Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 8974265, 10 pages
https://doi.org/10.1155/2021/8974265

Research Article

βLact-Pred: A Predictor Developed for Identification of Beta-Lactamases Using Statistical Moments and PseAAC via 5-Step Rule

Muhammad Adeel Ashraf,[1] Yaser Daanial Khan,[1] Bilal Shoaib,[2,3] Muhammad Adnan Khan,[4] Faheem Khan,[5] and T. Whangbo[5]

[1]*Department of Computer Science, University of Management and Technology, Lahore 54770, Pakistan*
[2]*Department of Computer Science, Minhaj University Lahore, Lahore 54770, Pakistan*
[3]*Centre of Research and Innovation in Marytime Affairs (CRIMA), Lahore 54770, Pakistan*
[4]*Pattern Recognition and Machine Learning Lab, Department of Software, Gachon University, Seongnam 13120, Republic of Korea*
[5]*Department of Computer Engineering, Gachon University, Seongnam 13120, Republic of Korea*

Correspondence should be addressed to Muhammad Adnan Khan; adnan@gachon.ac.kr and T. Whangbo; tkwhangbo@gachon.ac.kr

Received 22 September 2021; Accepted 22 November 2021; Published 17 December 2021

Academic Editor: Mario Versaci

Beta-lactamase (β-lactamase) produced by different bacteria confers resistance against β-lactam-containing drugs. The gene encoding β-lactamase is plasmid-borne and can easily be transferred from one bacterium to another during conjugation. By such transformations, the recipient also acquires resistance against the drugs of the β-lactam family. β-Lactam antibiotics play a vital significance in clinical treatment of disastrous diseases like soft tissue infections, gonorrhoea, skin infections, urinary tract infections, and bronchitis. Herein, we report a prediction classifier named as βLact-Pred for the identification of β-lactamase proteins. The computational model uses the primary amino acid sequence structure as its input. Various metrics are derived from the primary structure to form a feature vector. Experimentally determined data of positive and negative beta-lactamases are collected and transformed into feature vectors. An operating algorithm based on the artificial neural network is used by integrating the position relative features and sequence statistical moments in PseAAC for training the neural networks. The results for the proposed computational model were validated by employing numerous types of approach, i.e., self-consistency testing, jackknife testing, cross-validation, and independent testing. The overall accuracy of the predictor for self-consistency, jackknife testing, cross-validation, and independent testing presents 99.76%, 96.07%, 94.20%, and 91.65%, respectively, for the proposed model. Stupendous experimental results demonstrated that the proposed predictor "βLact-Pred" has surpassed results from the existing methods.

1. Introduction

The advent of penicillin was a great revolution of the last century in the medical history of mankind. It was a very effective treatment for many incurable diseases of that time and led to the discovery of more effective remedies for other fatal diseases. After this substantial discovery, a large number of antibiotics were discovered to kill disease-causing bacteria. As the application of such advanced drugs increased, bacteria also acquired resistance to these antibiotics by producing enzymes capable of breaking down these antibiotics [1]. One example of such an antibiotic-resistant enzyme is beta-lactamase which hydrolyzes the beta-lactam ring found in antibiotics, thus destroying its structure. Consequently, effective antibiotic medications are formed by administering the β-lactam antibiotic drug along with a beta-lactamase inhibitor to cure a bacterial infection [2]. In this perspective, β-lactam antibiotics and β-lactamases are of great consideration in clinical set up for the treatment of skin infections, respiratory tract infections, eye infections,

gonorrhoea, soft tissue infections, bronchitis, meningitis, urinary tract infections, pneumonia, and others. A lot of work has been done to understand the structure and the action mechanism of these enzymes in order to elucidate the acquired immunity of microbes against different drugs [3]. β-Lactamase enzymes are produced from bacteria such as cephamycins, penicillins, cephalosporins, and carbapenems [4, 5]. Its action mechanism works by breaking down the beta-lactam ring present in all broad-spectrum antibiotics through hydrolysis, thus deactivating the antibacterial nature of the drug. These antibiotics are used to treat a vast spectrum of Gram-negative and Gram-positive bacterial infections though β-lactamases are produced only from Gram-negative and anaerobic bacteria [5].

Figure 1 depicts the chemical structure of different β-lactam antibiotics. The ring of β-lactam is can be seen as a quad-edge shape for each antibiotic [6]. Three classes of these enzymes, i.e., A, C, and D hydrolyze the substrate by making an acyl-enzyme with the active involvement of serine residue. While class B enzyme uses Zn+ for carrying out its normal function [6].

The initial work of Yildirim et al. studied a ligand based on network model to cluster proteins. A network was created, and the target protein network was connected to their node if there was at least one ligand common**.** However, the study demonstrated results pertaining to only common networks and not for different compounds [7]. Keiser et al. used ligand-based chemical resemblance and formulated subsets of ongoing classes [8]. Cheng et al. used a bipartite network to represent the target node and the protein compound on the basis of similarity sharing protein and ligand [9, 10]. In 2009, Bailey et al. worked on uses of MEME-MAST to extract motifs on the amino acid sequence in β-lactamase [11]. Both works do not concern chemical applications. But since the fuzzy techniques are "data independent," they can also be exploited for the problem under study by the authors [12, 13]. Recently, a predictor named Blapred has been proposed for the classification and identification β-lactamases with its respective classes, i.e., A, B, C, or D by using a three-tier identification computation model via Chou's PseAAC [14].

