A Recurrent Neural Network Linear B-Epitope Predictor: BIRUNI

S.A.R. Abidi

Faculty of Engineering and Natural Sciences,
International University of Sarajevo International University of Sarajevo,
HrasnickaCesta 15, Ilidža 71210 Sarajevo,
Bosnia and Herzegovina
azraabidi34@yahoo.com

Article Info

Abstract

Experimental methods used for characterizing epitopes that play a vital role in the development of peptide vaccines, in diagnosis of diseases, and also for allergy research are time consuming and need huge resources. There are many online epitope prediction tools are available that can help scientists in short listing the candidate peptides. To predict B-cell epitopes in an antigenic sequence, Jordan recurrent neural network (BIRUNI) is found to besuccessful. To train and test neural networks, 262.583 B epitopes are retrieved from IEDB database. 99.9% of these epitopes have lengths in the interval 6-25 amino acids. For each of these lengths, committees of 11 expert recurrent neural networks are trained. To train these experts alongside epitopes, non-epitopes are needed. Non-epitopes are created as random sequences of amino acids of the same length followed by a filtering process. To distinguish epitopes and non-epitopes, the votes of eleven experts are aggregated by majority vote. An overall accuracy of 97.23% is achieved. Then these experts are used to predict the Linear Bepitopes of five antigens, Plasmodium Falciparum, Human Polio Virus Sabin Strain, Meningitis, Plasmodium Vivax and Mycobacterium Tuberculosis. The success of BIRUNU is compared with the five online prediction tools ABCPRED, BCPRED, K&T, BEPIPRED, and AAP.It is seen that BIRUNI outperforms all of them in the average.

1. INTRODUCTION

In all living organisms, B-cells of the immunesystem recognize the pathogen's antigens by their membrane-bound immunoglobulinreceptors and, in response, produce antibodies specific to these antigens. Antigens have the capacity to bindby either a B-cell receptor or an antibody molecule. The part of an antigen that binds to an antibody iscalled a B-cell epitope. If an antigen is a protein, an epitope maybe either a short peptide from the protein sequence or a patch of atoms on the protein surfacein the three-dimensional structure.

In the process of antibody detection, and production, design of a molecule that can replace an antigen is needed. This molecule is antigen's epitope. If this epitope is a protein, its gene can be cloned, if it is a molecule, it can be synthesized. These designed molecules are inexpensive and noninfectiousin contrast to viruses or bacteria.

A peptide is called a continuous epitope of a protein when it corresponds to a short continuous stretch from the sequence of this protein, and binds the antibody raised against this protein. However, it is observed through crystallographic studies of antibodies that, most epitopes consist of atoms from distant residues joined on the protein surface in the three-dimensional space as a patch. In that sense, they are discontinuous. A third kind is nonprotein epitopes that are polysaccharides, DNA, glycoproteins, and other molecules, that expands the range of potential pathogensthat can be targeted by vaccines and diagnostics.
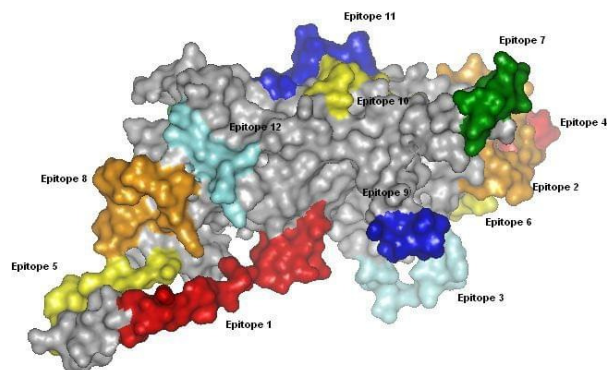

Figure 1. Predicted B cell epitopes of Human papillomavirus type 16 major late capsid protein L1

Prediction of immunogenic epitopes using bioinformatics tools is a challenging task because of the inherentcomplexity of antigen recognition (Flower, 2003).At the beginning, to predict continuous B-cell epitopes, a number of algorithms havebeen developed based on physico-chemical properties of aminoacids (Pellequer, et. al., 1991) but their accuracies were very low. The commonly used properties are hydrophilicity (Parker, 1986) flexibility (Karplus, and Schulz 1985), surface accessibility (Emini, et., al. 1985), and β-turns (Pellequer, et. al., 1993). All ofthe prediction calculations arebased on the propensity tables for each of the 20 aminoacids.

