Available online at www.ijpab.com

DOI: http://dx.doi.org/10.18782/2320-7051.2586

ISSN: 2320 – 7051

Int. J. Pure App. Biosci. **5 (1):** 759-770 (2017)

Research Article

International Journal of Pure & Applied Bioscience

OPEN ACCESS

Peptide Markers based Prediction of Antigen Sequence using Neural Network

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

[1]Student (M. Tech.), Information Technology, Dharmsinh Desai University, Gujarat, India

[2]Professor & Head, Research & Development Center, Dharmsinh Desai University, Gujarat, India

*Corresponding Author E-mail: raginioza1206@gmail.com

Received: 13.02.2017 | Revised: 25.02.2017 | Accepted: 26.02.2017

ABSTRACT

Bioinformatics has witnessed considerable progression in recent years; the prediction of antigen sequence in big data environment still remains challenging. A novel approach is proposed here to generate and evaluate tri-peptide markers, where a combination of high frequency tri-peptides can signify a characteristic of target antigen sequence. A dataset of Plasmodium falciparum antigen sequences is extracted from benchmark uniref100 protein sequence database; Training and test set are generated from extracted P. falciparum dataset. Genetic Algorithm (GA) is used here to identify an optimal set of tri-peptide markers from training set. Through different generations of GA, markers are evaluated using approximate selection function. A total 100 tri-peptides are identified using GA and the rest 150 are extracted by examining fitness function using iterative convergence algorithm. A back propagation neural network is trained to predict target antigen sequences using selected tri-peptide markers. The algorithm is tested on a test set which is non-inclusive in training set and the prediction result obtained shows 93% accuracy. This algorithm can also be useful to synthesis new sequence as possible drug antigen for given target protein.

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

INTRODUCTION

Antigen (Ag) is a foreign substance that enters in human body and persuades an immune system, which affects the production of antibodies (Abs); in other words Ag causes production of Abs against itself in immune system[20,21]. *P. falciparum* is protozoan parasite that causes malaria in human body. Since the protein is the candidate for parasite Ag it can be used to design vaccine against parasite Ag[2]. *P. falciparum* proteins are immunogenic according to analysis of Katarzyna *et al*[3]. This *P. falciparum* causes the most dangerous form of diseases like malignant malaria[1]. In 2015, World Health Organization has reported that there are around 214 million cases of malaria worldwide which resulted into 438000 deaths[4].

Cite this article: Oza, R.V. and Mazumdar, H.S., Peptide Markers based Prediction of Antigen Sequence using Neural Network, *Int. J. Pure App. Biosci.* **5(1):** 759-770 (2017). doi: http://dx.doi.org/10.18782/2320-7051.2586

Over 90 % of the malarial death is caused by *P. falciparum*[23]. Recent research in vaccine design for liver stage *P. falciparum* Ag reported >90 % immunity based on genomic sequence by Joao Aguiar *et al*.[5] which stats that an effective development of drug for malaria is possible since there is no Food & Drug Administration (FDA) approved drug available[5]. Jack Richards *et al*. shows an approach to identify biomarkers for Ag specific response to immunity[6]. A polypeptide marker sequence can be used to identify protein sequence is represented by Thomas *et al*[8].

The Genetic Algorithm (GA), as described by Jihoon Yang and Vasant Honavar[7], performs optimal feature subset selection for optimization problem. Individuals in current population represents features which can be selected for next generation by selection function based on its fitness calculated by fitness function. GA is a distinctive approach from other statistical methods, which performs parallel population based randomized search in objective to optimize a solution as stated by Goldberg and Holland[9]. Jane Liu *et al*[10]., have shown how GA can be used to generate predictive features as biomarker of tumour automatically. Deutsch[11] has also used evolutionary algorithm to find an optimal set of marker genes to diagnose a specific type of cancer.

Machine learning approaches to predict epitope in protein sequence using Decision tree and nearest neighbour classifier is compared by Johannes and Bernd Mayer[12], which reported an accuracy of 72%. Similarly Yasser *et al*[13]., has 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 a results by Manoj and Raghava[14]. Li Li *et al*[15]., has shown an approach to predict a microarray data using a hybrid method of GA and k-nearest neighbours which states 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 research team of R&D Center, DDU, which index the 15 residue words of protein sequence and provides parallelism for searching in huge database as described in[16]. This concept further extended for multi sequence alignment in[17]. The tool for prediction of protein secondary structure using 5 residue words and neural network is developed by our research team[18]. As discussed earlier it is necessary to diagnose a *P. falciparum*, which can cause a most dangerous diseases for human body, for which accurate prediction of Ag sequence is desirable and which can be used to design a new drug for such disease. This encouraged us to propose such algorithm which can accurately predict target Ag sequence in a large dataset. A proposed approach reports about 93% accuracy.

MATERIALS AND METHODS

This research aims to predict Ag sequence of *P. falciparum* based on set of peptide (n-residue) markers, which are selected using GA approach and prediction using Backpropagation Neural Network (BPNN) is discussed in this work. Identification of an optimal set of peptide markers is very challenging task. Initially the length of peptide marker was considered as 5 residue word but based on statastic shown in section 5, three residue words (tri-peptide) were selected. These statistics are generated based on distribution of peptide frequency in *P. falciparum* and UniRef100[22] protein sequence database. A datasets are separated in training and testing *P. falciparum* dataset and 10 different random uniref training datasets as discussed in section 2.1. An approach to optimize set of tri-peptide marker using GA is discussed in section 2.2, in which set of tri peptide is evaluated based on fitness function at every generation and fitness of marker is decided by frequency distribution in *P. falciparum* and uniref training dataset. In our approach selection is done based on probability of survival of peptide marker in

maximum generations which distinguish our approach from other traditional GA approach. Total 250 tri-peptide markers are evaluated which has dominating presence in *P. falciparum* training dataset. Different parameters like number of hidden neurons, avoid overtraining etc. for BPNN is discussed in section 3. Following figure shows general flow of our method.

