International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

Prediction of Protein Toxicity by Analyzing DNA Sequence

Md. Mynul Hasan
Department of Computer Science & Engineering
Chittagong University of Engineering and Technology
Chittagong, Bangladesh

Dr. Muhammad Ibrahim Khan
Department of Computer Science & Engineering
Chittagong University of Engineering and Technology
Chittagong, Bangladesh

***Abstract*— Toxic Protein classification plays an important role in the analysis and prediction phases of drug designing task which is costly and time consuming in case of batch processing. Accurate prediction of toxic protein is an essential goal in bioinformatics because of the effect of protein toxicity in human body. This task is more challenging due to the variation in proteins as well as the lacking of distinct features supported by toxic protein sequence. In this paper, a machine learning based computational tool is proposed which can facilitate the automatic identification of rapid growing toxic sequences. A set of machine learning classifiers with various physical and chemical features have been used in our processed corpus, consisting of 55000 protein sequences as FASTA format where 38500 sequences used for training and 16500 sequences for testing purpose. The performance of the proposed tool is compared with different ML techniques including some existing techniques. The Random Forest Classifier with selected features provide a simple and consistent classification of toxic protein with the highest accuracy of 98%.**

Keywords—Bioinformatics, Protein Classification, Machine Learning, Toxic Protein.

I. INTRODUCTION

As proteins are the workhorse of a cell, they perform a vast array of functions within living organisms, including catalyzing metabolic reactions, replicating DNA, responding to stimuli etc. Proteins are large biological molecules consisting of one or more chains of amino acids. Protein involves most of the body's function and life processes. Each protein has its own unique amino acid sequence that is specified by the nucleotide sequence of the gene encoding this protein. Any type of irregularities in DNA sequence can impact on the function of a protein as protein sequences generate from DNA sequences. But this protein may become a curse when it turns into toxic one. Protein toxicity occurs when the body is unable to get rid of the potentially toxic wastes that are generated as a result of protein metabolism. In every disease of a human body there is a contribution of a specific protein. When any abnormalities occur in the body, a specific protein secreted in the affected region. The classification task of toxic and non-toxic protein identification is advantageous in various applications such as toxin annotation, drug designing, synthetic biology, healthcare etc. Machine learning based approaches can address the
challenges of toxin identification and provide effective solution to discover unknown features of toxins. In our proposed method, we adopt supervised machine learning approaches to predict toxic protein sequences.

II. RELATED RESEARCH

Over the past few decades, some researches have been done to detect malfunctioning protein. One of the bottlenecks for developing protein-based therapies to treat various diseases is protein toxicity. At present some methods to detect toxic protein from their amino acid sequence are available. There have been some specific machine learning based approaches proposed for toxin prediction. ToxinPred [2] tool uses in silico method for predicting only small peptide-based toxins. ToxClassifier[3] uses ensemble of classifiers with feature engineering including amino acid pair frequency, BLAST and HMMER based similarity score. ClanTox[4] is a web based application for ML classifier of small animal toxins. There exists a computational prediction method based on probabilistic measure. But this approach uses only a few features for detection and the accuracy is low [1]. Another approach [6] uses databases of toxic protein to compare a protein sequence among the declared toxic proteins whether there is a match. But this approach is not efficient in the case of a new protein which is not in the existing databases. There is also some method which detect specific toxic protein that is responsible for specific disease [5]. There are several online tools which only analyze the protein but they can't make any prediction [8]. We are detecting the whole group of toxic protein using supervised machine learning. We consider most of the parameters that can be calculated from the protein sequence. For the purpose of better performance we analyze four suitable supervised learning algorithms. The performance analysis shows that our method of detection using Random Forest[10] classifier provides the higher accuracy.

III. SYSTEM ARCHITECTURE

The system architecture depicts the procedure of the detection process. It is combined with two phases. The training phase represents how to train the learning algorithms using the features calculated from protein sequence. Test phase represents the testing process of an unknown protein sequence for the prediction. The combined system architecture is shown in Figure 1.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