A few computer programsare also developed to make the job predicting epitopes inan antigenic sequence online. PREDITOP (Pellequer, andWesthof 1993) uses22 normalized scales, corresponding to hydrophilicity,accessibility, flexibility, and secondary structure propensities. PEOPLE (Alix, 1999) uses physico-chemical properties, surface accessibility,hydrophilicity, and flexibility.

BEPITOPE (Odorico, and Pellequer, 2003)searches for patterns either in a singleprotein or on a complete translated genome. BcePred uses also various physicochemical properties (Blythe, and Flower2005). The predictive value of algorithms varies between 40–68% (VanRegenmortel, and de Marcillac1988; Saha, and Raghava2006).

In 1980s, direct prediction methods based on structural and sequential analysis of T cell epitopes are also developed (DeLisi, and Berzofsky, 1985; Cornette, 1993; Spouge 1987; Stille, 1996). DeLisi and Berzofsky(1985) proposed that the critical requirement of T cell epitopes is its ability to form stable amphipathic structure. Based on this hypothesis, a program AMPHI was developed (Cornette, 1993;Stille, 1996). Another algorithm SOHHA was developed based on the assumption that T cell epitopes consist of a helix of 3–5 helical turns with a narrow strip of hydrophobic residues on one side.

A number of indirect methods have also been developed that predict MHC binders instead of T cell epitopes. They are based on structure, binding motifs, propensity matrices and Artificial Neural networks (ANNs) (Brusic, et. al. 1994;Gulukotaet. al. 1997; Rammenseeet. al. 1996;Bhasin, and Raghava,2004; Parker, et. al., 1994; Singh, and Raghava, 2001; 2003;Bahasin, and Raghava, 2003; 2004).

In 2004, machine learning techniques like artificial neuralnetworks (ANN)s to predict Bcellepitopes started to emerge (Saha and Raghava 2006;). In machine learningtechniques, the input window length has to befixed. 262.583 B epitopes inIEDB database, 99.9% of these epitopes have lengths in the range of 6-25 amino acids. To deal with this abundance, (Saha and Raghava 2006; Bhasin, and Raghava,2004) take a fixed length of 20 amino acids although 77% of B epitopes inIEDB database have lengthof 15 amino acids as seen in Table 1. Furthermore, they have mistaken in assuming that adding residues taken from the parent/original antigenic sequences or removing residues at the terminals, does not change their character as being of Bcell epitopes. In this research, eleven neural networks are trained for each of the 20 epitope lengths between 6 and 25.

2.MATERIALS AND METHODS

2.1 Recurrent Artificial Neural Networks (RNN)

Variables in the data are dependent nonlinearly to the outcome, and hence recurrent neural networks (RNN) are particularly useful. They are constructed by taking a feedforward network and adding feedback connections from output and/or hidden layers to input layers. The standard backpropagation algorithm also trains these networks conditional that patterns must always be presented in time sequential order. The one difference in the structure is that there are extra neurons in the input

layer that is connected to the hidden layer and/or output layer just like the other input neurons. These extra neurons hold the contents of one of the layers as it existed when the previous pattern was trained. In this way, the network takes into account previous knowledge it has about previous inputs. These extra neurons are called the context unit and it represents the network's long-term memory (Balkin, 1997).

To train neural networks, one needs both positive (e.g. B-cell epitopes) as well as negative (e.g. non-B-cell epitope) datasets. We downloaded the positive B-cell epitope data set from the iedb.org. database. Since there is no proven non-epitopes database, peptides with desired number of amino acids are taken randomly. Since these random peptides may also have B-cell epitopes, they are filtered according to a similarity threshold.

Jordanian recurrent neural networks (JRNN)s are used in Akcesme, et. al. 2017 to predict the B-cell epitopes in an antigenic sequence. In that research work for each window length, in the domain 6–25, committee machines with eleven JRNNs in each committee are trained and votes of experts are aggregated by simple majority voting. It is seen there that committee machine voting boosts accuracy at least 10%.

2.2. Blind Data

To evaluate performances of experts in committee machines, they are tested on five antigens, Plasmodium Falciparum, Human Polio Virus Sabin Strain, Meningitis, Plasmodium Vivax and Mycobacterium Tuberculosis.