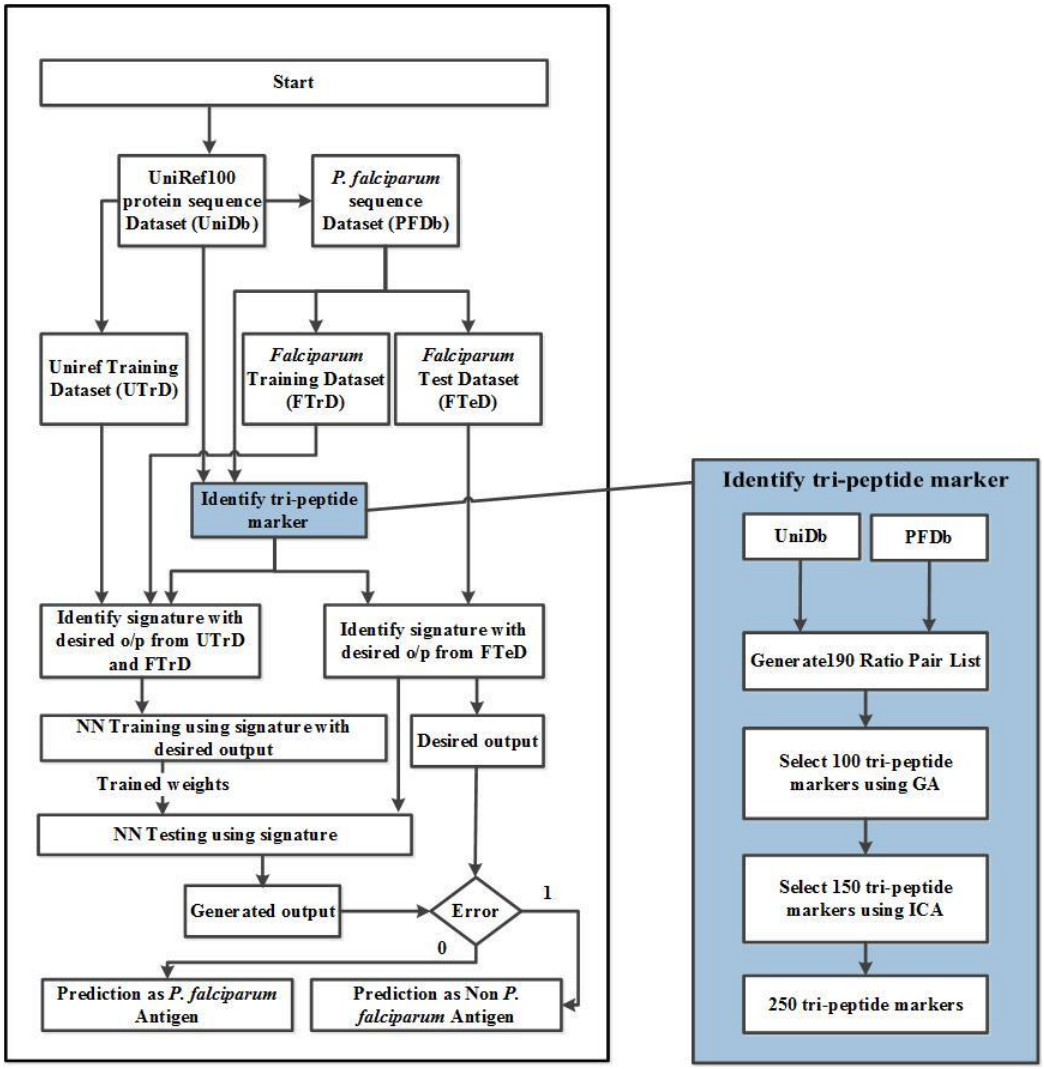

Fig. 1 (a) **Fig. 1 (b)**

Fig. 1 (a) Steps for NN based Antigen (Ag) prediction system. **Fig. 1 (b)** Steps to identify tri-peptide marker set (TPMS). UniDb: UniRef100 Benchmark Dataset. PFDb: *Plasmodium Falciparum* Dataset. FTrD: *Falciparum* Training Dataset. UTrD: Uniref Training Dataset. FTeD: *Falciparum* Testing Dataset.

Database

A benchmark database of protein sequence, UniRef100 (UniDb) is downloaded from Universal Protein Resource[22]. UniDb is clustered database in FASTA format, which has header and sequence as a single record. This dataset is used in earlier work conducted at R&D Center, DDU[16,17,18]. *Plasmodium Falciparum* Database (PFDb) is created by extracting sequences from UniDb by accessing cluster name in header of records. PFDb has approximately 67×10^3 sequences which is further ramdomly divided in *Falciparum* training dataset (FTrD) and *Falciparum* Testing Dataset (FTeD) contains 50×10^3 and 17×10^3 respectively. Ten different set of Uniref Training Dataset (UTrD) are generated from UniDb which contains random 50×10^3 different sequences in each dataset. Tri-peptide markers are extracted from FTrD and UTrD datasets and tested in FTeD dataset.