In the past, chemists and biologists used traditional methods to identify and differentiate of a protein in the laboratory with the utilization of costly equipment which is time-consuming, operator-dependent, costly, and laborious. Besides this, the predictors previously available to classify and identify β-lactamase do not have higher accuracy [14]. There is a need to construct a computational model for the differentiation and classification of β-lactamase enzymes from non-β-lactamase enzymes. The objective of the research is to develop a computational model βLact-Pred by collecting a benchmark dataset, extracting the features and then training the model via Chou's PseAAC [15]. For the purpose of identification and differentiation of proposed model, Chou's five steps are employed which entails [16, 17] (i) construction or selection of an effective benchmark dataset for training and testing the sequence-based statistical predictor, (ii) using mathematical expression, finding a correlation in the dataset, which is called feature extraction; (iii) implementing an algorithm for learning and prediction; (iv) performing numerous kind of persuasive verification and validation testing to factually assess the projected precision of the predictor. This tells that how much our method is effective and trustworthy; (v) developing of a comprehensible and foolproof webserver that will be user-friendly, to ensure its receptiveness and accessibility to the public.

2. Methods and Materials

Consecutively, to develop a vigorous computational model, it is prerequisite to acknowledge an accurate and explicit scale dataset for the sake of training and testing the model. An inoperative dataset may lead the computational model to produce capricious results with untrustworthy validation and unyielding verification testing. It is of uttermost suggestive that the gathered dataset is an accurate, pertinent, nonredundant, related, and comprehensive. Protein's sequence dataset is collected to construct the βLact-Pred computational model. Important and relevant statistical feature vectors are extracted in the form of numerical from the essential protein structure/primary sequences. The computational model is trained on these extracted features using the neural network to accomplish the convergence. Here, Chou's first 3-steps will remain tended, as illustrated in Figure 2.

2.1. Collection of Benchmark Data Set. A database which is publicly available and well-known named Uniport is the major fount to collect the protein sequences of beta-lactamase and non-beta-lactamase. To acquire the concerning positive sequences "beta-lactamase" named keyword was used. An accurate and meticulously process is used to collect dataset in which ambiguous, dubious, and uncertain sequences are excluded, by probability or similarity. Furthermore, for the purpose of accurate and valid results, complete sequences which should not be annotated with fragment-like words are selected. These sequences are annotated with different class names, e.g., class A, B, C, or D. To exclude the redundant and homology-biased sequences, CD-HIT [17] is used with ≥60% resemblance. In consequence, a great quality and an excellent data set is collected which includes the most up-to-date beta-lactamase protein sequences.

After applying CD-HIT, 2172 beta-lactamase sequences were derived. By following the same procedure, 3463 non-beta-lactamase were derived from the same database named UniProt. By considering the Chou's rule [18], any protein sequence can be illustrated as

$$K_{\rho}(\beta) = M_0 M_1 \ldots M_{(n-1)} M_n. \quad (1)$$

Considering all, a minimized dataset was obtained by the following equation:

$$T = T^{+} \cup T^{-}. \quad (2)$$

Here, T^{+} contains 2172 positive beta-lactamase sequences, T^{-} contains 3463 negative beta-lactamase

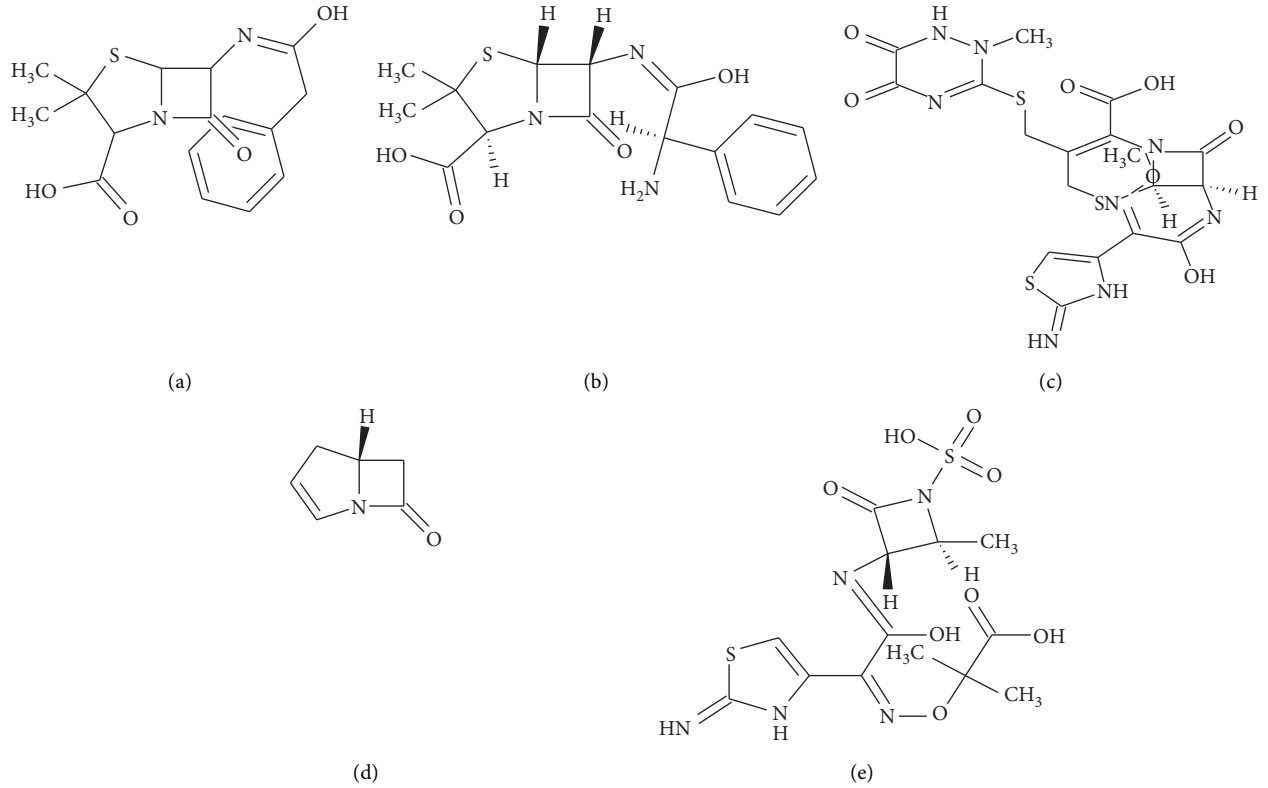

Figure 1: Chemical structures of the β-lactam antibiotics. (a) Penicillin. (b) Ampicillin. (c) Cephalosporin. (d) Carbapenem. (e) Monobactam.