To have an idea about accuracies and reliabilities of the above-mentioned online tools, we have chosen five antigens, whose linear b-epitopes are known and compiled through wet lab experiments. i.e.Plasmodium Falciparum, Human Polio Virus Sabin Strain, Meningitis, Plasmodium Vivax and Mycobacterium Tuberculosis.

2.2.1   Mycobacterium Tuberculosis

Members of the genus Mycobacterium are characterized by a verycomplex cell wall envelope that is responsible for the remarkablelow permeability of their cells as well as the characteristicdifferential staining procedure (known as Zhiel-Neelsen acid-faststain), which specifically stains all members of the genera. Bothfeatures are due to the presence of long chain a-alkyl, β-hydroxyfatty acids in their cell wall. TheMycobacterium genus is usuallyseparated into two major groups on the basis of their growthrate.Tuberculosis remains the most devastating bacterial cause of human mortality (1). Despite of improved diagnosis, surveillance, and treatment regimens, the incidence of TB increases annually (2). The ability to combat this deadly pathogen hinges on the dissection and understanding of the mechanisms of pathogenesis for Mycobacterium tuberculosis. Central to the ability of the microbe to cause disease is the capability to survive and replicate within macrophages by avoiding lysosomal fusion with the mycobacteria-containing phagosome. M. tuberculosisinteracts with and invades various human and animal epithelial cells in culture and appears to possess multiple mechanisms of entry into macrophages. Furthermore, the specific bacterial adhesions involved in the complex interplay between M. tuberculosisand the human host are largely unknown. For Mycobacterium Tuberculosis 13 linear B-epitopes are reported (Young, et. al. 2013).

2.2.2 Plasmodium Falciparum:

Plasmodium falciparum is a protozoan parasite that causes an infectious disease known as malaria. P. falciparum is the most severe strain of the malaria species correlated with almost every malarial death. The other 3 species that cause malaria include: P. vivax, P. ovale, and P. malariae. Humans become infected by a female Anopheles mosquito which, transfers a parasitic vector through its saliva into the blood stream. The parasite then infects the liver and undergoes asexual reproduction followed by insertion into red blood cells where an additional round of replication takes place. P. falciparum changes the surface of an infected red blood cell causing it to adhere to blood vessels, cytoadherence, as well as to other red blood cells.In severe cases this leads to obstructions of microcirculation resulting in dysfunction of many organs. Symptoms depend on severity of infection and can present a range of signs such as flulike symptoms, vomiting diarrhea, shock, kidney failure, coma, and death. Plasmodium falciparum mostly infects children under the age of 5 as well as pregnant women. An important virulence property of P. falciparum is the expression of parasite-derived antigens on the surface of IEs, generally known as variant surface antigens, and its strong propensity to adhere in the vasculature.

Sickle cell individuals have shown to rarely contract malaria. Research has shown that this is partially due to weakened binding of parasite-infested sickle cell erythrocytes to micro vascular endothelial cells when compared to normal hemoglobin parasite erythrocytes binding. The virulence factor PfEMP1 that normally conducts cytoadherence is altered creating a weekend attachment between it and the epithelial wall. Due to the ability to attach lacking, sequestration would also not occur limiting the severe malarial response. The mechanism for how this is done is still unknown and needs further research.The 26 wet Lab reported linear B Cell epitopes of Plasmodium Falciparum are listed in Appendix A2(Isea, 2017).

2.2.3   Meningitis

Viral meningitis is contagious and infectious disease in which there is an inflammation of themembranes of cerebrospinal fluid (CSF). The membranes and cerebrospinal fluid (CSF) encase and bath the brain and

spinal cord. Viral meningitis is the most common type of meningitis. Bacterial meningitis is less common. Viral meningitis is also sometimes called aseptic meningitis.Meningitis is by far the most common neurological manifestation of mumps virus infection. Before widespread immunization, mumps was a common cause of meningitis, which occurred in 15% of patients with mumps. Mumps meningitis can precede or follow the parotid swelling, and 50% of cases occur in the absence of parotitis. Meningitis is more common in males than in females. Diagnostic tests include a lumbar puncture, also called a spinal tap. A lumbar puncture involves withdrawing a small sample of cerebrospinal fluid (CSF) from the spine with a needle. The sample of CSF is tested to rule-out bacterial meningitis and diagnose viral meningitis.Meningitis may be accompanied by mucocutaneous manifestations of enterovirus infection, including localized vesicles such as in hand, foot, and mouth disease; herpangina; and generalized maculopapular rash. Most cases that present clinically with meningitis are self-limiting and carry a good prognosis. Nevertheless, enteroviral meningitis causes considerable morbidity, with moderate or high fever despite antipyretics and several days of severe headache warranting opiate analgesia. Abrupt deterioration in mental status or seizures may be caused by progression from meningitis to meningoencephalitis.No specific antiviral treatment is available, and management is conservative. Immunoglobulin replacement has a role in patients with hypogammaglobulinemia, who are prone to severe and chronic enteroviral disease. For meningitis 9 linear B-epitopes are listed in Appendix A3(Chandra, and Singh, 2012).