Table 1: Datasets used for prediction of *P. falciparum* Ag sequences

S.No	Name of DataBase (Fasta Format)	No. of Records	Downloaded/ Created Date	Source
1	UniRef100.fasta	83,050,155	20/06/2016	Universal Protein Resource [21]
2	FalciparumDB.txt	66,662	29/09/2016	Generated at R&D, DDU from Uniref100.fasta
3	*Falciparum* Training Dataset (FTrD)	50,000	02/10/2016	Generated at R&D, DDU from FalciparumDB.txt
4	*Falciparum* Testing Dataset (FTeD)	16,662	02/10/2016	Generated at R&D, DDU from FalciparumDB.txt
5	Uniref Training Dataset (UTrD)	10 sets of 50,000 records	15/10/2016	Generated at R&D, DDU from Uniref100.fasta

Tri-Peptide Marker Selection

Tri-peptide marker selection process is described in this section. This process is divided in three steps (a) 190 ratio pair list generation (b) Select 100 tri-peptide markers using GA (c) Rest 150 using iterative convergence algorithm as explained in following sections.

190 Ratio Pair List Generation

The selected tri-peptide marker should have higher occurance in PFDb and low in UniDb. Thus ratio of tri-peptide is calculated between occurance frequency in PFDb and UniDb. It is highly desirable that markers in selected set should have higher ratio as well as minimum co-occurance (markers which occurs in same records) in training dataset. Markers which occurs in different records in dataset is consider as orthogonal markers. To achieve an orthogonality in selected set, we distributed every tri-peptide in to 190 different sets. Tri-peptide in each set contains a unique amino acid pair of length 2 such as AC, AD, CD, DL etc. There are total 190 pairs possible from 20 amino acids. We followed this procedure for both Tri-peptide list of UniDb and PFDb, and generated two sets of 190 pair list. From these two sets of pair list we created another pair list which has 190 pairs and each pair contains tri-peptides which are common in both pair list with their occurrence frequency ratio (*PFDb occurance frequency : UniDb occurance frequency*). Figure 2 shows the flow to generate 190 ratio pair list.

Select 100 Tri-Peptide Markers using GA

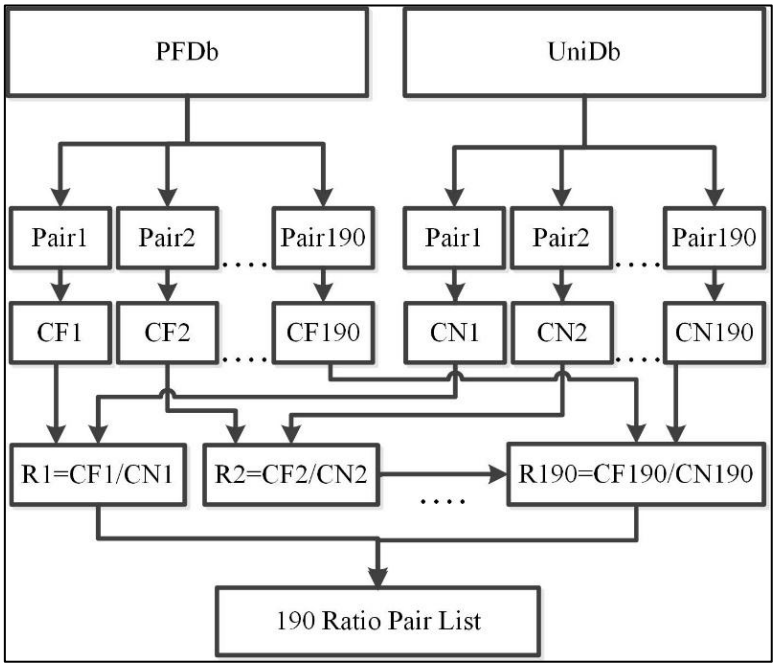

Fig. 2: Flow chart to Generate 190 Ratio Pair List. CF_i: occuurance frequency of tri-peptide in PFDb, CU_i: occurance frequency of tri-peptides in UniDb, $R_i = CF_i / CU_i$: ratio pair of occurance frequency of particular tri-peptide in ith pair; where i ranges from 1 to 190

In this section selection procedure for 100 tri-peptide markers is discussed as in Figure 3. In every generation of GA, UTrD is generated randomly by selecting 50×10^3 records from UniDb to provide robustness in evaluating markers. FTrD is same throughout every generation as PFDb has very few records as compared to UniDb. Occurance frequence of tri-peptide in FTrD (CF) and Occurance frequence of tri-peptide in UTrD (CU_g) generated at every generation. Ratio of CF and CU_g is denoted by RG_i (Ratio of Generation). Generation-wise tri-Peptide List (GPL) is the list of most occuring marker through out generations.

Different paramenters of Genetic Algorithm is as shown below,

Individual: tri-peptide marker

Population: 190 ratio pair list is considered as population for tri-peptide markers. 100 tri-peptides are selected randomly from 190 ratio pair list for the first generation of GA cosidering only one peptide from one pair of 190 ratio pair list.

Fitness Function: Fitness function for the survival of marker is that occurrence frequency should be dominating in FTrD dataset. Thus tri-peptide with higher ratio has chance to survive in that generation.

Tournament Selection: Individuals selected for next generation follows the tournament selection criteria where individuals should be in top 50 marker based on RG_i, where i represents the generation number.

Mutation: Mutation is performed by replacing 50 markers from population, which were not selected for next generation. Here, 50 new markers were selected randomly from ratio pair list (population).

Selection Function: Tri-peptide marker's survival on the generation is consider as selection criteria of good marker. Markers in GPL are sorted based on their occurrence in generations and top 100 highest occurring markers are selected at the end of GA. Example of ten tri-peptide markers is shown in Table 2.