Figure 2: Graphical illustration of the computational model using the Chou's first three stages.

sequences, and $\cup$ shows the "union of two set." A total of 5635 $(2172 + 3463 = 5635)$ sequences comprised dataset.

2.2. Sample Formulation. A specific sequence is constructed by using the amino acids polypeptide chain. These sequences contain biophysical characteristics of proteins. Minor absence or presence of amino acids could not control the characteristics of protein. Behavior of protein is contrived by many constituents, e.g., positioning of amino acids residues and their composition. By observing data and the behavior of different models, it is noted that minor change in comparative composition or ordering of amino acids residue change the characteristics of protein by great extent. Due to all these facts, feature vectors are extricating from primary or core building/blocks of protein by using the computational model which contains both of amino acids relative positions and protein constituents. An extended technique from the technique [18, 19] is used to extract features for βLact-Pred.

2.2.1. Statistical Moment Calculation. Quantitative measures to describe the collection of data are known as statistical moments. Different statistical moments order renders nonidentical data properties. Some statistical moments are helpful in evaluation of the data size, some demonstrate data eccentricity, and some are related to the alignment of proteins. These moments formed by some mathematicians and statisticians contain certain polynomials and distribution functions. βLact-Pred explained by using the moments which include Central, Raw, and Hahn moments. Raw moments, most fundamental moments, contain different properties of a distribution, e.g., mean, variance, and asymmetry. Raw moments do not represent the location, rotation, and scale invariants. To calculate location, rotation, and scale invariants, central moments are calculated deliberately. Central moments again did not calculate the scale and location variants. To calculate scale and location variant properties, another well-liked set of moments named Hahn moment is computed. Hahn moment obtained by using Hahn polynomials exhibits scale and location variants. Major keys to choose these moments are to inspect the composition and composition of residues as they are important factors as per initial discussion. Calculated values yielded from the all above techniques describe in data in their distinctive way. Furthermore, variance is described in terms of moments by using numerical values for capricious datasets [20].

To make protein synthesis, solely 20 amino acids are useable. To compute the moments, distinctive integer index is allocated to each and every amino acids residue. If the allocated index is unique, consistent, and integral, then it barely makes any distinction that what a particular esteem is substituted. Initially, a mapping conversion tool is discovered to convert 1-D (one-dimensional) essential structure into a 2-D (two-dimensional) illustration by equation.

Let S be a sequence of the proteins. The format of S is given as follows:

$$S = \{\beta_1, \beta_2, \beta_3, \ldots, \beta_{m-1}, \beta_m\}. \tag{3}$$

In above, m is surplus in primary protein

$$Z = \lceil \sqrt{m} \rceil, \tag{4}$$

where Z represents the features of S' matrix in the following equation.

All amino acids S that are computed given by $m * m$

$$S' = \begin{bmatrix} K_{11} & K_{12} & \cdots & K_{1n} \\ K_{21} & K_{22} & \cdots & K_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ K_{m1} & K_{m2} & \ldots & K_{mm} \end{bmatrix}. \tag{5}$$

The 2-D matrix S' refers to matrix S. It can be converted by using mapping function as ν.

$$\nu(\beta_x) = \alpha_{pq}, \tag{6}$$

where p and q signify the index of K in S'.

Moments can be computed till 3-degree by using two-dimensional S', and consequent equation is utilized for computing raw moments.

$$Z_{mm} = \sum_{x=1}^{l} \sum_{y=1}^{l} x^{m y^n a_{xy}}, \tag{7}$$

where $m + n]$ indicates the order of moments, l describes the aspects of matrix, which should be the same, i.e., Z. Moments till 3-degree are computed as $Z_{00}, Z_{01}, Z_{02}, Z_{10}, Z_{11}, Z_{12}, Z_{20}, Z_{21}$, and Z_{22}.

Data center is like center of gravity. Distribution of data is fair along with the data's central point w.r.t the average weight of data. It computes the following raw moments and known as an argument $(\overline{v}, \overline{w})$, where

$$\begin{aligned} \overline{v} &= \frac{Z_{10}}{Z_{00}}, \\ \overline{w} &= \frac{Z_{01}}{Z_{00}}. \end{aligned} \tag{8}$$

Central moments are calculated by point where the centroid is acting. The following equation is employed to compute the central moments such as

$$B_{st} = \sum_{k=1}^{m} \sum_{l=1}^{m} (k - \overline{v})^s (l - \overline{w})^1 a_{kl}. \tag{9}$$

For Hahn moments calculation, 1-D analysis S was transferred to a square matrix analysis S'. The Hahn polynomials in n order can be employed as

$$\omega_m^{a,b}(p, M) = (M + b - 1)_m (M - 1)_m \times \sum_{l=0}^{m} (-1)^l \frac{(-m)_l (-p)_l (2M + a + b - m - 1)_l}{(M + b - 1)_l (M - 1)_l} \frac{1}{l!}. \tag{10}$$

The above polynomial uses Pochhammer mark as

$$(b)_l = b, (b + 1) \ldots (b + 1 - 1). \tag{11}$$

Simple form of the above can be represented by using a delta operator:

$$(b)_l = \frac{\Delta(b + l)}{\Delta(b)}. \tag{12}$$

Hahn moments are calculated by weighing function and square rule such as

$$\widetilde{\beta}_n^{c,d}(q, N) = \beta_n^{c,d}(q, N) \sqrt{\frac{o(q)}{c_n^2}} \quad n = 0, 1, \ldots N - 1, \tag{13}$$

whereas

$$o(q) = \frac{\phi(c + q + d)\phi(d + q + 1)(c + d + q + 1)_N}{(c + d + 2q + 1)n!(N - q - 1)!}. \tag{14}$$

The logical data for 2-dimentional discrete data is calculated by using the following equation:

$$G_{ef} = \sum_{c=0}^{N-1} \sum_{d=0}^{N-1} \alpha_{c\ d} \widetilde{J}_t^{g,h}(c, N) \widetilde{J}_s^{u,v}(b, N), \quad n = 0, 1, \ldots N - 1. \tag{15}$$

In order, Han and Central moments can be calculated up to 3.

