	0.678	90.13	84.82	0.926
Random Forest	80.88	85.02	0.572	84.27	82.95	0.905
Naïve Bayes	76.47	83.06	0.512	81.87	79.77	0.881
LibD3C	66.18	88.60	0.515	84.53	77.39	0.859

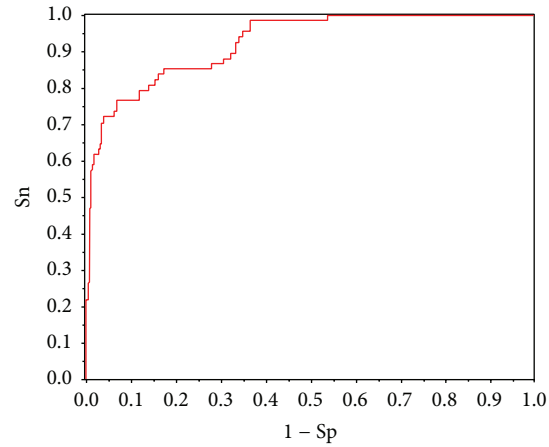

Figure 2: The ROC curve for the proposed model with the 63 optimal features in jackknife cross-validation using SVM.

3.2. Online-Server Guide. A user-friendly online server called *Lypred* was established. A simple guide about the server was given below in order to further make it easier for the users.

Lypred has five pages. Users can browse the server at http://lin.uestc.edu.cn/server/Lypred/ and see the home page on the screen as shown in Figure 3. The Read Me page provides a brief introduction about *Lypred* and the caveat when being used. The Data page shows a brief description about the benchmark dataset and the optimal feature subset used in this work and provides links for downloading. The relevant paper about the detailed development and algorithm of *Lypred* can be seen by clicking the Citation button. Example sequences in FASTA format can be found by clicking the Example button right above the input box.

Users can not only type or copy/paste the query protein sequences into the input box, but also upload FASTA/txt file containing the query protein sequences at the center of the home page of *Lypred*. Note that *Lypred* also has some constraints so as to guarantee the reliability of the results: firstly, protein sequences must be in FASTA format consisting of a single initial line beginning with a greater-than symbol (">") in the first column, followed by lines of sequence data, and the sequence is deemed to end if there is another line starting with ">"; secondly, the query protein sequence should only contain 20 kinds of amino acids; thirdly, the length of a query protein sequence should be no less than eight.

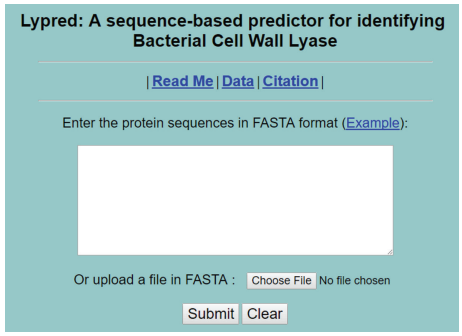

Figure 3: A semiscreenshot to show the home page of Lypred. Its website address is http://lin.uestc.edu.cn/server/Lypred/.

4. Conclusions

With growing drug resistance of pathogenic bacteria, great effort has been placed on the study and development of lyases. Effective identification of lyases will provide convenience for development of new antimicrobials. In this work, we used an improved PseAAC including g-gap dipeptide compositions and correlation factors of the physicochemical properties to extract the characteristics of protein sequences. A feature selection technique based on ANOVA was used to optimize features. The results of AA of 84.82% and auROC of 0.926 make us believe that *Lypred* will become a powerful and useful tool for the experimental study of bacterial cell wall lyase.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

This work was supported by the Applied Basic Research Program of Sichuan Province (nos. 2015JY0100 and LZ-LY-45), the Scientific Research Foundation of the Education Department of Sichuan Province (11ZB122), the Nature Scientific Foundation of Hebei Province (no. C2013209105), the Fundamental Research Funds for the Central Universities of China (nos. ZYGX2015J144 and ZYGX2015Z006), and the Program for the Top Young Innovative Talents of Higher Learning Institutions of Hebei Province (no. BJ2014028).

References

[1] D. Trudil, "Phage lytic enzymes: a history," *Virologica Sinica*, vol. 30, no. 1, pp. 26–32, 2015.

[2] Y. Li, C. Wang, Z. Miao et al., "ViRBase: a resource for virus–host ncRNA-associated interactions," *Nucleic Acids Research*, vol. 43, no. 1, pp. D578–D582, 2015.

[3] E. Hankin, "L'action bactericide des eaux de la Jumna et du Gange sur le vibrion du cholera," *Annales de l'Institut Pasteur*, vol. 10, pp. 511–523, 1896.

[4] V. A. Fischetti, "Bacteriophage lytic enzymes: novel anti-infectives," *Trends in Microbiology*, vol. 13, no. 10, pp. 491–496, 2005.

[5] D. C. Osipovitch, S. Therrien, and K. E. Griswold, "Discovery of novel *S. aureus* autolysins and molecular engineering to enhance bacteriolytic activity," *Applied Microbiology and Biotechnology*, vol. 99, no. 15, pp. 6315–6326, 2015.