Stop Condition: Stopping condition for our GA is that GPL should have enough number of highest occurring markers to cover all records in FTrD. We have achieved 96.8% coverage using 100 highest occurring markers until 26 generations.

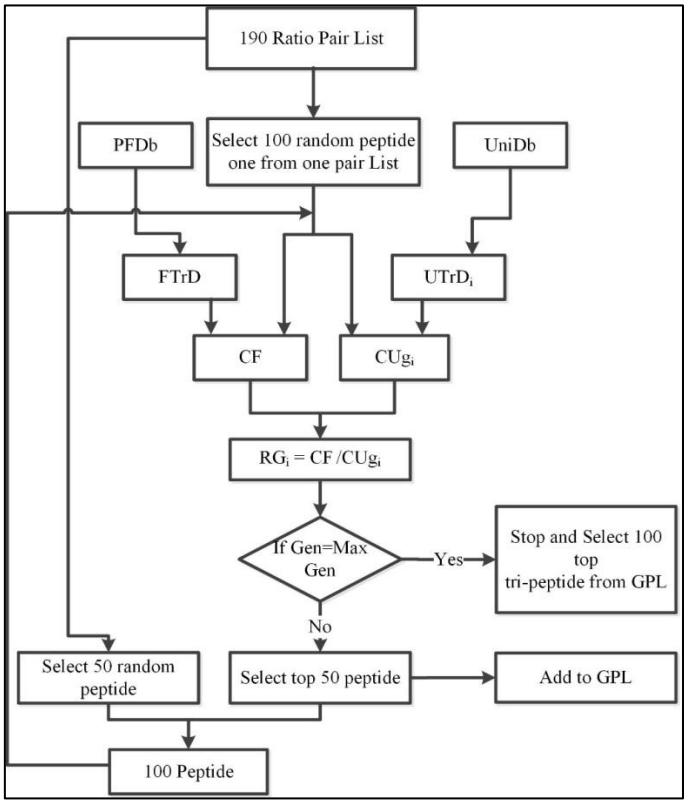

Fig. 3: Flow chart of Genetic Algorithm to generate 100 Tri-peptide Marker Set (TPMS). CF: occurance frequency of tri-peptide in FTrD. CUg_i: occurance frequency of tri-peptide in UTrD generated at ith generation from UniDb. $RG_i = CF/CUg_i$: Ratio of occurance frequency of particular tri-peptide; where i is the number of generation. GPL: Generation-wise tri-peptide List.

Iterative Convergence Algorithm

Iterative Convergence Algorithm (ICA) is used to handle the remaining records which are not characterized by selected 100 TPMS. Exception List (ExL) is generated which contains unlearned records i.e. the record which are not characterized using TPMS. Further more tri-peptides are generated from ExL. Frequency ratio is calculated between CF and CU, and top 30 peptides are selected and added to TPMS. These steps are repeated iteratively to reduce number of records in ExL as shown in Figure 4. At the end total 150 tri-peptids are added to TPMS and left with 153 record in ExL as shown in Figure 5. ICA helps to reduce number of records which are not recognized in FTrD by increasing TPMS length, hence improves the accuracy of prediction.

Table 2: Presence of 10 tri-peptide markers in every generation

Generation	Presence of tri-petide markers in every generation									
	KNA	NMN	MCS	KGF	NNI	NKN	HHA	QGK	DHN	KEM
G1	0	1	0	1	0	0	0	0	0	1
G2	1	1	0	1	1	1	0	0	0	1
G3	1	1	0	1	1	0	0	0	0	1
G4	1	1	0	1	1	0	0	0	0	1
G5	1	1	0	1	0	0	0	0	0	1
G6	1	1	0	1	0	1	0	0	0	1
G7	1	1	0	1	0	1	0	1	0	1
G8	1	1	0	1	1	1	1	1	0	1
G9	1	1	1	1	1	1	1	1	0	1
G10	1	1	1	1	1	1	0	1	0	1
G11	0	1	1	1	1	0	0	0	0	1
G12	1	1	1	1	1	1	0	1	1	1
G13	1	1	1	1	1	1	0	1	1	1
G14	1	0	1	1	1	1	0	1	1	1
G15	1	0	1	1	1	1	1	0	0	1
G16	1	0	1	1	1	1	1	0	0	1
G17	1	0	1	1	1	1	0	1	0	1
G18	1	1	1	1	1	1	1	1	1	1
G19	0	1	0	0	1	0	0	1	1	1
G20	0	1	1	1	1	0	0	1	1	0
G21	0	1	1	1	1	0	0	1	1	1
G22	0	1	0	1	1	1	1	1	1	1
G23	0	0	1	0	1	1	1	1	1	1
G24	0	1	1	1	1	1	1	1	1	1
G25	1	1	1	0	1	1	1	1	1	1
G26	1	1	1	0	1	1	1	1	1	1
Total	18	21	16	22	22	18	10	17	12	25