2.2.2. Generation of Position Relative Index Matrix. Information regarding the composition/arrangements is the foundation of any computational model that is used to predict protein functions. Physical properties of the proteins can be determined by assuming a key function for the area of amino acid. Relative positioning of amino acid in polypeptide chain is very important as position relative index matrix (PRIM) divulges information about the relative position of amino acids in polypeptide chain. Position relative index matrix (PRIM) excerpts the amino acid's location information in polypeptide chain [20]. A matrix of 20×20 dimensions related to PRIM matrix is given as follows:

$$Z_{\text{PRIM}} = \begin{bmatrix} Q_{1\longrightarrow 1} & Q_{1\longrightarrow 2} & Q_{1\longrightarrow 3} & Q_{1\longrightarrow b} & \cdots & Q_{1\longrightarrow 20} \\ Q_{2\longrightarrow 1} & Q_{2\longrightarrow 2} & Q_{2\longrightarrow 3} & Q_{2\longrightarrow b} & \cdots & Q_{2\longrightarrow 20} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ Q_{d\longrightarrow 1} & Q_{d\longrightarrow 2} & Q_{d\longrightarrow 3} & Q_{d\longrightarrow b} & \cdots & Q_{d\longrightarrow 20} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ Q_{U\longrightarrow 1} & Q_{U\longrightarrow 2} & Q_{U\longrightarrow 3} & Q_{U\longrightarrow b} & \cdots & Q_{U\longrightarrow 20} \end{bmatrix}. \tag{16}$$

An element of matrix such as $Q_{d\longrightarrow b}$ contains the aggregate of b^{th} residue in contradiction of the first index of d^{th} residue. It makes 400 coefficients which show a large number. Dimensions of PRIM matrix are curtailed by computing the three moments, i.e., raw, central, and Hahn.

2.2.3. Generation of Reverse Position Relative Index Matrix (RPRIM). Reverse position relative index matrix (RPRIM) is used to extract hidden features from protein sequences which have the ambiguity of homologous sequences. RPRIM has a 20×20 dimension matrix containing 400 coefficients same as in the PRIM, but it is used in a reverse order of the PRIM [20].

$$Q_{\text{RPRIM}} = \begin{bmatrix} R_{1\longrightarrow 1} & R_{1\longrightarrow 2} & \cdots & R_{1\longrightarrow k} & \cdots & R_{1\longrightarrow 20} \\ R_{2\longrightarrow 1} & R_{2\longrightarrow 2} & \cdots & R_{2\longrightarrow k} & \cdots & R_{2\longrightarrow 20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ R_{t\longrightarrow 1} & R_{t\longrightarrow 2} & \cdots & R_{t\longrightarrow k} & \cdots & R_{t\longrightarrow 20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ R_{z\longrightarrow 1} & R_{z\longrightarrow 2} & \cdots & R_{z\longrightarrow k} & \cdots & R_{z\longrightarrow 20} \end{bmatrix}. \tag{17}$$

Like PRIM, the dimension of the RPRIM matrix is also curtailed by computing the three moments, i.e., raw, central, and Hahn.

2.2.4. Frequency Matrix. Frequency matrix is a technique used to determine the structure and how frequently proteins are occurring. This plays a significant role in sequencing of proteins. PRIM holds the series information of amino acids, while frequency matrix does not hold that series information [20]. The following expression is used to compute the frequency of the matrix as

$$\xi = \{\tau_1, \tau_2, \tau_3, \tau_4, \ldots, \tau_{20}\}. \tag{18}$$

Here, τ_i denotes the frequency of i^{th} essential amino acid.

2.2.5. Generation of Accumulative Absolute Position Index Vector. Frequency matrix contains the protein formation related information and the total occurrence of protein information. Frequency matrix did not contain the information related to the occurrence of amino acid residues in a polypeptide chain. Accumulative absolute position incidence vector (AAPIV) is used to compute the information related to the position of amino acid residue in the polypeptide chain. AAPIV contains position relevant information in a vector form. A vector with 20 elements in which each component encompasses a numerical ordered value to represent the amino acid position relevant information from the residue [20]. Native sequence shows the specific residue occurrence in a protein structure which is given as follows:

$$v^k_{\mu^1} \ldots v^k_{\mu^2} \ldots v^k_{\mu^3} \ldots v^k_{\mu^n}. \tag{19}$$

It represents v^k residue which is placed at a position of $\mu^1, \mu^2, \mu^3, \ldots \mu^n$

Let accumulative absolute position index vector represented as

$$T = \{\nu_1, \nu_2, \nu_3, \nu_4, \ldots, \nu_{20}\}. \tag{20}$$

Hence, i^{th} element of the accumulative absolute position index vector is computed by

$$\nu_i = \sum_{u=1}^{n} su. \tag{21}$$

2.2.6. Generation of Reverse Accumulative Absolute Position Index Vector. As per earlier discussion, detecting ambiguous patterns using feature extraction is an efficient technique. RAAPIV did the same task as AAPIV performs, but it finds the patterns in a reverse order [20]. It also contains 20 elements which can be represented as follows:

$$\delta = \{o_1, o_2, o_3, o_4, o_5, \ldots, o_{20}\}. \tag{22}$$

Reversed sequence in RAAPIV is shown as

$$\omega^k_{m1} \ldots \omega^k_{m2} \ldots \omega^k_{m3} \ldots \omega^k_{mn}. \tag{23}$$

The amino acid residue $\overset{k}{\omega}$ that occurs in the reverse order sequence and the term $m_1, m_2, m_3, \ldots, m_n$ represents their ordered position. The significance of any residue is calculated as

$$\ell_i = \sum_{m=1}^{n} t_m. \tag{24}$$

All of these abovementioned features have specific biological significance. These methods help in extracting position and composition relative features from the amino acid sequence which is a very pivotal aspect while dealing with proteins. Each amino acid, in its surrounding, plays a role in describing the physiochemical characteristics of that molecule; thus, these features help in extracting such information. For example, the frequency of amino acids in molecule, position relative occurrence of amino acids, composition of a specific peptide, and absolute positioning of residues.

