International Journal of Engineering & Technology, 7 (1.1) (2018) 435-438

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Enterobacteria virulence factor prediction server

M. Thirunavukkarasu [1], K. Dinakaran [2], E.N. Sathishkumar [3], S. Gnanendra [4] *

[1] Department of Computer Science & Applications, Mahendra Arts & Science College (Autonomous), Kalippatti, Research and Development Center, Bharathiar University, Coimbatore, Tamil Nadu, India
[2] Department of Computer Science and Engineering, PMR Engineering College, Adayalampattu, Chennai, Tamil Nadu, India
[3] Department of Computer Science, Periyar University, Salem, Tamil Nadu, India
[4] Department of Biotechnology, Yeungnam University, Gyeongsan, Gyeongbuk, 38541, South Korea
**Corresponding author E-mail: gnani.science@gmail.com*

Abstract

The continuous usage of antibiotics has resulted in the increase of multidrug resistance in the bacteria. The drastic increase in the bacterial genome projects has paved a path for the identification of potentially novel virulence-associated factors and their possibility as novel drug targets. Thus in the present study, we have implemented SMO classifiers for the better prediction of proteins based on individual protein sequences amino acid composition (AAC) and the performance of evaluation was checked via threshold dependent parameters such as Sensitivity, Specificity, Accuracy, and Mathew correlation coefficient. The predictions are based on the dataset incorporated in the database of five major virulence factors from six pathogens of *Enterobacteriacae*. This comprehensive database can serve as a source for the selection of significant virulence factor based on the intellectual Gene ontology terms that play a critical role in the pathogenesis and its surveillance in the host.

Keywords: *Virulence prediction; amino acid composition; data mining; classifiers; SMO.*

1. Introduction

The continuous usage of antibiotics has resulted in the increase in the emergence of multidrug resistance in the bacteria. This has led an alarming concern to all researchers to significantly identify the potential virulence factors as a most promising and alternative candidate drug targets that can overwhelm the worse situations of multi-drug resistant bacteria that has a concerning issue of public health sectors [1].

The drastic increase in the bacterial genome projects has paved a path for the identification of potentially novel virulence-associated factors and their possibility as a novel drug target [2]. However, we have many significant sequence databases holds genome and protein sequences resulting in huge amount of data we are unfortunate to have the virulence factors that determine the pathogenesis of the bacteria.

In the present work, we emphasized a database of five major virulence factors from six pathogens of Enterobacterial. This comprehensive database can serve as a source for the selection of significant virulence factor based on the intellectual Gene ontology terms that play a critical role in the pathogenesis and its surveillance in the host [3].

The current release of this database holds the 5 virulence factors, Capsule, Cell wall, Flagella, Pilli and Toxins from 6 noteworthy emerging multi-drug resistant bacteria of Enterobacterial namely, Enterobacter spp., Escherichia spp., Klebsiella spp., Proteus spp., Salmonella spp., and Shigella spp., The collective informations such as, Virulence factors, Protein names, Family, Domains, related genes, Gene ontology and GO-ID, Uniprot_ID, Interpro_ID, Keywords, Pathways and Sequences [4]. The warehouse is designed as a comprehensible for researchers with Gene Ontology terms as a query to access the intellectual virulence factors and their related pieces of information.

This database is committed to envisage the functional and structural aspects which can ignite researchers to elucidate its role as pathogenesis and its mechanism to emerge as a most dreadful infectious disease. This can also serve as a specialized source to retrieve the novel virulence factors that are ignored in due course of time also virulence-associated factors employed by bacterial pathogens to effectively inhibit the host niche [5]. These understandings can ecstasy the researchers to attain the immediate investigations for the development of novel approaches and drugs to treat and prevent the diseases caused by multiple drug resistance (MDR) bacteria.

2. Materials and methods

2.1. Source of data

The data of this server is mainly concerned with the prediction from six important pathogens and the sequences were collected from UniProt database [6]. The protein sequence data of virulence factors such as Capsule, Cell wall, Flagella, Pilli, and Toxins were considered as most important factors that can contribute and notify the bacteria as a pathogen. A total of 1706 sequences, out of which 827 sequences were found to be virulent sequences and 879 non-virulent sequences were used as training dataset and in the development of final database.