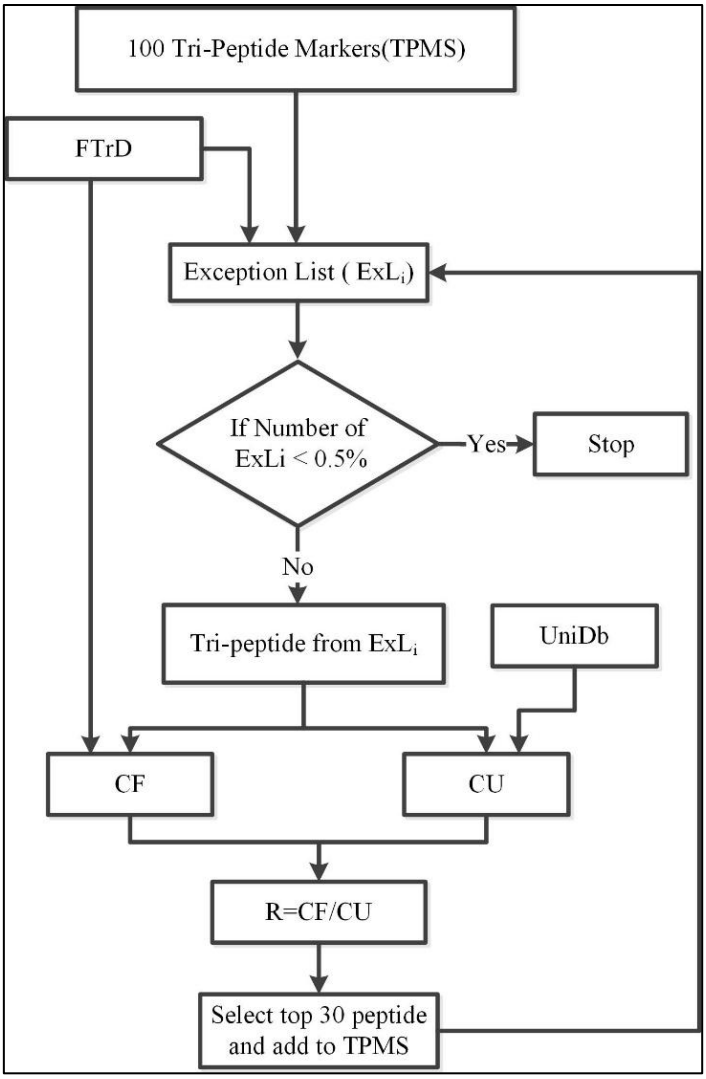

Fig. 4: Flow chart of Iterative Convergence Algorithm (ICA). ExL$_i$: Exception List containing unlearned records using TPMS in FTrD; where i is number of iteration. CF: occurance frequency of tri-peptide in FTrD, CU: occurance frequency of tri-peptide in UniDb. R = CF/CU: ratio of frequencies of particular tri-peptide.

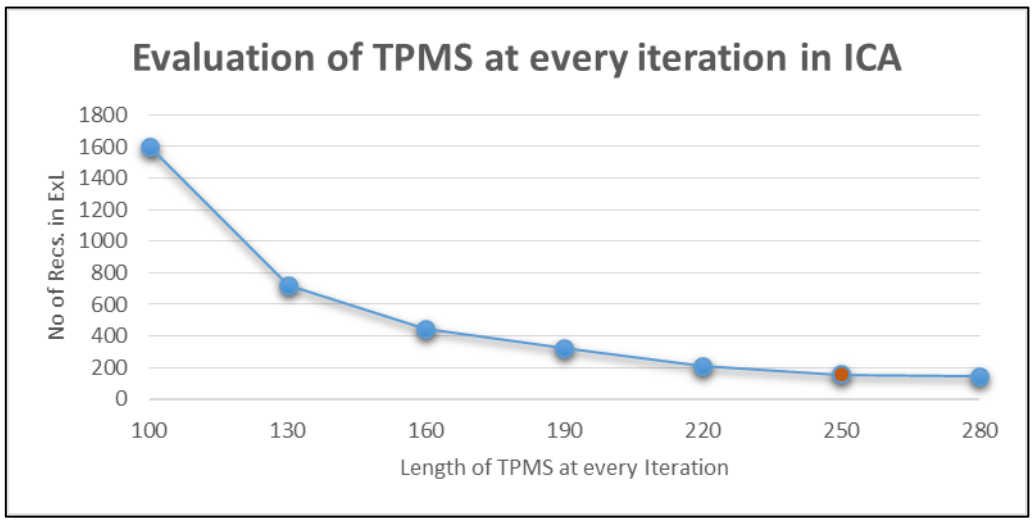

Fig. 5: Number of records in ExL for different size of TPMS at every iteration in ICA

Back-propagation Neural Network

As shown in Figure 6, back propagation neural network is used as a supervised learning algorithm. This network has 250 input neurons, 20 hidden neurons and one binary output neuron. Signature input to the neural network is prepared by checking presence of 250 tri-peptide markers in sequence of FTrD and 10 different UTrD datasets. Markers in sequence is represented as '1' for presence and '0' for absence, thus input signature is a binary signature. Then input signature is feeded to the network untill the error drops to minimum (below 10 % in our case). Output neuron gives binary result '1' if predicted as *Falciparum* Ag sequence and '0' if predicted as non-*Falciparum* Ag sequence. Once network is trained it is tested on test dataset FTeD which is not present in training process. Result shows that our network predicts *falcipaum* antigen with 93% accuracy as shown in Table 3.