3. Operational Algorithm via Neural Network

Artificial neural network is one of the most significant tools for tackling the issue examined in this paper, it mimics preparing data as depicted in Figure 3. Neural network clarifies the fundamental shape of every residue within a

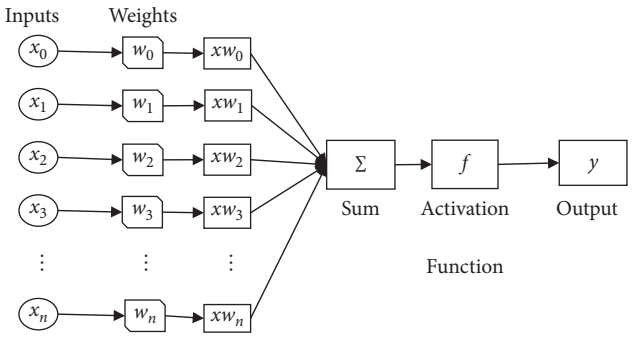

Figure 3: Graphical representation of the artificial neural network for βLact-Pred.

protein. To train the model, composition of positive and negative feature vectors which are extracted in above section are used. These feature vectors depict the two-dimensional structure of protein by using central, raw, and Hahn moments. Here, in this study, the neural network was considered as neural network which is represented by directed graph similar to the biological neuron system in brain. Back propagation ANN was used instead of SVM because of many reasons that ANN performs better than SVM. First of all, ANN is a parametric model, while SVM is not. As in ANN, there can be many hidden layers depending on features and parameters [20]. In SVM, we have support vectors that are acquired by training data. In some cases, support vectors can have many support vectors with weight of each vector. ANN can also have one or many outputs, while SVM can have only one output. In case of a n-ary classifier, ANN can be trained in one step, while SVM needs to train n support vectors one by one that is time-consuming [20].

ANN is fast and flexible. ANN can be reached at global optimal point, and we do not face any issue regarding choosing the number of parameters, but in case of SVM, we need to select hyperparameters. Less amount of memory is required to store ANN, but SVM requires much memory because it needs to store support vectors as well. Results in ANN are more readable and interpretable [21, 22].

4. Formulation of Results and Discussion

4.1. Estimated Accuracy Metrics. The unbiased assessment of newly constructed computational model is the most key aspect that aids to estimate the accomplishment of that computational model [22, 23]. Conversely, for such kind of an unbiased assessment, two important aspects one must keep in mind that (i) the choice of metrics accuracy and (ii) the test method deployed for the validation of the computational model. Here, first classify the measurements for the unbiased assessment and then use the numerous validation and verification techniques.

4.2. Mathematical Formulation of Metrics. It is obvious that, for any machine learning problem, some collective and important metrics are used for formulation of the metrics, which are (1) Acc (accuracy) is the percentage of correctly classified samples from total input dataset; (2) MCC (Matthews correlation coefficient) is used in case of binary classification, and it is also considered as balanced measure even in multiple classes of different sizes; (3) S_n (sensitivity) is the percentage of true positive or those samples that are correctly classified as positive, and it is also called true positive recognition rate. (4) S_p (specificity) is the percentage of true negative or those samples that are correctly classified as negative, and it is also called true negative recognition rate.

Predominantly, these four metrics were introduced in 2001, and an accurate set of four measures was obtained in [24] for all of these measures.

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{\mathrm{N}_-^+}{\mathrm{N}^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \dfrac{\mathrm{N}_+^-}{\mathrm{N}^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = 1 - \dfrac{\mathrm{N}_-^+ + \mathrm{N}_+^-}{\mathrm{N}^+ + \mathrm{N}^-} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{Mcc} = \dfrac{1 - \left(\left(\mathrm{N}_-^+/\mathrm{N}^+\right) + \left(\mathrm{N}_+^-/\mathrm{N}^-\right)\right)}{\sqrt{\left(1 + \left(\mathrm{N}_+^- - \mathrm{N}_-^+/\mathrm{N}^+\right)\right)\left(1 + \left(\mathrm{N}_-^+ - \mathrm{N}_+^-/\mathrm{N}^-\right)\right)}} & -1 \le \mathrm{Mcc} \le 1 \end{cases}. \tag{25}$$

Here N^- signifies non-β-lactamases data, predicted as non-β-lactamases correctly by βLact-Pred. N_+^- signifies the non-β-lactamases aggregate number which are anticipated inaccurately as β-lactamases by βLact-Pred. Additionally, N^+ is the β-lactamases aggregate number which are predicted correctly as β-lactamases by βLact-Pred, and N_-^+ is the β-lactamases aggregate number which are identified inaccurately as non-β-lactamase by βLact-Pred. Accordingly,

equation (25) provides the information regarding Sn, Sp, Acc, and consistency more relaxed to recognize and innate, especially when we discourse about MCC [25, 26].

These accuracy metrics have been used/identified by a numerous researchers [27, 28], but merely for binary class data labelled. Multiclass data labelled identification is a utterly diverse problem, which has been supplementary prominent in computational biology [29] and biomedicine [30]. Consequently, it entails a diverse kind of accuracy metrics for formulation [29].