2.2. Annotations

The data pertaining to UniProt ID were used to retrieve virulence factors related pieces of information such as Function, PDB-ID,

Secondary structure, 3D structure, Motifs, Domains, Sequences, Gene Ontologies, Related Genes, and Pathways were collected from various sources [7].

2.3. Amino acid composition (AAC)

The compositions of amino acid representing a protein are considered as 20 dimensions feature vectors. The amino acid composition is calculated by AAC = Total occurrence of i^{th} amino acid in the sequence, Where i is single amino acid.

2.4. Classification algorithms

To perform data modeling, two important classification methods such as Bayes methods and functions methods were implemented in WEKA [8]. The tests were performed using 5-fold internal cross validation. Both the classification algorithms were tested and compared in terms of Accuracy, Sensitivity, Specificity and Mathews Correlation Coefficients (MCC) [9].

2.5. Cross validation and performance evaluation

An objective based statistical method, Leave-One-Out Cross-Validation (LOOCV) also known as jackknife cross-validation is was used as an effective method to evaluate a classifier for its effectiveness [10]. From the training dataset, the virulent proteins and non-virulent proteins are defined as positive and negative respectively and determined the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The performance of both classification algorithms were measured with standard parameters like Accuracy, Sensitivity, Specificity, and MCC.
Sensitivity (Sn): Sensitivity measures the ability of the process to predict correct results

$$Sn = \frac{tp}{tp+fn} \times 100\%$$

Specificity (Sp): Specificity measures the ability of a process to predict incorrect results.

$$Sp = \frac{tn}{tn+fp} \times 100\%$$

Accuracy (Acc): Accuracy measures the degree of correctness of the predicted results to its actual value or the experimental value

$$Acc = \frac{tp+tn}{tp+fn+tn+fp} \times 100\%$$

Mathews Correlation Coefficient (MCC): In the machine learning MCC measures the degree to which the binary classification is correct.

$$MCC = \frac{(tp)(tn)-(fp)(fn)}{\sqrt{(tp+fp)(tp+fn)(tp+fp)(tn+fn)}} \times 100\%$$

3. Result and discussion

The availability of complete bacterial genomes of pathogens is the rich source of information to determine the virulence factors and its associated proteins. However, due to the complexity in determining the virulence factors, the urgency in computational tools to predict the virulence factors are desperately in need [11]. In line with this several predicting algorithms have been proposed by many research groups. In connection with this the machine learning algorithms has existed as a choice to deal these prediction strategies. In this work we have evaluated the statistical parameters such as total instances considered in the study, correctly classified and in-correctly classified instances, Kappa Static value, Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error, Coverage of cases (0.95 level), Mean rel. region size (0.95 level), True Positive, False Positives, Precision, false positive rate and false negative rate for both virulent and non-virulent protein sequences were predicted for both the classifier methods.

3.1. Performance of classification methods

The entire dataset of 1706 number of protein sequences was divided into training (1500) and testing sets (206) for which two classification methods were calculated by Cross Validation method. The cross validation parameters such as sensitivity and specificity were optimized over the training set by maximizing the accuracy and minimizing the training error. For 5-fold cross validation, the entire dataset is classified into five datasets, in which 1 dataset is used as test set while the other remaining set are used as training sets. It is iterated till that every sequence in the datasets has been considered as test sequence. At every iteration, the best parameter values were noted for the final average values of the classification method.
The performance of bayes methods and functions methods was evaluated by using amino acid composition as input features. The evaluation of the performance was carried out using five-fold cross validation. Among the Bayes classification methods, the Naïve Bayes method exhibited the Sensitivity (66.67%), Specificity (72.72%), Accuracy (69.70%) and MCC (38.89) while compared to the other two methods of the same classification such as Bayes Net and Naïve Bayes updatable methods. It is observed that the accuracies and MCC values of both the methods were almost similar however the difference is observed in sensitivity and specificity.
Among the various functions methods such as logistic, multilayer perceptron, SGD, Simple Logistic, SMO and Voted Perceptron methods, the remarkable Sensitivity (86.67%), Specificity (71.43%), Accuracy (81.82%) and MCC (58.10) was shown by SMO method. These results significantly confirmed that optimized SMO is able to learn crucial features responsible for the accurate classification and helps to understand the prediction accuracy which can also be increased by providing more comprehensive information of a protein sequence. The performance evaluation of significant classifiers from both the bayes and functions methods is given in Table.1.