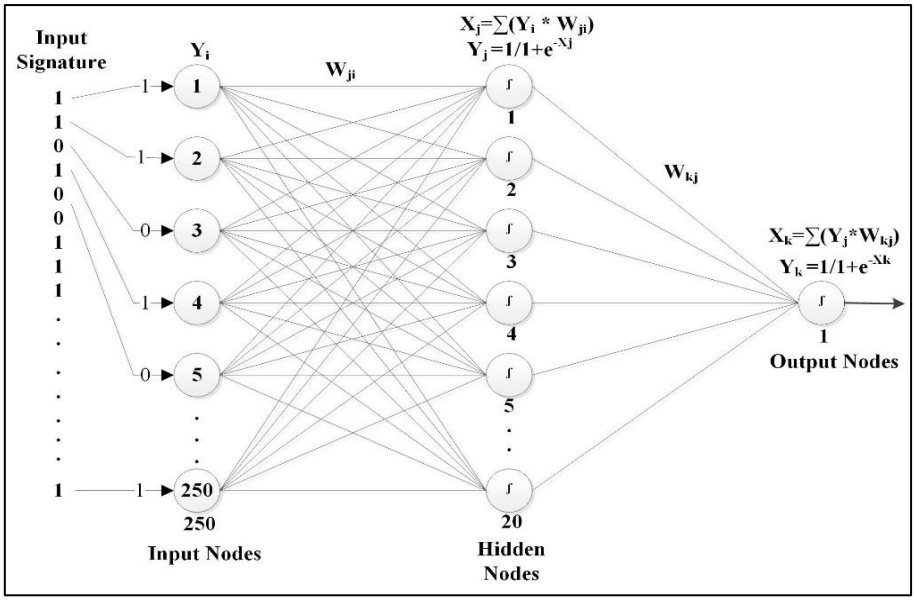

Fig. 4: Back-propagation Neural Network Architecture

Experiment and Results

A novel approach is proposed here to extract an optimal set of peptide markers, 250 tri-peptide markers are extracted from working database for *P. falciparum* Ag prediction. Selected tri-peptide occurance frequency (CF) in FTrD and occurance frequency(CU) in 10 UTrD datasets is shown in Figure 7 and Figure 8 respectively. It is observed from both the charts that TPMS has more presence in FTrD, thus this set is used as a combination to separate *P. falciparum* Ag sequences from other protein sequence.

Now using this 250 tri-peptide markers, binary signature is created as described in section 3 and feeded in neural network. In Table 3 results are compared for different number of hidden neurons in neural network.

Table 3 Accuracy of the Neural Network for different Hidden Neurons

Number of Hidden Neurons	Sensitivity (Sn)	Specifity (Sp)	Accuracy
10	91.91	92.71	92.31
20	92.07	93.87	93.30
30	91.84	94.01	92.93
40	91.32	93.52	92.42
50	90.82	92.50	91.68

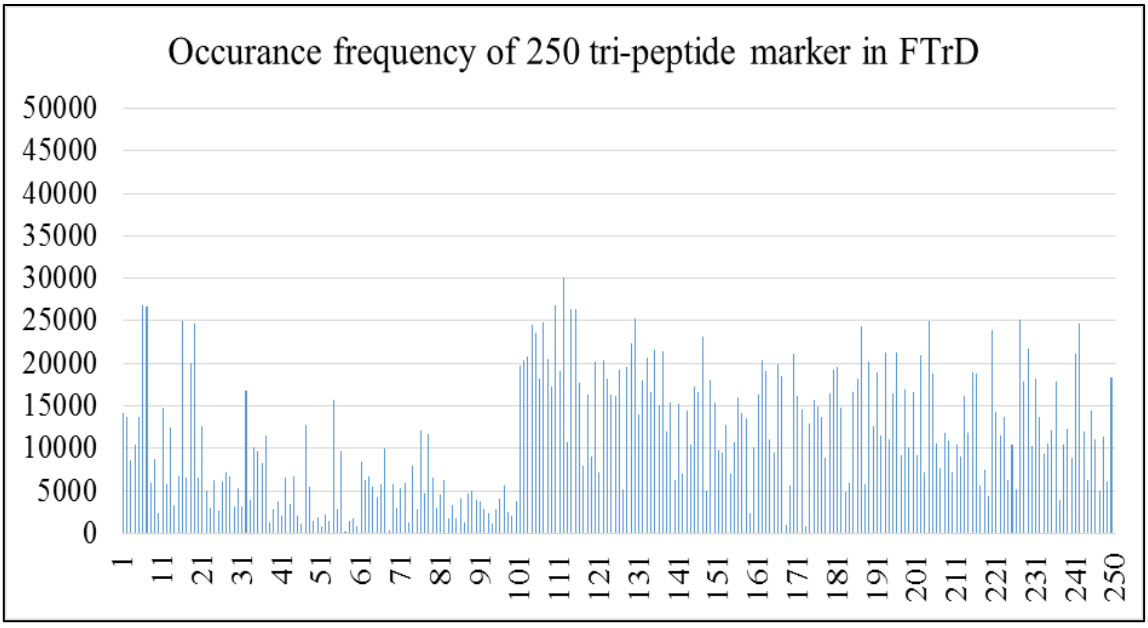

Fig. 5: Occurrence frequency of identified 250 markers in *Falciparum* training dataset-FTrD