4.3. Self-Consistency Testing. The self-consistency testing is a term referred as the ultimate test for the validation of efficiency and efficacy of the prediction model using the test cases by training the data set. The reason behind the implementation of self-consistency is that the obtained results are individual and the actual true positive rate of the benchmark dataset is also known. Self-consistency results are revealed in Table 1; it can be observed that the βLact-Pred has the 99.76% Acc, 99.76% Sp, 99.76% Sn, 0.99 MCC, and 0.99 AUC.

4.4. Validation of Model via Leave-One-Out. Validation is a significant step that comes toward the end of the process. Its motivation is to discover that how much the model is proficient. A few validation techniques are utilized to validate the model. To validate the model, data are portioned into two parts; (1) training set and (2) testing set. The model is trained on training data, and then its performance is measured on testing data. As the validation techniques select the data haphazardly for predicting the model, there is not well-defined technique that expresses how to partition the data from the given dataset. Generally, the predictive model can be tested using numerous types of testing, i.e., k-folds (subsampling), independent testing, and leave-one-out (jackknife) [27, 30]. Jackknife testing is amongst the most frequently used validation techniques. Jackknife works by overlooking each observation from the data and set up the model on residual data. At the end, average is calculated of all calculations and the output is unique. Issues like sampling or sub-sampling are alleviated.

Jackknife is used to quantify the quality of the predictor, and it is likewise generally utilized in these sorts of problems. It is an iterative technique that computes the accuracy of the model for all variations of the sample of size $n-1$. The jackknifing technique trains the predictor on left-out data and estimates overall accuracy by meticulously leaving out every observation from a dataset. It is more efficient as it overwhelms the issues that are triggered by data independency and subsampling [31]. Results of jackknife validation testing is 96.07% which is higher than the BlaPred [12] and are revealed in Table 2.

4.5. K-Fold Cross-Validation Testing. Cross-validation is a method to thrive an expectancy for the proposed model as an exemplary method in the absence of validation set. Cross validation tests the model on given training dataset and prevents underfitting and overfitting. In k-fold cross validation, the dataset is portioned into k sets and k is picked at start, and afterward, it is kept constant. Generally, k is kept 5 or 10; however, in the proposed method, k is set to 10. The model is tested k times and, in each iteration, 9 sets (k-1) are used for training set and the one set (k set) is treated as testing set. Subsequent to performing k iterations, the accuracy of model is computed by the sum of each iteration and then divided by k. This average accuracy is considered as a result of cross validation. The overall 10-fold validation was repeated 20 times, so that the credibility of results is increased, as illustrated in Table 3.

Table 1: Performance analysis of self-consistency for βLact-Pred.

Predictor/identifier	Precision metrics				
	Acc (%)	Sp (%)	Sn (%)	MCC	AUC
βLact-Pred	99.76	99.76	99.76	0.99	0.99

4.6. Independent Dataset Testing. To evaluate the precision of βLact-Pred, independent testing was performed, in which the training/testing split method was used for validating the model. Out of 2172 positive and 3463 negative samples, three different train/test split ratios were used which were 90/10, 80/20, and 70/30. After sufficient training, the left-out samples were used for testing, and subsequent evaluation of the accuracy of the proposed prediction technique was performed. Based on the ability and inability of the model to recognize the test samples accurately, all the described metrics in equation (25) were computed, which are mentioned in Table 4.

4.7. Comparative Analysis. βLact-Pred uses a composition and position variant feature extraction method for classification besides neural network. The other existing prediction models discussed in text use type-1 PseAAC, type-2 PseAAC, and classic PseAAC for feature extraction combined with SVM (support vector machine). Both the techniques (type I and type II) and classic are based on the PseAAC model, presented in [32]. The method of feature extraction for such kind of problems has extreme significance. The proficiency to uncover deeply obscure patterns within a specified set of data is highly anticipated for a feature extraction algorithm. The capability of a model to translate deeply obscure patterns in the primary structure into coefficients is dependent on a variable λ. The value of λ not only determines the size of the feature vector but also plays a significant role in sieving out the correlation among residues within a peptide chain. The factors produced by βLact-Pred are not reliant on such a variables. The vector size of the feature is adjusted and carefully calculates all possible interactions between all possible residues in the peptide chain in the form of succinct. βLact-Pred used both assorted sequences of β-lactamase and non-β-lactamase which is subsequently used as a dataset for the purpose of training and testing. As illustrated in Table 1, βLact-Pred reveals a greater sensitivity, specificity, accuracy, and MCC for

Table 2: Jackknife testing results for βLact-Pred.

Predictor/identifier	Precision metrics				
	Acc (%)	Sp (%)	Sn (%)	MCC	AUC
βLact-Pred	96.07	97.39	96.96	0.92	0.93
BlaPred [12]	93.57	94.00	89.24	0.70	—

Table 3: Performance analysis of 10-fold cross-validation results (20 iterations) for βLact-Pred.

10-fold (iterations)	Precision metrics				
	Acc (%)	Sp (%)	Sn (%)	MCC	AUC
1	93.92	97.23	99.74	0.97	0.99
2	96.11	97.97	99.90	0.98	1.00
3	93.87	97.00	99.12	0.98	0.99
4	94.68	97.26	99.32	0.98	0.99
5	95.03	97.58	97.22	0.98	0.99
6	96.26	98.72	97.59	0.98	1.00
7	93.38	99.00	98.32	0.98	0.98
8	94.04	97.00	97.23	0.97	0.99
9	96.24	98.30	97.57	0.99	1.00
10	93.34	96.00	96.97	0.99	0.98
11	94.94	97.32	96.63	0.97	0.99
12	93.41	99.80	99.01	0.98	0.99
13	93.72	99.00	99.91	0.98	0.99
14	93.90	95.11	99.89	0.98	0.99
15	94.09	96.44	99.23	0.98	0.99
16	96.12	98.00	99.12	0.98	1.00
17	94.25	96.79	98.90	0.98	0.99
18	95.15	97.70	97.26	0.98	0.99
19	94.17	96.57	99.32	0.98	0.99
20	95.60	97.82	97.34	0.97	1.00
Average	94.61	97.80	99.89	0.98	1.00

Table 4: Results for independent dataset testing of three different methods.