Table 1: Performance Evaluation Parameters of Bayes and SMO Methods

Measures	Bayes Method	SMO
Sensitivity	66.67	86.67
Specificity	72.22	71.43
Accuracy	69.70	81.82
MCC	38.89	58.10
False Positive rate	27.78	28.57
False Negative rate	33.33	13.33

Hitherto, the above mentioned parameters are threshold-dependent. So to calculate threshold-independent performance, a Receiver Operating Characteristic (ROC) curve was plotted between TP rate (sensitivity) and FP rate (1-specificity) [12]. The Area under the Curve (AUC) describes inherent between sensitivity and specificity and thereby measures the accuracy of the SVM model. The AUC obtained from ROC curve (Figure 1) was best for SMO that significantly reflects that SMO is highly comprehensive but condensed and meaningful information is required to attain such high prediction accuracy.

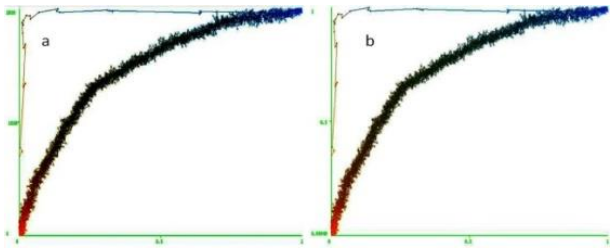

Fig. 1: Receiver Operating Characteristic (ROC) Curve Plotted between TP Rate (Sensitivity) and FP Rate (Specificity) A. Bayes Method B. SMO.

Implementation of SMO classifiers and prediction tool in EvirDB

Further, the server was implemented using the MySQL relational database system (http://www.mysql.com). It was built by a PHP programmes to communicate with the MySQL system. Most of the 3D structures were computed with MODELLER [13-15]. Secondary structure assignments were made by GOR. Motifs and Domains were assigned by using InterProscan and Prosite databases. The server accepts a protein sequence inputs in FASTA format. Then it calculates the amino acid composition and uses SMO classifier for the further prediction. This server is accessible at www.evirdb.info. The HTML interface allows the user to query the database by gene ontology terms, virulence factors or by organism names. The information of virulence factors and related gene and protein sequences can be retrieved, from MySQL database using appropriate query. The hypertext pre-processor scripting language was used to retrieve the data and also as a server-side HTML embedded scripting language for building dynamic pages based on the user's query [16-18]. The server Homepage is shown in figure.2.

3.2. Composition of database

In brief, the current release of this server hold information of six pathogens namely Enterobacter spp., Escherichia spp., Klebsiella spp., Proteus spp., Salmonella spp., and Shigella spp.,that belongs to Enterobacteria. The 5 noteworthy virulence factors, Capsule, Cell wall, Flagella, Pilli and Toxins data that mount to ~36,000 were included in the database.

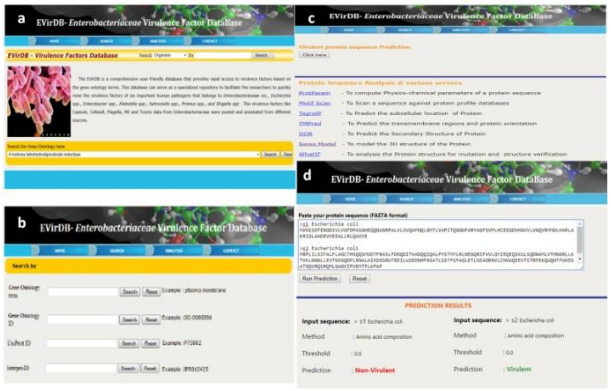

Fig. 2: The Screen Shot of the Implemented Evirdb Prediction Server. A. Home Page of the Server B. the Simple Search Page of the Server with Various Search Options C. Various Analysis Tools D. Sample Input and Prediction Result.

