EUROPEAN JOURNAL OF BIOMEDICAL AND PHARMACEUTICAL SCIENCES
http://www.ejbps.com

ISSN 2349-8870
Volume: 4
Issue: 11
514-519
Year: 2017

AGPRED: ANTIGEN PREDICTION USING NEURAL NETWORK BASED ON TRI-PEPTIDE MARKERS

Himanshu S. Mazumdar[1]* and Ragini V. Oza[2]

[1]Professor and Head, Research and Development Center, Dharmsinh Desai University, Gujarat, India.
[2]Assistant Professor, Deparment of MCA, Parul University, Gujarat, India.

***Corresponding Author: Himanshu S. Mazumdar**
Professor and Head, Research and Development Center, Dharmsinh Desai University, Gujarat, India.

Article Received on 16/09/2017 **Article Revised on 06/10/2017** **Article Accepted on 26/10/2017**

ABSTRACT
The protein sequence plays an important role to understand the function and feature of protein. Antigen prediction from the huge amount of protein primary sequence is a challenging problem. A novel approach is proposed here to characterize an antigen sequence using a set of features which can describe characteristics of target antigen group. The proposed system uses a combination of evolutionary algorithm and proposed ordering algorithm to identify the set features. Here the features are the tri-peptides. An algorithm ensures that the unique combination of the tri-peptide separates target antigen sequence from other protein sequences. We have preprocessed the datasets of *Plasmodium Falciparum, Leptospira Interrogans*, *Pseudomonas Aeruginosa*, *Streptococcus Pneumonia* and *Bacillus Thuringiensis* to use it in our system. These datasets are extracted from Uniref100 protein sequence database which contains 83 million records. Neural networks are trained using training sets from all species and results are compared here. A prediction result gives 98 % of peak accuracy for *P.Falciparum* using identified tri-peptide features which are tested on the test dataset.

KEYWORDS: *Plasmodium Falciparum*; Tri-peptide Residue; Occurrence Frequency; Population Ratio; Genetic Algorithm; Back-propagation Neural Network.

1. INTRODUCTION
Some major challenges in bioinformatics are 1. A huge data environment 2. Heuristic search 3. Knowledgebase operations. In order to full fill above requirements, protein sequence analysis is performed. Recent Advancement in Protein sequence analysis creates enormous accessible information.[1] Protein sequence data contain intrinsic dependencies between their constituent elements, the task of the researcher is to find the dependencies between neighboring elements by generating all the contiguous (potentially overlapping) sub-sequences of a certain length(n-residue words)[2] and predict functional and auxiliary comparability of proteins using this n-residue words. As a rule, useful comment of proteins is exceedingly dependent on grouping comparability to other known proteins.[1] Antigen sequence prediction is one of the challenges in protein sequence analysis.

Machine learning approaches to predict epitope in protein sequence using Decision tree and the nearest neighbour classifier is compared by Johannes and Bernd Mayer[3], which reported an accuracy of 72%. Similarly, Yasser et al.[4] have reported 75% accuracy using Support Vector Machine (SVM) for prediction of epitope based on string kernels. A proposed approach to predict epitope using three methods quantitative matrix (QM), Support Vector Machine (SVM) and Artificial Neural Network (ANN) and compared results by Manoj and Raghava.[5] Li Li et al.[6] has shown an approach to predict a microarray data using a hybrid method of GA and k-nearest neighbours which state that a key feature genes can be accurately identified using proposed approach.