Splits	Precision metrics				
	Acc (%)	Sp (%)	Sn (%)	MCC	AUC
90/10	95.27	94.50	96.90	0.8990	0.92
80/20	91.57	92.60	93.40	0.8310	0.89
70/30	88.10	91.34	92.10	0.8120	0.86
Average	91.65	92.81	94.13	0.8473	0.89

prediction of β-lactamases and non-β-lactamases than the other previous predictors. Experiments prove that it is a highly efficient technique as compared to previous ones. Rigorous validation in diverse scenarios elucidates that the method is less noisy and more effective for the prediction of beta-lactamases. Subsequently, it is also established that the presented methodology provides higher throughput and accuracy than the previous predictors. To quantitatively evaluate and compare the βLact-Pred, an independent dataset of 75 β-lactamases, previously reported by [12], was used in (Table 5).

In addition to this, the results of βLact-Pred were also compared with CNN-BLPred [33], which performs the functional and molecular classification of β-lactamases by employing a deep learning method/technique called the convolutional neural network (CNN). The study performs classification of β-lactamases at molecular and functional level; however, for comparison with βLact-Pred, only molecular classification (level 1) results were considered. Comparative analysis is provided in Table 6.

Furthermore, βLact-Pred applies numerous types of approach and uses composition and positioning features of sequences of protein to accomplish the identification of β-lactamases. In first, it uses PseAAC, and then it calculates the statistical moments, AAPIV, RAAPIV, PRIM, and RPIRM using the relative positioning features of protein; thus, βLact-Pred outperforms its counterparts.

5. Web Server

Final step of Chou is the enlargement of user-friendly and publicly accessible webserver for the comfort of chemists and biologists as an enlightened in [34, 35]. Publicly accessible and user-friendly webserver development and

Table 5: Comparative performance of βLact-Pred as compared to the previous predictors.

Predictor/identifier	Total number of β-lactamases	Predicted β-lactamases
βLact-Pred	75	62
BlaPred [12]	75	58
PredLactamase [32]	75	40

Table 6: Comparative analysis of 10-fold cross-validation results with CNN-BLPred.

Predictor/identifier	Precision metrics			
	AUC	Sp (%)	Sn (%)	MCC
CNN-BLPred [33]	1.00	95.73	99.90	0.96
βLact-Pred	1.00	97.80	99.89	0.98

establishment signifies the direction of the future in order to develop prediction methodologies [34, 35]. For this purpose, various computational analysis and research findings have been reported. Therefore, useful and practical webserver has significantly enhanced the overall impacts of computational biology on medical sciences directing medicinal chemistry into an unsurpassed revolution [12]. In this view, the webserver shall be established for βLact-Pred as described in the paper.

6. Conclusion

Multidrug-resistant strains of bacteria have posed a great threat to human health nowadays. Bacteria have cleverly and speedily acquired resistance against most of the antibiotics of the time and are creating hurdles in an effective cure for diseases. It is believed that, within few years, all prevailing antibiotics would lose their efficacy against these multidrug-resistant bugs. β-Lactamase is one of the safeguards produced by bacteria which protects it from the adverse action of β-lactam antibiotics. Various data preprocessing techniques are used to calculate the feature vector including raw, Hahn, and central Moments and position and composition variant features. For this purpose, an artificial neural network is used for training and predicting the sequences. The results for the proposed computational model was validated by employing numerous types of approaches, i.e., self-consistency testing, jackknife testing, cross-validation, and independent testing. The overall accuracy of the predictor for self-consistency testing, jackknife testing, cross-validation, and independent testing by using paradigm metrics presents 99.76%, 96.07%, 94.20%, and 91.65%, respectively, for the proposed model. Stupendous experimental results demonstrated that the proposed predictor "βLact-Pred" has surpassed results from the existing methods.

Data Availability

The data used in this paper are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the GRRC program of Gyeonggi province (GRRC-Gachon2020 (B04), Development of AI-Based Healthcare Devices).

References

[1] E. Grudzien-Nogalska and M. Kiledjian, "New insights into decapping enzymes and selective mRNA decay," *Wiley Interdisciplinary Reviews: RNA*, vol. 8, no. 1, p. e1379, 2017.

[2] M. D. Barnes, M. L. Winkler, M. A. Taracila et al., "*Klebsiella pneumoniae* carbapenemase-2 (KPC-2), substitutions at ambler position Asp179, and resistance to ceftazidime-avibactam: unique antibiotic-resistant phenotypes emerge from β-lactamase protein engineering," *mBio*, vol. 8, no. 5, 2017.

[3] S. Leclercq, A. Derouaux, S. Olatunji et al., "Interplay between penicillin-binding proteins and SEDS proteins promotes bacterial cell wall synthesis," *Scientific Reports*, vol. 7, no. 1, p. 43306, 2017.

[4] R. L. Oehler, A. P. Velez, M. Mizrachi, J. Lamarche, and S. Gompf, "Bite-related and septic syndromes caused by cats and dogs," *The Lancet Infectious Diseases*, vol. 9, no. 7, pp. 439–447, 2009.

[5] S. A. Hasan and K. S. Abass, "Prevalence of gram negative bacteria isolated from patients with burn infection and their antimicrobial susceptibility patterns in Kirkuk city, Iraq," *Indian Journal of Public Health Research & Development*, vol. 10, no. 8, pp. 2197–2201, 2019.

[6] D. Lee, S. Das, N. L. Dawson, D. Dobrijevic, J. Ward, and C. Orengo, "Novel computational protocols for functionally classifying and characterising serine beta-lactamases," *PLoS Computational Biology*, vol. 12, no. 6, p. e1004926, 2016.