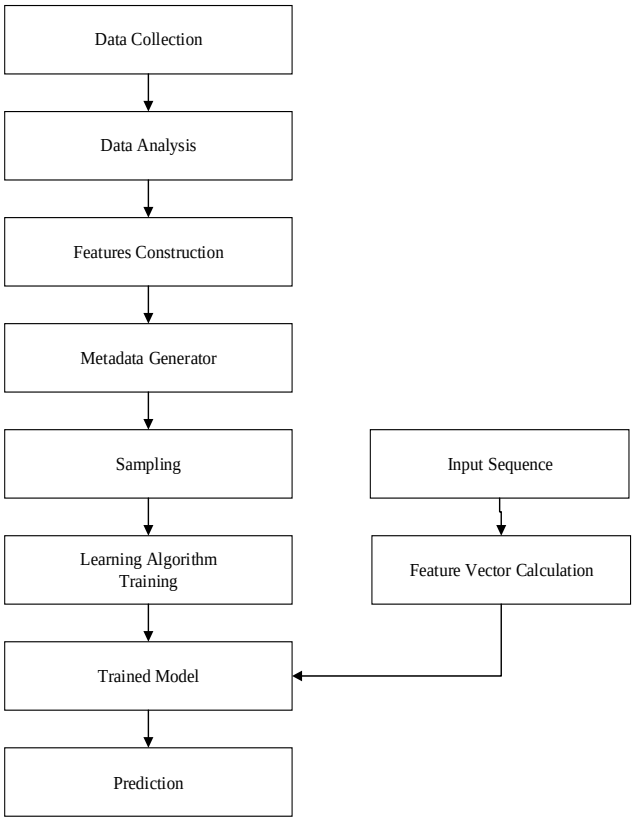


Figure 1: System Architecture of Toxic Protein detection

A. *Data Collection*

Data for training and testing of machine learning classifiers used in our prediction technique is obtained from NCBI database [7]. We collect approximate 55000 declared toxic and non-toxic protein sequences.

B. *Data Analysis*

We create two datasets for training and testing the classifier. Positive dataset includes toxic protein sequences and negative dataset includes non-toxic protein sequences. We remove all duplicate entries with identical sequence and sequence identifier. As the toxic protein are shorter in length, we ignore the sequences having length greater than 500. The result of the normalization process shows that 8,093 sequences in positive dataset and 47,144 sequences in negative dataset.

C. *Features Construction*

In this section, we calculate some parameters for all individual protein in the whole dataset. Information gain of each parameter is calculated using the value of entropy measure. According to calculated information gain fifteen parameters are selected as feature. These features are used to train and test the learning algorithms.

D. *Metadata Generation*

Metadata is a representation of data in a form of table in which row represent a feature vector for a protein and each column represent a distinct feature. We generate metadata for our raw data and store in a csv file. Metadata generator calculate all feature values for each sequence in dataset which is used to train the learning algorithms. The following fields were defined as belonging to the Standard Metadata set: Amino Acid, PI, Scale Value, Molecular Weight, Half-life, Aliphatic Index, Instability Index, Extinction Coefficient, Absorbance, Grand Average, Positive Residues, Negative Residues, Total Atoms, Atomic Composition, Peptide Composition and Class Value.

E. *Sampling*

We randomly split our raw data into a training and a test dataset. Training dataset is used to train the learning algorithm and test dataset is used to test the trained model to evaluate the performance of the model at very end. Finally we determine the suitable size of training and test dataset based on overfitting point. We used 70% data for training and 30% for testing.

F. *Learning Algorithm Training*

The training process involves providing a machine learning algorithm with training data to learn from and training parameter to control the learning algorithm. This training process takes metadata of training dataset as input by means of feature vector for learning and generate a model for prediction. We consider SVM, Decision Tree, Naïve Bayes and Random Forest as learning algorithms for training.

G. *Trained Model*

A model is prepared through a training process in which it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data. Four models are prepared and the suitable model is selected based on performance measures. This model is used to predict whether an input protein sequence is toxic or non-toxic.

H. *Input Sequence*

To predict a protein sequence as a toxic or non-toxic protein we need to input the sequence. The input sequence must be in FASTA format.

I. *Feature Vector Calculation*

We calculated all selected features for the input protein sequence using our tool. All features are saved in a vector format known as feature vector which is passed through the trained model for the purpose of prediction.

J. *Prediction*

This is the final phase of our system architecture. The trained model take the feature vector as input and predict the input protein as toxic or non-toxic based on learning experience.

IV. FEATURES OF DETECTION

A brief details and formula for calculation of some features such as Isoelectric point (PI), Aliphatic index, Instability index, Extinction coefficient, Grand average, Absorbance, Half-life etc. are described in this section.