[6] C. C. Kietzman, G. Gao, B. Mann, L. Myers, and E. I. Tuomanen, "Dynamic capsule restructuring by the main pneumococcal autolysin LytA in response to the epithelium," *Nature Communications*, vol. 7, article 10859, 2016.

[7] H. Oliveir, L. D. R. Melo, S. B. Santos et al., "Molecular aspects and comparative genomics of bacteriophage endolysins," *Journal of Virology*, vol. 87, no. 8, pp. 4558–4570, 2013.

[8] R. Brunoghe and J. Maisin, "Essais de therapeutique au moyen du bacteriophage du staphylocoque," *Journal des Comptes Rendus de la Société de Biologie*, vol. 85, pp. 1029–1121, 1921.

[9] W. R. Maxted, "The active agent in nascent phage lysis of streptococci," *Microbiology*, vol. 16, no. 3, pp. 584–595, 1957.

[10] R. M. Krause, "Studies on the bacteriophages of hemolytic streptococci. II. Antigens released from the streptococcal cell wall by a phage-associated lysin," *The Journal of Experimental Medicine*, vol. 108, no. 6, pp. 803–821, 1958.

[11] V. A. Fischetti, E. C. Gotschlich, and A. W. Bernheimer, "Purification and physical properties of group C streptococcal phage-associated lysin," *The Journal of Experimental Medicine*, vol. 133, no. 5, pp. 1105–1117, 1971.

[12] R. Schuch, D. Nelson, and V. A. Fischetti, "A bacteriolytic agent that detects and kills Bacillus anthracis," *Nature*, vol. 418, no. 6900, pp. 884–889, 2002.

[13] O. Salazar and J. A. Asenjo, "Enzymatic lysis of microbial cells," *Biotechnology Letters*, vol. 29, no. 7, pp. 985–994, 2007.

[14] H. J. Rogers, H. R. Perkins, and J. B. Ward, *Microbial Cell Walls and Membranes*, Chapman and Hall London, 1980.

[15] M. McCarty, *The Transforming Principle: Discovering That Genes Are Made of DNA*, W. W. Norton & Company, New York, NY, USA, 1986.

[16] J. M. Sanchez-Puelles, C. Ronda, J. L. Garcia, P. Garcia, R. Lopez, and E. Garcia, "Searching for autolysin functions. Characterization of a pneumococcal mutant deleted in the lytA gene," *European Journal of Biochemistry*, vol. 158, no. 2, pp. 289–293, 1986.

[17] K.-C. Chou and H.-B. Shen, "Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms," *Nature Protocols*, vol. 3, no. 2, pp. 153–162, 2008.

[18] M. K. Leong and T.-H. Chen, "Prediction of cytochrome P450 2B6-substrate interactions using pharmacophore ensemble/support vector machine (PhE/SVM) approach," *Medicinal Chemistry*, vol. 4, no. 4, pp. 396–406, 2008.

[19] B. Liu, D. Zhang, R. Xu et al., "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[20] D. Li, Y. Ju, and Q. Zou, "Protein folds prediction with hierarchical structured SVM," *Current Proteomics*, vol. 13, no. 2, pp. 79–85, 2016.

[21] A. Reinhardt and T. Hubbard, "Using neural networks for prediction of the subcellular location of proteins," *Nucleic Acids Research*, vol. 26, no. 9, pp. 2230–2236, 1998.

[22] X. Zhao, Q. Zou, B. Liu, and X. Liu, "Exploratory predicting protein folding model with random forest and hybrid features," *Current Proteomics*, vol. 11, no. 4, pp. 289–299, 2014.

[23] H. Shen and K.-C. Chou, "Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types," *Biochemical and Biophysical Research Communications*, vol. 334, no. 1, pp. 288–292, 2005.

[24] C. Yan, J. Hu, and Y. Wang, "Discrimination of outer membrane proteins using a K-nearest neighbor method," *Amino Acids*, vol. 35, no. 1, pp. 65–73, 2008.

[25] T.-L. Zhang, Y.-S. Ding, and K.-C. Chou, "Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern," *Journal of Theoretical Biology*, vol. 250, no. 1, pp. 186–193, 2008.

[26] A. Bulashevska, M. Stein, D. Jackson, and R. Eils, "Prediction of small molecule binding property of protein domains with Bayesian classifiers based on Markov chains," *Computational Biology and Chemistry*, vol. 33, no. 6, pp. 457–460, 2009.

[27] A. Bulashevska and R. Eils, "Using Bayesian multinomial classifier to predict whether a given protein sequence is intrinsically disordered," *Journal of Theoretical Biology*, vol. 254, no. 4, pp. 799–803, 2008.

[28] H. Lin and Q.-Z. Li, "Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components," *Journal of Computational Chemistry*, vol. 28, no. 9, pp. 1463–1466, 2007.

[29] H. Lin, "The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 252, no. 2, pp. 350–356, 2008.