2.2.4   Human Polio Virus

Poliovirus, the causative agent of paralytic poliomyelitis, is an enterovirus spread by the oral route. The principal infection associated with the poliovirus is enteritis with the prodromal illness of fever, headache, arthralgia, vomiting, and diarrhea lasting 3–4 days. About half of the patients do not develop paralytic manifestations. In the remaining, a biphasic course evolves. As the initial enteritis subsides, the paralysis begins. Severe back and limb pain, headache, and meningismus develop, accompanied by severe and disabling muscle spasms. Paralysis tends to occur in a patchy, multifocal distribution. Weakness of individual muscles comes on rapidly over days and typically reaches a maximum within 1 week. The virus has a specific tropism for the motor neurons, resulting in motor neuron death. Virtually any of the skeletal muscles, including bulbar, limb, and respiratory muscles, can be affected. The time from being infected with the virus to developing symptoms of disease (incubation) ranges from 5 - 35 days (average 7 - 14 days). Most people do not develop symptoms.Outbreaks can still occur in the developing world, usually in groups of people who have not been vaccinated. Some victims develop neurological

complications, including stiffness of the neck and back, weak muscles, pain in the joints, and paralysis of one or more limbs or respiratory muscles. In severe cases it may be fatal, due to respiratory paralysis. Despite the eradication of acute poliomyelitis, there remains a large population of patients with significant motor deficits who were infected before the onset of the vaccination programs.

The World Health Organization has now eradicated wild-type polio from all but four countries limited to central Africa. It is hoped that if mass vaccination programs are allowed to continue in central Africa, complete eradication will be resulted within a few more years.

In our insilico research project we have taken FASTA OF Human Polio Virus Sabin strain from GenBank with identification number AAN85444.1 polyprotein [Human poliovirus 3]. 64 linear B-epitopes are listed in Appendix A4(Nomoto et. al. 1982).

2.2.5   Plasmodium Vivax

Plasmodium vivaxis a protozoal parasite and a human pathogen. The most frequent and widely distributed cause of recurring (Benign tertian) malaria, P. vivax is one of the six species of malaria parasites that commonly infect humans. It is less virulent than Plasmodium falciparum, the deadliest of the six, but vivax malaria can lead to severe disease and death due to splenomegaly (a pathologically enlarged spleen). P. vivax is carried by the female Anopheles mosquito, since it is only the female of the species that bite.Plasmodium vivax malaria is prevalent in many regions of the world. It accounts for more than half of all malaria cases in Asia and Latin America. Despite the high prevalence of disease caused by this parasite, research regarding its effects has lagged disproportionately.Organ dysfunction which is seen in P. falciparum malaria, is not seen in P.vivax infections. Thus, severe malaria is reported with P. falciparum but not with P. vivax infection. 26 linear b cell epitopes are listed in Appendix A5 (Caro-Aguilar, et. al. 2002).

3.   THE FIVE MOST PROMINENT LINEAR B-EPITOPE PREDICTION TOOLS

The success of BIRUNU is compared with the five online prediction tools ABCPRED, BCPRED, K&T, BEPIPRED, and AAP.