A. *Amino Acid Composition*

The amino acid composition is defined as the fraction of each amino acid in a peptide and it can be calculated by the following equation:

$$\text{Comp}(i) = \frac{R_i}{N} \times 100 \qquad (1)$$

Where, Comp(i) is the percent

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

composition of amino acid (i); R_i is the numbers of residues of type i and N is the total number of residues in the peptide.

B. Dipeptide Composition

Dipeptide composition is advantageous over simple amino acid composition as it provides a composition of a pair of residues (e.g. Gly-Gly, Gly-Leu etc.) present in the peptide. Dipeptide composition can be calculated using the following formula:

$$\text{Dipeptide Comp (i)} = \frac{\text{Total number of Dipeptide (i)}}{\text{Total number of all possible dipeptides}} \quad (2)$$

Where, Dipeptide (i) is one out of 400 dipeptides.

C. Atomic Composition

Atomic composition of a sequence is determined by counting total number of Carbon, Hydrogen, Nitrogen, Oxygen and Sulfur in a given sequence.

D. Isoelectric point(pI)

Isoelectric point is a measure of pH in which net charge of a protein is zero. Proteins isoelectric point mostly depends on seven charged amino acids from more than 20 amino acids. These are glutamate (δ-carboxyl group), aspartate (ß-carboxyl group), cysteine (thiol group), tyrosine (phenol group), histidine (imidazole side chains), lysine (ε-ammonium group) and arginine (guanidinium group). Each of them has unique acid dissociation constant (pK). We used Henderson-Hasselbach equation to calculate protein charge.

For negative charged macromolecules:

$$\sum_{i=1}^{n} \frac{-1}{1 + 10^{pKn - pH}} \quad (3)$$

Where, pKn is acid dissociation constant of negative charged amino acid.

For positive charged macromolecules:

$$\sum_{i=1}^{n} \frac{1}{1 + 10^{pH - pKp}} \quad (4)$$

Where pKp is the acid dissociation constant of positive charged amino acid.

E. Aliphatic Index

The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine) [8]. It may be regarded as a positive factor for the increase of thermo stability of globular proteins. The aliphatic index of a protein is calculated according to the following formula:

$$AI = X(Ala) + a \times X(Val) + b \times (X(Ile) + X(Leu)) \quad (5)$$

Where, X(Ala), X(Val), X(Ile) and X(Leu) are mole percent (100 X mole fraction) of alanine, valine, isoleucine and leucine

a=relative volume of valine side chain (2.9)

b=relative volume of Leu/Ile side chain (3.9) to the side chain of alanine.

F. Instability Index (II)

Instability index is an estimate of the stability a protein in a test tube [9]. Statistical analysis of 12 unstable and 32 stable proteins has revealed that there are certain dipeptides, the occurrence of certain dipeptides is different in the unstable proteins compared with those in the stable ones. We use the weight value of instability to each of the 400 different dipeptides (DIWV). Using these weight values it is possible to compute instability index (II) which is defined as:

$$II = (10/L) \times \sum_{i=1}^{i=L-1} DIWV\,(x[i]x[i+1]) \quad (6)$$

Where L is the length of sequence.

DIWV(x[i]x[i+1]) is the instability weight value for the dipeptide starting in position i.

A protein whose instability index is smaller than 40 is predicted as stable, a value above 40 predicts that the protein may be unstable.

G. Half Life

Half of a proteins life time is known as protein half-life. It is a prediction of the time it takes for half of the amount of protein in a cell to disappear after its synthesis in the cell. We use N-end rule to determine the half-life of a protein [8]. N-terminal amino acid of a protein sequence determine the half-life according to this rule. We also use the half-life of all amino acids in a protein sequence.

H. Extinction Coefficient

The extinction coefficient indicates the amount of light a protein absorbs at a certain wavelength [6,8]. It is possible to estimate the molar extinction coefficient of a protein from its amino acid composition. From the molar extinction coefficient of tyrosine, tryptophan and cystine at a given wavelength, the extinction coefficient of protein can be computed as

$$E(Prot) = Numb(Tyr) \times Ext(Tyr) + Numb(Trp) \times Ext(Trp) + Numb(Cystine) \times Ext(Cystine) \quad (7)$$

Where (for proteins in water measured at 280 nm):

Ext(Try) = 1490, Ext(Trp) = 5500, Ext(Cystine) = 125

I. Grand Average

In short form we called it Gravy, a grand average of hydrophathicity. The GRAVY value for a peptide or protein is calculated as the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence[6].

$$Hv = Sum(N\text{-acid} \times scale.val) \quad (8)$$

$$Grav = Hv/n \quad (9)$$

Where Hv = Total hydropathy value, Scale.value = Hydropathy scale value for corresponding amino acid and n = Sequence length.