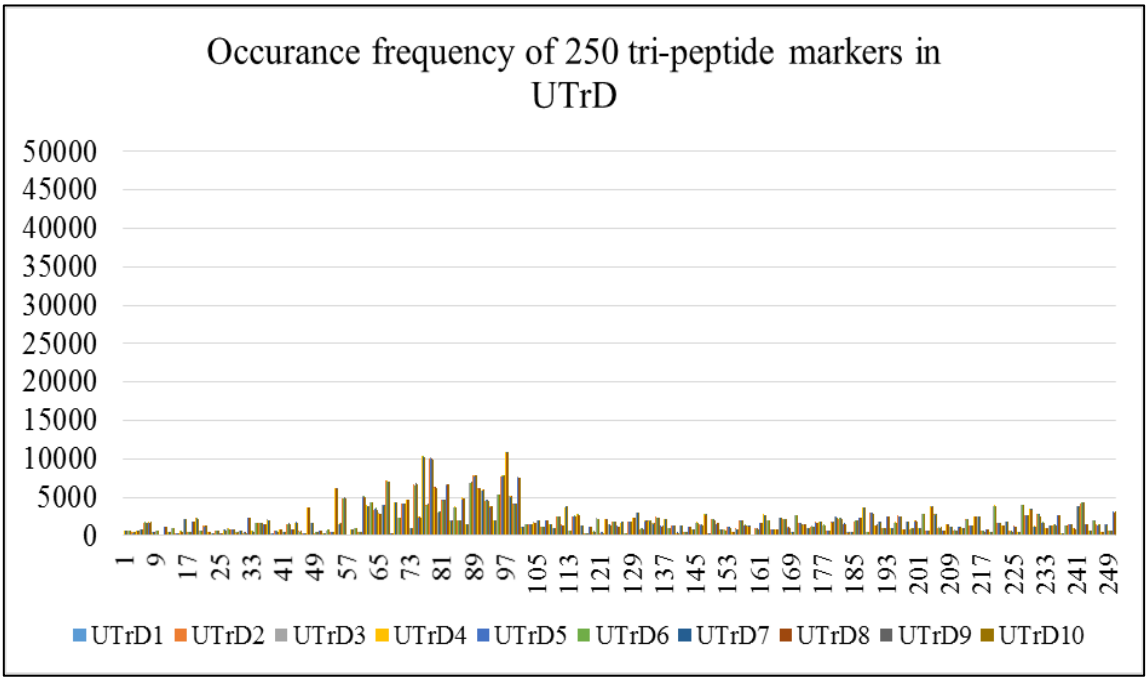

Fig. 6: Occurrence frequency of identified 250 markers in uniref training datasets-UTrD

DISCUSSION

Selecting residue size for marker is necessary as well as challenging task. It is important that n-residue word should represent largest possible data in respective dataset. There are total 20^n combinations possible for n-residue words as protein sequence is made up of a combination of 20 amino acids. Hence it is necessary to optimze the memory usage. As shown in Table 4 di-peptide has omnipresent in UniDb and hexa-peptide holds very large memory thus we have not used di-peptide and hexa-peptide. We found the occurance frquency of tri-peptide as 30,000, tetra-peptide as 18000 and penta-peptide as 8000. Hence tri-peptides are selected due to its considrable higher occurrence compare to tetra-peptide and penta-peptide.

Table 4: n-residue words and their occurance frequency in UniDb; where n ranges from 2 to 6

Type of Peptide	Residue Size	Number of Combinations possible	Size of Files	Frequency in UniDb / peptide
Di-Peptide	2	4×10^2	10 KB	297061849,LA
Tri-Peptide	3	8×10^3	118 KB	36572557, LAA
Tetra-Peptide	4	16×10^4	2360 KB	3386524, LAAS
Penta-Peptide	5	32×10^5	53125KB	223386, LAASA
Hexa-Peptide	6	64×10^6	1062500KB	3500, LAASAG

A key issue in Selecting a minimal number of tri-peptides in TPMS is that, it should be optimal. As disscussed in section 2.3.2, GA generates 50 tri-peptides in every generation which are stored in separate file GPL. TPMS is prepared by selecting top 100 tri-peptide based on most occuring tri-peptides. Optimality is decided by number of records in ExL generated using TPMS and FTrD, which should be minimal with less number of tri-peptides. An optimal 100 tri-peptide set is selected as it has minimum 1597 records in ExL. Rational for selecting 100 tri-peptide is shown in Figure 9.

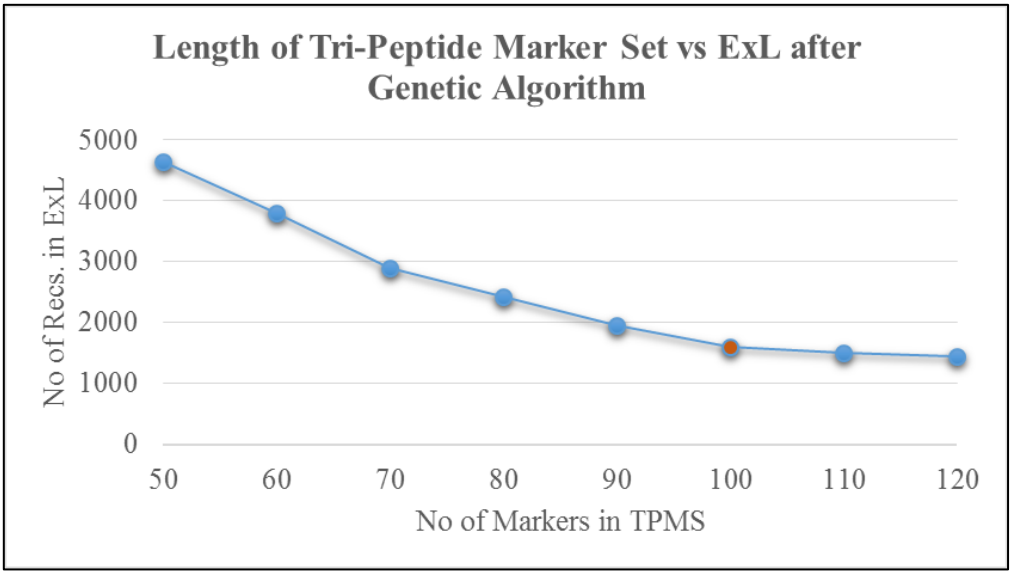

Fig. 7: Number of records in ExL for different size of TPMS after GA

Accuracy of the neural network depends on hidden neurons, network with more neurons are more complex and leads to the overfitting problem, and network with less neurons are not sufficient to learn very accuratly[19]. As we can see in Table 3 network with 20 hidden neurons gives 93% accuracy for 250 input neurons. Neural network starts overefitting as we increase number of neurons in hidden layer network. Once neural network starts overfitting error in the training dataset reduces (i.e the network starts memorizing instead of modelling) but in test dataset error increases and reduces the accuracy.