The virulence factors of each organism stored in this server were shown in Figure.3. The total Gene Ontology terms used to retrieve the Virulence factor information were alphabetically sorted to make the query search as user-friendly.

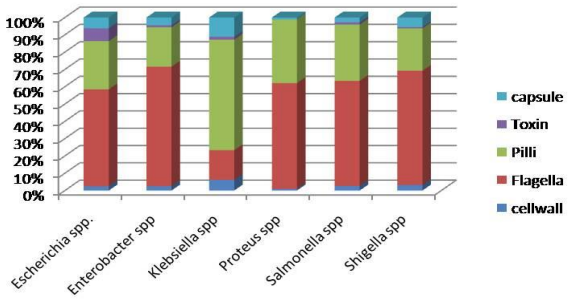

Fig. 3: The Total Number of Virulence Factors of Each Organism in the Server.

3.3. Amino acid prediction

The virulence of protein is predicted based on individual protein sequences amino acid composition (AAC). The sequences from the user are considered as query and by using javascripts, the amino acids compositions of the input sequences were determined and displayed. Further, predicted amino acids compositions were considered for further classification by using SMO and the output will be displayed as virulent or non-virulent protein. The amino acid composition prediction of the entire dataset is shown in figure.4.

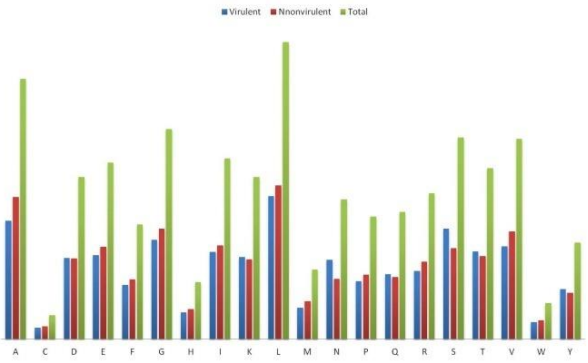

Fig. 4: The Amino Acids Composition Analysis for the Entire Dataset Used in the Analysis. (Total: 1706 Sequences).

While using the protein sequences (1706) as training data set, the total Amino Acid Composition analysis revealed the presence of basic amino acids such as leucine (L), alanine (A), glycone (G), serine (S) and valine (V) as in most frequently occurring amino acids in the sequence. Interestingly the high number of Leucine is observed in both virulent and non-virulent sequences. The high composition of amino acid such as Alanine (A), Cystine (C), Glutamic acid (E), Phenyl alanine (F), Glycine (G), Histdine (H), Isoleucine (I), Leucine (L), Methionine (M), Proline (P), Arginine (R), Valine (V) and Tryptophan (W) were observed in the protein sequences that are classified as non-virulent proteins, while the remaining amino acids such as Aspargine (N), Glutamine (Q), Serine (S), Threonine (T) and Tyrosine (Y) were high in the protein sequences predicted as virulent which significantly implies that these residual compositions plays a major role in the successful classification of models that can predict the sequence as virulent and non-virulent.

4. Conclusion

This database is committed to providing the rational information on virulence factors of most important bacterias. The Database is enhanced to accept the new submissions by researchers from worldwide. In future, the database is aimed to serve as a specialized database that focuses on the 5 important virulence factors such as Capsule, Cell wall, Flagella, Pilli and Toxins of all human bacterial pathogens. As the expression of virulence genes and their pathways plays a significant role in the regulation of pathogenicity

of the bacteria and also to colonize the host system, we have future plan to include much details on regulation and host interactions of the virulence factors. It is suggested that implementing SMO classifiers of functions methods would result in better prediction of virulence factors based on the dataset implemented in this predictions.