To perform a similarity search in large protein sequence database a novel approach is proposed by the research team of R&D Center, DDU in.[7] This concept further extended for multi-sequence alignment in.[8] The tool for prediction of protein secondary structure using 5 residue words and the neural network is developed by our research team.[9] The identification and characterization of Antigen (Ag) sequence play an important role in diagnostic tests, vaccine design, and antibody production. Therefore, computational tools for reliably predicting Ag sequence are highly desirable. We evaluated 250 tri-peptide using Genetic Algorithm and trained a back propagation neural network for 66,662 *P. Falciparum* Ag sequences in our earlier approach.[11] Based on the results of our previous computational experiments[11], we propose AgPred, a novel method for predicting Ag sequence using the peptide sequence. We showed that the predictive performance of AgPred (AUC

= 0.98) outperforms our previously evaluated method[11] (AUC = 0.93). Furthermore, we also tested AgPred on other species, *Leptospira Interrogans*, *Pseudomonas Aeruginosa*, *Streptococcus Pneumonia* and *Bacillus Thuringiensis*. Analysis of the used datasets and the results of this comparison shows conclusion about the relative performance.

The paper is organized as follows: the framework of proposed method and resources used are discussed in section 2 and 3. Section 4 and 5 describes experimental results and discussion on parameters considered, respectively. Finally, we concluded the discussion in Section 6.

2. MATERIALS AND METHODS

As this paper focuses on the prediction of five different antigen group, a general flow is designed to extract tri-peptides from target antigen group and train a back propagation neural network. Fig. 1 represents methodology to extract tri-peptide for any target antigen group.

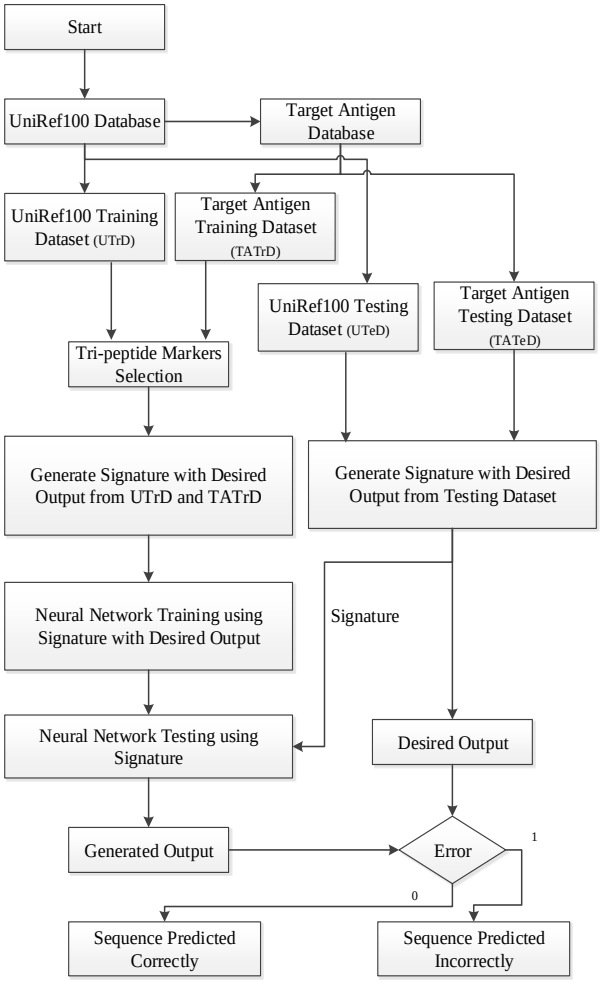

Figure 1: General flow of Antigen Prediction system.

2.1. Dataset Collection

UniRef100 database combines identical sequences and sub-fragments from any source organism into a single UniRef entry.[10] All the small datasets for *Plasmodium Falciparum*, *Leptospira Interrogans*, *Pseudomonas Aeruginosa*, *Streptococcus Pneumonia* and *Bacillus Thuringiensis* are extracted from UniRef100. Table 1 shows the description of all extracted databases.

Table 1: Database for different species.