J. Absorbance

It is a measure of molar absorption of ultraviolet (UV) light of a protein sequence (in solution). We use Beer-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

Lambert law for calculating molar absorbance of a protein sequence.

Absorbance = E.L.c (10)

Where, E = Molar absorption coefficient, L = Length of the sequence and c = Concentration of the sequence.

V. Performance Analysis

Classifiers were tested using those protein sequences from positive and negative datasets that were not used in training phase. The following performance measures were used to evaluate the models on each of the datasets.

A. Confusion Matrix

The confusion matrix depicts the ways in which classification model is confused when it make predictions. The number of true and false predictions are summarized with count values and broken down by each class.

- *Number of True Positives (TP):* Number of toxic sequences correctly predicted as toxic.
- *Number of True Negatives (TN):* Number of non-toxic sequences correctly predicted as non-toxic.
- *Number of False Positives (FP):* Number of non-toxic sequences incorrectly classified as toxins.
- *Number of False Negatives (FN):* Number of toxic sequences incorrectly classified as non-toxic.

Table 1: Confusion matrix of Classifiers

Classifier	Confusion Matrix			
	TP	FP	FN	TN
SVM	2240	154	2793	11284
DT	2232	162	2034	12043
NB	2135	259	1870	12207
RF	2263	131	119	13958

B. Accuracy (ACC)

Accuracy indicates the proportion of correctly predicted sequences including both toxic and non-toxic. The following formula is used to calculate the accuracy measure.

ACC= (TP + TN)/(TP + TN + FP + FN) (11)

C. Classification Report

Classification report generates a report to provide a quick idea about the model accuracy using some measures. This report shows the precision, recall, f1-score and support for each class.

- *Precision:* Precision is the number of True Positives divided by the number of True Positives and False Positives. It is also called the Positive Predicted Value (PPV). Precision is calculated as

  PPV = TP/(TP + FP) (12)
- *Recall:* Recall is the number of True Positives divided by the number of True Positives and the number of False Negatives. It is also called Sensitivity or the True Positive Rate. It is calculated using the formula:

  SENS = TP/(TP + FN) (13)
- *F1 Score:* F1 Score is harmonic mean of precision and sensitivity and represents weighted average of precision and recall. It is also called the F Score or the F Measure. It is calculated as

  F1 = (2× PPV× SENS)/(PPV + SENS) (14)
- *Support:* Support indicates the total number of data of positive and negative dataset which are used for testing.

D. Specificity (SPEC)

Specificity means the proportion of correctly predicted non-toxins (true negatives). It is also called true negative rate, which is calculated as

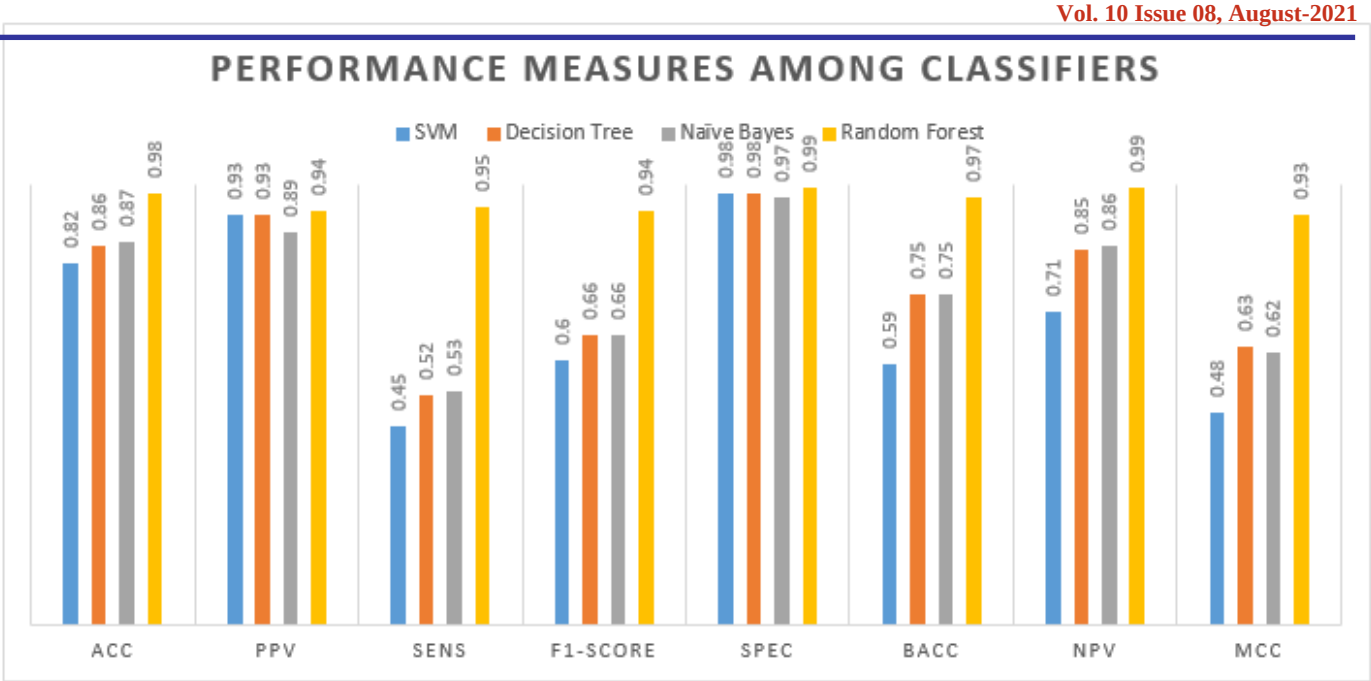
SPEC = TN/(TN + FP) (15)