[7] R. Sharma, Impact of comorbidity on severity of Covid-19 patients: a network of target coding genes perspective, 2020.

[8] M. J. Keiser, J. J. Irwin, and B. K. Shoichet, "The chemical basis of pharmacology," *Biochemistry*, vol. 49, no. 48, pp. 10267–10276, 2010.

[9] F. Cheng, C. Liu, J. Jiang et al., "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Computational Biology*, vol. 8, no. 5, p. e1002503, 2012.

[10] F. Cheng, Y. Zhou, W. Li, G. Liu, and Y. Tang, "Prediction of chemical-protein interactions network with weighted network-based inference method," *PLoS One*, vol. 7, no. 7, p. e41064, 2012.

[11] T. L. Bailey, M. Boden, F. A. Buske et al., "Meme suite: tools for motif discovery and searching," *Nucleic Acids Research*, vol. 37, pp. W202–W208, 2009.

[12] F. C. Morabito, M. Versaci, G. Pautasso, C. Tichmann, and A. U. Team, "Fuzzy-neural approaches to the prediction of disruptions in ASDEX upgrade," *Nuclear Fusion*, vol. 41, no. 11, pp. 1715–1723, 2001.

[13] M. Versaci, G. Angiulli, P. Di Barba, and F. C. Morabito, "Joint use of eddy current imaging and fuzzy similarities to

assess the integrity of steel plates," *Open Physics*, vol. 18, no. 1, pp. 230–240, 2020.

[14] A. Srivastava, R. Kumar, and M. Kumar, "BlaPred: predicting and classifying β-lactamase using a 3-tier prediction system via Chou's general PseAAC," *Journal of Theoretical Biology*, vol. 457, pp. 29–36, 2018.

[15] X. Cui, Z. Yu, B. Yu, M. Wang, B. Tian, and Q. Ma, "UbiSitePred: a novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components," *Chemometrics and Intelligent Laboratory Systems*, vol. 184, pp. 28–43, 2019.

[16] S. Khanum, M. A. Ashraf, A. Karim et al., "Gly-LysPred: identification of lysine glycation sites in protein using position relative features and statistical moments via Chou's 5 step rule," *Computers, Materials and Continua (CMC-COMPUT MATER CON)*, vol. 66, no. 2, 2020.

[17] S. Ilyas, W. Hussain, A. Ashraf, Y. D. Khan, S. A. Khan, and K. C. Chou, "iMethylK-PseAAC: improving accuracy of lysine methylation sites identification by incorporating statistical moments and position relative features into general PseAAC via Chou's 5-steps rule," *Current Genomics*, vol. 20, no. 4, pp. 275–292, 2019.

[18] 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.

[19] K.-C. Chou, "Using subsite coupling to predict signal peptides," *Protein Engineering Design and Selection*, vol. 14, no. 2, pp. 75–79, 2001.

[20] S. J. Malebary, M. S. U. Rehman, and Y. D. Khan, "iCrotoK-PseAAC: identify lysine crotonylation sites by blending position relative statistical features according to the Chou's 5-step rule," *PLoS One*, vol. 14, no. 11, pp. e0223993–11, 2019.

[21] S. Siranush and H. Anna, "One approach to the problem of the existence of a solution in neural networks," *American Journal of Mathematical and Computer Modelling*, vol. 5, no. 3, pp. 83–88, 2020.

[22] S. Haykin, *Neural Networks: A Comprehensive Foundation, 1994*, Mc Millan, Trenton, NJ, USA, 2010.

[23] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, vol. 27, pp. 3104–3112, 2014.

[24] Y. Xu, L. Li, J. Ding, L.-Y. Wu, G. Mai, and F. Zhou, "Gly-PseAAC: identifying protein lysine glycation through sequences," *Gene*, vol. 602, pp. 1–7, 2017.

[25] P.-M. Feng, H. Ding, W. Chen, and H. Lin, "Naive Bayes classifier with feature selection to identify phage virion proteins," *Computational and mathematical methods in medicine*, vol. 2013, Article ID 530696, 2013.

[26] P. Charoenkwan, N. Anuwongcharoen, C. Nantasenamat, M. Hasan, and W. Shoombuatong, "In silico approaches for the prediction and analysis of antiviral peptides: a review," *Current Pharmaceutical Design*, vol. 27, no. 18, 2020.

[27] F. Ali and M. Hayat, "Classification of membrane protein types using voting feature interval in combination with Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 384, pp. 78–83, 2015.

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

[29] K. Ahmad, M. Waris, and M. Hayat, "Prediction of protein submitochondrial locations by incorporating dipeptide composition into Chou's general pseudo amino acid composition," *Journal of Membrane Biology*, vol. 249, no. 3, pp. 293–304, 2016.

[30] Z. Chen, Y.-Z. Chen, X.-F. Wang, C. Wang, R.-X. Yan, and Z. Zhang, "Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs," *PLoS One*, vol. 6, no. 7, p. e22930, 2011.

[31] L. Nanni, S. Brahnam, and A. Lumini, "Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 360, pp. 109–116, 2014.

[32] Y. Dou, B. Yao, and C. Zhang, "PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine," *Amino Acids*, vol. 46, no. 6, pp. 1459–1469, 2014.

[33] S. Mondal and P. P. Pai, "Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction," *Journal of Theoretical Biology*, vol. 356, pp. 30–35, 2014.

[34] R. Kumar, A. Srivastava, B. Kumari, and M. Kumar, "Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine," *Journal of Theoretical Biology*, vol. 365, pp. 96–103, 2015.

[35] C. White, H. D. Ismail, H. Saigo, and D. B. Kc, "CNN-BLPred: a convolutional neural network based predictor for β-Lactamases (BL) and their classes," *BMC Bioinformatics*, vol. 18, no. 16, p. 577, 2017.