Species Name	File Name	Database Name	No of Records	Size
Plasmodium Falciparum	FalciparumDB.txt	PFDb	66662	54.42 Mb
Leptospira Interrogans	InterrogansDB.txt	LIDb	68643	27.88 Mb
Pseudomonas Aeruginosa	AeruginosaDB.txt	PADb	218816	113.06 Mb
Streptococcus Pneumonia	StreptococcusDB.txt	SPDb	242868	110.9 Mb
Bacillus Thuringiensis	ThuringiensisDB.txt	BTDb	80175	33.34 Mb

All the databases are further separated in training and testing datasets. Each training dataset contains 50×10^3 records randomly populated from its respective database and testing datasets contain 16,662 records, here dataset size is restricted to a minimum of all datasets. Ten different datasets of a protein sequence are generated from UniRef100 for training purpose, each dataset contains 50×10^3 random records in training. Testing set of non-antigen sequence is also extracted from UniRef100, no care is taken to exclude antigen from non-antigen training and testing dataset as the occurrence frequency is below 2%.

2.2. Tri-peptide marker selection

This section describes the process of selecting tri-peptide markers which are performed in three phase. The first phase includes the use of Evolutionary Algorithm to select 1000 tri-peptides based on dominating presence in

target antigen group. The second phase continues the process of selecting 500 tri-peptides from target antigen group using Iterative Convergence Algorithm. And finally, in the third phase, this 1500 tri-peptides are sorted based on their pair frequency distribution using proposed ordering algorithm. The detail description of each phase is described below.

2.3. Evolutionary Algorithm

In this phase, tri-peptide are extracted from target antigen training dataset as well as from UniRef100 training dataset. These tri-peptide markers are in order of occurrence frequency in its respective dataset. Now, Proposed Algorithm uses this two tri-peptide list to evaluate the tri-peptides based on dominating presence in target antigen training dataset w.r.t UniRef100 training dataset. It is achieved by updating a counter for every tri-peptide in a list, this counter is increased by one if the occurrence of tri-peptide is found in the sequence of target antigen training dataset and decreased by one if it occurs in a sequence of UniRef100 training dataset. Below Algorithm 1 describes steps to extract tri-peptides.

Algorithm 1 Evolutionary Algorithm

1. **Input:** OFa - Occurrence frequency of tri-peptide in target antigen training dataset,
   OFu - Occurrence frequency of tri-peptide in Uniref100 training dataset.
2. **Output:** List of sorted 1000 tri-peptide
3. Calculate ratio between OFa and OFu
4. Select 1000 tri-peptide based on ratio
5. Create counter for every 1000 tri-peptides and initialize its value to zero.
6. **for** i = every sequence of the target antigen training dataset **do:**
7. **for j=0 to 999 do:**
8. **if (** j^{th} tri-peptide is present in i^{th} sequence**):**
9. Increment counter of j^{th} tri-peptide by one
10. **end if**
11. **end for**
12. **end for**
13. **for** k = every sequence of the Uniref100 training dataset **do:**
14. **for j=0 to 999 do:**
15. **if (** j^{th} tri-peptide is present in k^{th} sequence**):**
16. Decrement counter of j^{th} tri-peptide by one
17. **end if**
18. **end for**
19. **end for**
20. Sort 1000 tri-peptide list according to the counter value of tri-peptide.

2.4. Iterative Convergence Algorithm

In this phase, a list of tri-peptide is evaluated further using proposed iterative convergence algorithm in.[12] Top 100 tri-peptide are selected from the list of 1000 tri-peptides and their presence is checked in the target antigen training dataset. The exception list is generated from training antigen sequence which does not have the presence of any tri-peptide from selected 100 tri-peptides. Now, new tri-peptides are generated from exception list and the ratio of tri-peptide frequency from the exception list and UniRef100 training dataset is calculated. After sorting tri-peptide based on the ratio, top 500 tri-peptides are selected and added to 1000 tri-peptide list.

2.5. Ordering Algorithm

A new approach is proposed to order the tri-peptides based on their pair frequency. Here, the pair frequency means a number of times the same pair of tri-peptides occurs in the different sequences. Most frequently co-occurring tri-peptides has a higher order in an ordered list. Below Algorithm 2 describes steps to order the list.