E. Balanced Accuracy (BACC)

Balanced accuracy is a mean value of specificity and sensitivity which is calculated using the following formula:

BACC = (SPEC + SENS)/2 (16)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021



F. Negative Predicted Value (NPV)

NPV is the proportion of negatives that are true negatives. It is calculated using the following formula:

$$NPV = TN/(TN + FN) \quad (17)$$

G. Matthew's Correlation Coefficient (MCC)

The MCC value represents a correlation measure between predicted and observed.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (18)$$

VI. CONCLUSION

We have developed a toxic protein prediction scheme using supervised machine learning techniques. At first we extracted some features for each protein sequence in positive and .negative dataset. Machine learning models are generated by training four learning algorithms (SVM, Decision Tree, Naive Bayes, Random Forest). After some performance analysis Random Forest is selected as suitable model which gives the highest accuracy for the prediction of toxic protein. Accuracy may be improved by considering structural features along with physical and chemical parameters. Also handling the problem of imbalanced dataset and incorporating deep learning techniques can be used to enhance the performance of toxic protein detection scheme.

REFERENCES

[1] M.A. Rahman and M.I. Khan, "Computational Prediction of Toxic Protein", 9th International Forum on Strategic Technology (IFOST), Cox's Bazar, 2014, pp. 499-502.

[2] S. Gupta, P. Kapoor, K. Chaudhary and G.P. Sing, "In silico approach for predicting toxicity of peptides and proteins", PLoS One, 2013, pp.18

[3] R. Gacesa, D. J. Barlow, and P. F. Long. "Machine learning can differentiate venom toxins from other proteins having non-toxic physiological functions". In: PeerJ Computer Science 2 (2016), e90.

[4] G. Naamati, M. Askenazi, and M. Linial. "ClanTox: a classifier of short animal toxins". In: Nucleic acids research 37.suppl_2 (2009), W363–W368.

[5] C. M. Yang, "A striking similarity between proteins involved in the early stages of Alzheimer's disease and mad cow disease", 220th national meeting of the American Chemical Society, LA, USA, 2000.

[6] J. Cui, Q. Liul, D. Puett and Y. Xul, "Computational prediction of human proteins that can be secreted into the bloodstream", 2nd IEEE International conference, 2012, pp. 230-236.

[7] National Center for Biotechnology Information, "Protein database: ToxicProtein".[Online].Availible:http://www.ncbi.nlm.nih.gov/protein/?term=toxic+protein [Accessed: May 2020]

[8] Expasy, Bioinformatics Resource Portal, "Protparam Tool," SIB Bioinformatics Resource Portal. [Online]. Availible: http://web.expasy.org/protparam/ [Accessed: October 2016]

[9] H. Seker and P. I. Haris, "Predicting a proteins melting temperature from its amino acid sequence", Annu. International Conference of the IEEE Engineering in Medicine and Biology, Buenos Aires, 2010, pp. 1820-1823.

[10] A. Liaw and M. Wiener, "Classification and Regression by random forest", 19th International Multi-Topic Conference, 2016, pp. 24-31.