References

[1] Mulder NJ, Mazandu GK & Rapano HA, "Using Host-Pathogen Functional Interactions for Filtering Potential Drug Targets in Mycobacterium tuberculosis", J. Mycobac. Dis, (2013). https://doi.org/10.7196/SAMJ.5437.

[2] Warner DF & Mizrahi V, "Approaches to target identification and validation for tuberculosis drug discovery: a UCT perspective", South African Medical Journal, Vol.102, (2012), pp.457-460.

[3] Waldvogel FA, "Infectious diseases in the 21st century: old challenges and new opportunities", Int. J. Infect. Dis., Vol.8, (2004), pp.5–12. https://doi.org/10.1016/j.ijid.2003.01.001.

[4] Weiss RA, "Virulence and pathogenesis", Trends in Microbiology, Vol.10, (2002), pp.314–317. https://doi.org/10.1016/S0966-842X(02)02391-0.

[5] Hogan D & Kolter R, "Why are bacteria refractory to antimicrobials?" Curr. Opin. Microbiol, Vol.5, (2002), pp.472–477. https://doi.org/10.1016/S1369-5274(02)00357-0.

[6] Byarugaba DK, "Antimicrobial resistance in developing countries and responsible risk factors", Int. J. Antimicrob. Agents, Vol.24, (2004), pp.105–110. https://doi.org/10.1016/j.ijantimicag.2004.02.015.

[7] Docampo R, "New and reemerging infectious diseases", Emerg. Infect. Dis., (2003), pp.1030– 1033. https://doi.org/10.3201/eid0908.030324.

[8] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P & Witten IH, "The WEKA data mining software: an update", ACM SIGKDD explorations newsletter, Vol.11, No.1, (2009).

[9] Saha S & Raghava GP, "Prediction of continuous B-cell epitopes in an antigen using recurrent neural network", Proteins: Structure, Function, and Bioinformatics, Vol.65, No.1, (2006). https://doi.org/10.1002/prot.21078.

[10] Kalita MK, Nandal UK, Pattnaik A, Sivalingam A, Ramasamy G, Kumar M, Raghava GP & Gupta D, "CyclinPred: a SVM-based method for predicting cyclin protein sequences", PloS one., Vol.3, No.7, (2008). https://doi.org/10.1371/journal.pone.0002605.

[11] Tsai CT, Huang WL, Ho SJ, Shu LS & Ho SY, "Virulent-GO: prediction of virulent proteins in bacterial pathogens utilizing gene ontology terms", Development, (2009).

[12] Matthews BW, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme", Biochimica ET Biophysica Acta (BBA)-Protein Structure, Vol.405, No.2, (1975), pp.442-451. https://doi.org/10.1016/0005-2795(75)90109-9.

[13] Morens DM, Folkers GK & Fauci AS, "The challenge of emerging and re-emerging infectious diseases", Nature, Vol.430, (2004), pp.242–249. https://doi.org/10.1038/nature02759.

[14] Pragash DS, Ragunathan L, Banoo S, Rayapu V & Shaker IA, "Occurrence of CTX-M and SHV Genes in ESBL Producing Gram Negative Organisms Causing Pyogenic Infections in a Tertiary Care Hospital in Puducherry", Int J Pharm Bio Sci, Vol.3, No.4, (2012).

[15] Toth IK, Pritchard L & Birch PR, "Comparative genomics reveals what makes an enterobacterial plant pathogen", Annu. Rev. Phytopathol., Vol.44, (2006). https://doi.org/10.1146/annurev.phyto.44.070505.143444.

[16] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT & Harris MA, "Gene Ontology: tool for the unification of biology", Nature genetics, Vol.25, No.1, (2000), pp.25-29. https://doi.org/10.1038/75556.

[17] Apweiler R, Bairoch A & Wu CH, "Protein sequence Databases", Curr. Opin Chem. Biol., Vol.8, (2004), pp.76-80. https://doi.org/10.1016/j.cbpa.2003.12.004.

[18] Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y & Jin Q, "VFDB: a reference database for bacterial virulence factors", Nucleic Acids Res., Vol.33, (2005).