[30] C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan, and Q. Zou, "LibD3C: ensemble classifiers with a clustering and dynamic selection strategy," *Neurocomputing*, vol. 123, pp. 424–435, 2014.

[31] X. Zeng, S. Yuan, X. Huang, and Q. Zou, "Identification of cytokine via an improved genetic algorithm," *Frontiers of Computer Science*, vol. 9, no. 4, pp. 643–651, 2015.

[32] L. Song, D. Li, X. Zeng, Y. Wu, L. Guo, and Q. Zou, "nDNA-prot: identification of DNA-binding proteins based on unbalanced classification," *BMC Bioinformatics*, vol. 15, article 298, 2014.

[33] B. Liu, J. Chen, and X. Wang, "Application of learning to rank to protein remote homology detection," *Bioinformatics*, vol. 31, no. 21, pp. 3492–3498, 2015.

[34] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, pp. 1–15, Springer, 2000.

[35] T. G. Dietterich, "Ensemble learning," in *The Handbook of Brain Theory and Neural Networks*, vol. 2, pp. 110–125, MIT Press, 2002.

[36] M. H. Smith, "The amino acid composition of proteins," *Journal of Theoretical Biology*, vol. 13, pp. 261–282, 1966.

[37] J. Cedano, P. Aloy, J. A. Perez-Pons, and E. Querol, "Relation between amino acid composition and cellular location of proteins," *Journal of Molecular Biology*, vol. 266, no. 3, pp. 594–600, 1997.

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

[39] S. Saha and G. P. S. Raghava, "BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties," in *Artificial Immune Systems*, G. Nicosia, V. Cutello, P. J. Bentley, and J. Timmis, Eds., vol. 3239 of *Lecture Notes in Computer Science*, pp. 197–204, Springer, New York, NY, USA, 2004.

[40] L. Wei, M. Liao, X. Gao, and Q. Zou, "An improved protein structural classes prediction method by incorporating both sequence and structure information," *IEEE Transactions on NanoBioscience*, vol. 14, no. 4, pp. 339–349, 2015.

[41] C.-S. Yu, C.-J. Lin, and J.-K. Hwang, "Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions," *Protein Science*, vol. 13, no. 5, pp. 1402–1406, 2004.

[42] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K. Chou, "Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Research*, vol. 43, no. W1, pp. W65–W71, 2015.

[43] H. Ding, L. Luo, and H. Lin, "Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition," *Protein and Peptide Letters*, vol. 16, no. 4, pp. 351–355, 2009.

[44] A. M. Bairoch, R. Apweiler, C. H. Wu et al., "The universal protein resource (UniProt)," *Nucleic Acids Research*, vol. 33, pp. D154–D159, 2005.

[45] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[46] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[47] H. Tang, W. Chen, and H. Lin, "Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique," *Molecular BioSystems*, vol. 12, no. 4, pp. 1269–1275, 2016.

[48] P.-P. Zhu, W.-C. Li, Z.-J. Zhong et al., "Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition," *Molecular BioSystems*, vol. 11, no. 2, pp. 558–563, 2015.

[49] W.-C. Li, E.-Z. Deng, H. Ding, W. Chen, and H. Lin, "IORI-PseKNC: a predictor for identifying origin of replication with pseudo k-tuple nucleotide composition," *Chemometrics and Intelligent Laboratory Systems*, vol. 141, pp. 100–106, 2015.

[50] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[51] C. Chen, X. Zhou, Y. Tian, X. Zou, and P. Cai, "Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network," *Analytical Biochemistry*, vol. 357, no. 1, pp. 116–121, 2006.

[52] M. J. Anderson, "A new method for non-parametric multivariate analysis of variance," *Austral Ecology*, vol. 26, no. 1, pp. 32–46, 2001.

[53] R. Wang, Y. Xu, and B. Liu, "Recombination spot identification Based on gapped k-mers," *Scientific Reports*, vol. 6, article 23934, 2016.

[54] J. Chen, X. Wang, and B. Liu, "IMiRNA-SSF: improving the identification of microrna precursors by combining negative sets with different distributions," *Scientific Reports*, vol. 6, article 19062, 2016.

[55] P. Feng, H. Lin, W. Chen, and Y. Zuo, "Predicting the types of J-proteins using clustered amino acids," *BioMed Research International*, vol. 2014, Article ID 935719, 8 pages, 2014.

[56] W. Chen and H. Lin, "Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine," *Computers in Biology and Medicine*, vol. 42, no. 4, pp. 504–507, 2012.

[57] W. Chen and H. Lin, "Prediction of midbody, centrosome and kinetochore proteins based on gene ontology information," *Biochemical and Biophysical Research Communications*, vol. 401, no. 3, pp. 382–384, 2010.

[58] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[59] K.-C. Chou and C.-T. Zhang, "Prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 4, pp. 275–349, 1995.

[60] Y.-C. Wang, X.-B. Wang, Z.-X. Yang, and N.-Y. Deng, "Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature," *Protein and Peptide Letters*, vol. 17, no. 11, pp. 1441–1449, 2010.

[61] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.