Algorithm 2 Ordering Algorithm

1. **Input:** 1500 tri-peptide list
2. **Output:** Ordered list of 1500 tri-peptide
3. Make every possible pair of tri-peptides and find occurrence frequency in antigen training dataset
4. sumMin = Maximum Integer value
5. error = 0
6. **for i=0 to 1499 do**
7. **for j=i+1 to 10 do:** // Finding minima at local region
8. **for k=0 to 1499 do**
9. sum = sum + (frequency of pair having i^{th} tri-peptide and k^{th} tri-peptide - frequency of pair having j^{th} tri-peptide and k^{th} tri-peptide)
10. **end for**
11. Find minima of sum from every pair
12. **end for**
13. **if (minima tri-peptide is not a neighbour of** i^{th} **tri-peptide)**
14. error++
15. swap $(i+1)^{th}$ tri-peptide with minima tri-peptide in a list
16. **end if**
17. **end for**
18. **if(error is not zero)**
19. repeat steps 5 to 20
20. **end if**

Uniqueness and occurrence are contradicting requirement for an entity (tri-peptide), To start with, we increase the occurrence with uniqueness using diagonal metrics occurrence of entities with random neighbours. Swapping of tri-peptide in a group of 10 is performed iteratively so that similar neighbour along with its occurrence are closer to each other. The subset of the consecutive cluster represents the random distribution of tri-peptide set. The tri-peptides are shifted till the end to orderly sort the high occurring unique tri-peptide pair to the left in clusters. This ordered set is used to select top 200 tri-peptide markers. These tri-peptides markers are selected from clusters of 10 tri-peptide (one tri-peptide from each cluster) in an ordered list. Finally, selected 200 tri-peptides has following properties, a. dominating presence in the target antigen training set, b.

Omnipresence in rare sequence also and c. High pairing occurrence with other tri-peptides.

2.6. Neural Network

This section describes the steps to create input signature for a neural network using selected 200 tri-peptide markers. The presence of tri-peptide marker in a sequence is denoted as 1 and absence is denoted as 0. Hence, the input binary string is generated of length 200, this input signature is created for every sequence in training set for both antigen and non-antigen dataset which represents desired output during training. Back propagation neural network is trained here using binary input signature. Fig. 2 shows the architecture of neural network with 30 hidden nodes. Here, the output of hidden nodes is calculated by equation(1) and final output at the output node is given by equation (2).

$$X_j = \sum(Y_i * W_{ji})$$
$$Y_j = 1/1+e^{-Xj} \quad (1)$$

Where Y_i is the output of input layer, X_j is input to the hidden layer and Y_j is the output of hidden layer.

$$X_k = \sum(Y_j * W_{kj})$$
$$Y_k = 1/1+e^{-Xk} \quad (2)$$

Where Y_j is the output of hidden layer, X_k is input to the output layer and Y_k is the generated output at the output layer.

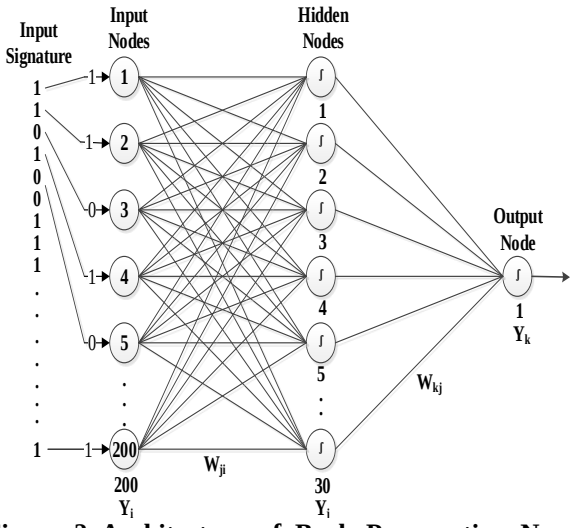

Figure 2 Architecture of Back Propagation Neural Network.