3.1 ABCpred

ABCpred server is used to predict linear B cell epitope regions in an antigen sequence. Saha and Raghava (2006a) have made available the data sets used to train and evaluate ABCpred. We have selected 20mers data set to analyze its performance. ABCpred uses recurrent artificial neural networks for predicting linear B-cell epitopes B-cell epitopes play a vital and significant role in the development of peptide vaccines. This is the first server developed based on recurrent neural network (machine-

based technique). Users can select window length of 10, 12, 14, 16 and 20 as predicted epitope lengths. ABCpred generates datasets of fixed length patterns by eliminating or adding residues at the terminal ends of the peptides. It presents the results in graphical and tabular frame. In case of graphical frame, this server plots the epitopes in blue color along protein backbone (black color), which assist the users in rapid visualization of B-cell epitope on protein. The tabular output is in the form of a table, which will provide the amino acids length from N-terminal to C-terminal in a protein predicted by the server. Experimental methods used for characterizing epitopes are time consuming and demand large resources. The availability of epitope prediction method(s) can rapidly help experimenters in simplifying and modifying this problem. In ABCpred all calculations done by using artificial neural network. This server will assist in locating and finding of epitope regions that are useful in selecting synthetic vaccine candidates, disease diagnosis, prediction of potential vaccine candidates and also in allergy research. ABCpred server is available at the web site (Saha, and Raghava2006);
http://osddlinux.osdd.net/raghava/abcpred/index.html

3.2 BCpred

The BCpreds server allows users to choose the method for predicting B-cell epitopes among several developed prediction methods. The current implementation of BCPREDS allows the user to select among three prediction methods: (i) implementation of AAP method (Chen et al., 2007); (ii) BCPred (EL-Manzalawy et al., 2008a); (iii) FBCPred (EL-Manzalawy et al., 2008b). The way to handle is that users provide an antigen sequence and optionally can specify and notify desired epitope length and specificity thresholds also. Results are returned in several user-friendly formats. BCPred server allows choosing prediction method among amino acids pair scaling method (AAP), BCPred, and FBCPred. AAP approach is based on the finding that particular amino acid pairs occur more frequently in epitope than nonepitope sequence. Combination of AAP propensity scale with turns, accessibility, antigenicity, hydrophilicity, and flexibility propensity scales improved the accuracy (72.5%). BCPred method employs subsequence kernel-based SVM classifier and was trained on homology-reduced dataset of linear B-cell epitopes (with <80% sequence identity) derived from dataset previously used to test ABCpred. BCpred server is available at the web site (EL-Manzalawy et al., 2008c).
http://ailab.ist.psu.edu/bcpred/

3.3 Kolaskar&Tongaonkar

The quality of Kolaskar andTongaonkar is that it'sanalysisof the data from experimentally determined antigenic sites on proteins has revealed that the hydrophobic residues Cys, Leu and Val, if they occur on the surface of a protein, are more likely to be a part of antigenic sites. A semi-empirical method which makes use of physicochemical properties of amino acid residues and their frequencies of occurrence in experimentally known segmental epitopes was developed to predict antigenic determinants on proteins. Application of this method to a large number of proteins has shown that our method can predict antigenic determinants with about 75% accuracy which is better than most of the known methods. This method is based on a single parameter and thus very simple to use. The B cell epitope prediction was performed using the program predicting antigenic peptides available online. The software for the detection of antigenic peptides is based on Kolaskar and Tongaonkar method previously described (Kolaskar and Tongaonkar, 1990). Kolaskar, andTongaonkar antigenicity is a semi-empirical method for the prediction of antigenic regions including information of surface accessibility and flexibility. Antigenicity was determined on the basis of well-balanced algorithmic method developed by Kolaskar and Tongaonkar for the prediction of antigenic determinants on protein antigens. This method is based on physicochemical properties of amino acid residues and their frequencies of occurrence in experimentally known segmental epitopes. KolaskarandTongaonkar online prediction tool is available at the web site (Kolaskar, and Tongaonkar, 1990):
http://tools.immuneepitope.org/bcell/

3.4 Bepipred

For each input sequence, BepiPred outputs not only a prediction score for every single residue but also it predicts the sequence of potential epitope. The positions of the linear B-cell epitopes are predicted to be located at the residues with the highest scores. Bepipred uses a hidden Markov model-based method along with amino acid propensity scales for accessibility, hydrophilicity, flexibility and polarity trained on a dataset of curated B cell epitopes. Lastly, the ABCpred prediction tool is an artificial neural network-based B cell epitope prediction server that recognizes that B cell epitopes have varying lengths (5 to 20 residues). In this paper we used IEDB Analysis Bepipred tool. The BepiPred method applies a hidden Markov model which takes two propensity scores as its inputs. A number of machine learning-based model were recently developed, from decision trees and k-nearest neighbor that utilized a combination of multiple propensities and sequence complexity as inputs, to neural network-based ABCpred that performs predictions directly from protein chain. The later method is designed to recognize epitopic peptides with 20 or fewer (i.e., 10,12,14,16 and 20) amino acids (AAs). The newest sequence-based predictors of continuous B-cell epitopes exclusively use support vector machine (SVM) models. BepiPred (http://tools.iedb.org/bcell/) is a combination method, produced by combining the predictions of a Hidden Markova model and the propensity scale by Parker et al. This method assigns a score value to each protein residue. Threshold was set at −0.2 (to obtain the sensitivity