CONCLUSION

We have proposed a novel aproach to predict an antigen (Ag) sequences using 3-residue words (tri-peptide marker). Tri-peptide markers are generated and evaluated for *P. falciparum* antigen group by Genetic Algorithm. Multiple iterative evaluation of markers by fitness function ensures that combination of 250 markers represents 99.7 % sequences in training dataset. A back propagation neural network is trained by training dataset with presence of selected markers in the sequence. The trained back-propagation neural network predicts unknown

sequence as *P. falciparum* or non *P. falciparum* antigen sequence with 93% of accuracy. The approach proposed here can also be used to predict any other target antigen group. we are looking forword to work on selecting orthogonal peptide markers using clustering based on there co-occurance and to develop a generic tool for antigen sequence characterization.

Acknowledgment

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

REFERENCES

1. *Alberts, B., Johnson, A. and Lewis, J. et al., The Adaptive Immune System"*. Molecular Biology of the Cell *(4th ed.). New York: Garland Science, (2002).*
2. Antigen, Retrieved from: https://medlineplus.gov/ency/article/002224.htm, [Accessed on: 5-July-2016]
3. Bhasin, Manoj, and Raghava, G.P.S., Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine*, **22(23):** 3195-3204 (2004).
4. David, E., Goldberg, and John, H., Genetic algorithms and machine learning. *Machine learning*, **3(2):** 95-99 (1988).
5. Gaurang Panchal, Amit Ganatra *et al*., Determination of over-learning and over-fitting problem in back propagation neural network. *International Journal on Soft Computing*, **2(2):** 40-51 (2011).
6. Himanshu S., Mazumdar, Ankita, C., Baravaliya, and Maulika, S.P., Keyword based Iterative Approach to Multiple Sequence Alignment. *Int. J. Pure App. Biosci*., **2(3):** 139-144 (2014).
7. Deutsch, J.M., Evolutionary algorithms for finding optimal gene sets in microarray prediction. *Bioinformatics*, **19(1):** 45-52 (2003).
8. Jack, S., Richards, Thangavelu, U., Arumugam, Linda Reiling, *et al*., Identification and Prioritization of Merozoite Antigens as Targets of Protective Human Immunity to *Plasmodium falciparum* Malaria for Vaccine and Biomarker Development, *The Journal of Immunology*, (2016).
9. Jane Jijun Liu, Gene Cutler *et al*., "Multiclass cancer classification and biomarker discovery using GA-based algorithms." Bioinformatics, **21(11):** 2691-2697 (2005).
10. Jihoon, Y. and Vasant, H., Feature subset selection using a genetic algorithm. Feature extraction, construction and selection. *Springer*, 117-136 (1998).
11. Joao Aguiar, Keith Limbach *et al*., *Plasmodium falciparum* sporozoite and liver stage antigens. U.S. Patent Application No. 14/219-390, (2014).
12. Johannes Sollner, and Bernd Mayer, Machine learning approaches for prediction of linear B-cell epitopes on proteins. *Journal of Molecular Recognition*, **19(3):** 200-208 (2006).
13. Katarzyna Krzyczmonik, Michał S witnicki, Szymon Kaczanowski, Analysis of immunogenicity of different protein groups from malaria parasite *Plasmodium falciparum* Infection, *Genetics and Evolution*, **12(8):** 1911-1916 (2012).
14. Kazuyuki Tanabe, Martin Mackay *et al*., Allelic dimorphism in a surface antigen gene of the malaria parasite *Plasmodium falciparum*, *Journal of molecular biology*, **195(2):** 273-287 (1987).
15. Li, L., Wei, J. and Xia Li, *et al*., A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics*, **85(1):** 16-23 (2005).
16. Maulika, S., Patel, and Himanshu, S., Mazumdar, Knowledge base and neural network approach for protein secondary structure prediction. *Journal of theoretical biology*, **361:** 182-189 (2014).
17. Maulika, S., Patel, and Himanshu, S., Mazumdar, Similarity search using pre-search in UniRef100 database.

International Journal of Hybrid Information Technology, **4(3):** (2011).

18. Stephen, M., Rich, Fabian, H., Leendertz, and Guang Xu, *et al*., The origin of malignant malaria. *Proceedings of the National Academy of Sciences*, **106(35):** 14902-14907(2009).
19. Thomas, P., Hopp, Kathryn, S. and Prickett, *et al*., A short polypeptide marker sequence useful for recombinant protein identification and purification. *Biotechnology*, **6(10):** 1204-1210 (1988).
20. UniRef100.fasta Database, Retrieved from: http://www.uniprot.org/downloads, [Downloaded on:20-June-2016]
21. Why is malaria Dangerous?, Retrieved from: http://www.malaria.com/questions/why-malaria-dangerous, [Accessed on: 8-August-2016]
22. World Health Organization. "World malaria report 2015". World Health Organization, (2015).
23. Yasser EL-Manzalawy, Drena Dobbs, and Vasant Honavar,mPredicting linear B-cell epitopes using string kernels. *Journal of molecular recognition*, **21(4):** 243-255 (2008).