The neural network is trained using backpropagation algorithm[11] using training dataset. The training is terminated when the error of desired output and actual output average falls below 3%. The network is tested using equation 1 and 2 with test dataset as describe above.

3. RESULTS AND DISCUSSIONS

In this section, the proposed tri-peptide marker extraction method is applied to five different datasets, as mentioned in section 2.1 to prove its correctness and effectiveness. The proposed method is used to generate input signature from training dataset as described in section 3. Finally, the input signature is fed to the neural network for the training purpose for individual species. The trained neural network is tested on a test dataset for respective species. Following Table 2 shows the accuracy obtained for each dataset. Here the performance measures are the sensitivity and specificity.

The highest occurring tri-peptide pair from ordering list should be present on left-hand side of the list. In Fig. 3, Y- axis represents the number of sequences in which set of tri-peptides are not present and X-axis represents set of tri-peptides from ordered list in an order from left to right. It is clearly observed from the Figure that the tri-peptide which are towards the left side of the list has very less number of sequences which are not recognized. This indicates the tri-peptide towards left are highest occurring tri-peptides.

Table 2: Sensitivity, Specificity and Accuracy on all datasets for different values of hidden neurons.

Datasets	Hidden Neurons	Sensitivity	Specificity	Accuracy
PFDb	15	92.12	93.02	92.57
	20	93.9	95.2	94.55
	30	97.7	98.4	98.05
LIDb	15	91.65	92.5	92.075
	20	94.8	96	95.4
	30	96.3	97.4	96.85
PADb	15	91.8	92.6	92.2

	20	94.8	95.1	94.95
	30	96.5	97.9	97.2
SPDb	15	92.6	93.9	93.25
	20	94.06	95.2	94.63
	30	95.5	96.9	96.2
BTDb	15	91.2	92.95	92.075
	20	93.35	94.06	93.705
	30	96.01	97.5	96.755

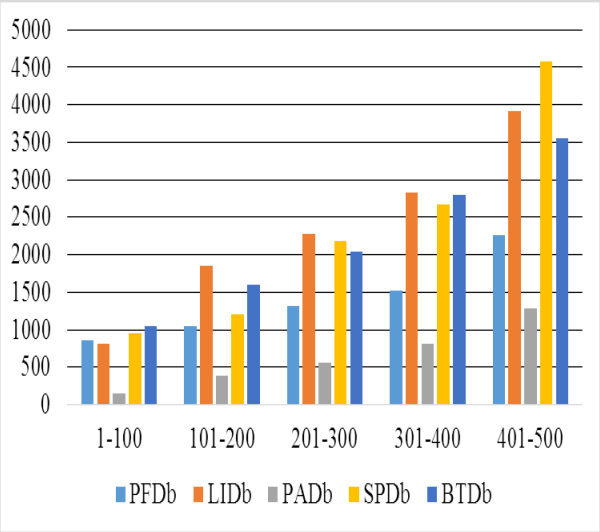

Figure 3: Goodness of ordering algorithm which increases from left to right after ordering.

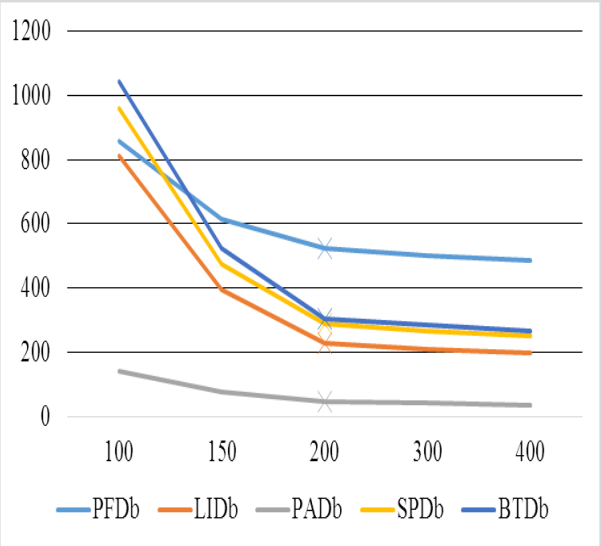

Figure 4 Selection of optimal set of tri-peptide.

The challenging task is to select the number of tri-peptide from ordered list. The coverage of set of tri-peptides in target antigen training dataset with different numbers of tri-peptides is checked. Fig. 4 represents the number of sequences which are not recognized by a set of tri-peptide. It was observed that as we increase the size of the set of tri-peptides, the exception sequences are also decreased noticeably for all five datasets but not much improvement is noticed for more than 200 tri-peptides. Thus we have selected 200 tri-peptide as an optimal number of tri-peptide.

4. CONCLUSION

We have proposed an evolutionary ordering algorithm that optimally extracts the tri-peptide markers for popular target antigen group. The evolutionary algorithm ensures that extracted tri-peptide has dominating presence in target antigen dataset w.r.t. non-antigen dataset. Ordering algorithm orders the tri-peptide in pairs based on their goodness. Pairs are found much more unique than individual and are homogeneously distributed among the antigen group. The tri-peptide pairing technique and its distribution is found an effective method for a wide range of a target antigen group. The proposed approach is found upto 98% accurate to predict the unknown sequence as target antigen or non-target antigen using neural network.

5. ACKNOWLEDGMENT

The research reported in this paper was performed as part of the project, guided and supported by Department of Science and Technology (DST) at R&D Center, DDU.

REFERENCE

1 Cornelia Caragea, Adrian Silvescu, Prasenjit Mitra “Protein Sequence Classification using feature hashing” Proteome Science, 2012; 10(1): S14.

2 Zhao-Hui Qi, Meng-Zhe Jin1, Hong Yang “A Measure of Protein Sequence Characteristics Based on the Frequency and the Position Entropy of Existing K -words” MATCH Commun. Math. Comput. Chem., 2015; 73: 731-748.

3 Johannes Sollner and Bernd Mayer. "Machine learning approaches for prediction of linear B-cell epitopes on proteins." Journal of Molecular Recognition, 2006; 19.3: 200-208.

4 Yasser EL-Manzalawy, Drena Dobbs, and Vasant Honavar. "Predicting linear B-cell epitopes using string kernels." Journal of molecular recognition, 2008; 21.4: 243-255.

5 Bhasin, Manoj and G. P. S. Raghava. "Prediction of CTL epitopes using QM, SVM and ANN techniques." Vaccine, 2004; 22.23: 3195-3204.

6 Li Li, Wei Jiang, Xia Li et al. "A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset." Genomics, 2005; 85.1: 16-23.

7 Maulika S. Patel and Himanshu S. Mazumdar. "Similarity search using pre-search in UniRef100 database." International Journal of Hybrid Information Technology, 2011; 4: 3.

8 Maulika S. Patel and Himanshu S. Mazumdar. "Knowledge base and neural network approach for protein secondary structure prediction." Journal of theoretical biology, 2014; 361: 182-189.

9 Himanshu S. Mazumdar, Ankita C. Baravaliya, and Maulika S. Patel. "Keyword based Iterative Approach to Multiple Sequence Alignment." Int. J. Pure App. Biosci, 2014; 2(3): 139-144.

10 UniRef100.fasta Database, Retrieved from: http://www.uniprot.org/downloads, [Downloaded on:20-June-2016].

11 R. Hecht-Nielsen, "Theory of the backpropagation neural network," International 1989 Joint Conference on Neural Networks, Washington, DC, USA, 1989; 1: 593-605.

12 Ragini V. Oza and Himanshu S. Mazumdar, “Peptide Markers based Prediction of Antigen Sequence using Neural Network”, is published in of IJPAB International Journal of Pure and Applied Bioscience, Jan. -Feb. 2017; 5(1): 771-776.