of 75% and specificity of 50%, similar to those of SVMTriP) or at 0.35 (the default). BepiPred online prediction tool is available at the web site (Larsen, and Nielsen, 2006):

http://www.cbs.dtu.dk/services/BepiPred-1.0/

3.5 AAP

The AAP antigenicity scale can reflect some special sequence-coupled feature in the prediction of B-cell epitopes, which is the essence why the new approach is superior to the existing ones. It is anticipated that with the continuous increase of the known epitope data, the power of the AAP antigenicity scale approach will be further enhanced and improved. AAP method 20 maps each peptide sequence into a set of fixed length numeric features and therefore it can be trained using datasets of flexible length sequences. However, the performance of this method had been reported using a dataset of 20-mer peptides. Identification of antigenic sites on proteins is of vital importance for developing synthetic peptide vaccines, immunodiagnostic tests and antibody production. Currently, most of the prediction algorithms rely on amino acid propensity scales using a sliding window approach. These methods are oversimplified and sometime yield poor predicted results in practice. In this paper, a novel scale, called the amino acid pair (AAP) antigenicity scale, is proposed that is based on the finding that B-cell epitopes favor particular AAPs. It is demonstrated that, using SVM (support vector machine) classifier, the AAP antigenicity scale approach has much better performance than the existing scales based on the single amino acid propensity. The path toward a balanced portfolio of ABCpred is a capable, safe, and transparent artificial intelligence basedsystem which draws on a broad spectrum of computing ideas and principles. ABCpred along with other online prediction tools are likely to become a driver for new advances in computing. AAP online prediction tool is available at the web site (Chen, et. al. 2007):

http://ailab.ist.psu.edu/bcpred/

4. COMPARISON OF BIRUNI WITHOTHER FIVE ONLINE TOOLS

FASTA sequences of five antigens are uploaded to five online tools. Number of wet lab reported b-epitopes alongside with the number of predicted b- epitopes are shown in table 2. To the last column the number of b-epitopes predicted by BIRUNI is added (Akcesme, et al. 2017).

Table 1. Second column content is the number of linear b-epitopes reported by wet lab experiments. Other columns are the numbers of peptides that are predicted as B-epitopes by the five online tools.

	WET	BIR	ABC	AAP	BC	K&T	BEPI
TUBER	13	21	6	3	2	3	5
PLFAL	26	85	95	31	28	21	26
MENIN	9	150	29	8	9	10	11
POLIO	64	250	194	60	45	54	75
PLVIV	26	62	15	6	4	5	7

The predicted peptides are compared with the wet lab reports with a similarity threshold of 80%. If predicted peptide has a longest common subsequence with the wet lab reported epitopes, with a length not less than 80% of the wet lab reported epitope, then this peptide is recorded as a success. Table 2 shows the number of successfully predicted b-epitopes by the five online predictors, and BIRUNI for each of the five antigens.

The accuracy of predictors is defined as:

ACC=Number of correctly predicted epitopes by a predictor/ Number of epitopes in wet lab report

(1)

Table2. Successfully predicted wet lab reported b-epitopes by the five predictors, and BIRUNI for each of the five antigens, and average success of each.

	WET	BIR	ABC	AAP	BC	K&T	BEPI
TUBER	13	10	9	6	4	2	4
PLFAL	26	24	23	17	17	1	7
MENIN	9	3	0	0	0	2	1
POLIO	64	55	57	24	26	18	9
PLVIV	26	19	20	9	5	7	7
SUM	138	111	109	56	52	30	28
%	100	80	79	41	38	22	20

From Table 2 it is seen that the highest accuracy 80% is obtained by the artificial neural network predictor BIRUNI. ABCpred. is the second best with79% of average accuracy.

5. DISCUSSION

During the last three decades, many algorithms have been developed to predict the location ofcontinuous B cell epitopes in antigen proteins. Among them are machine learning techniques like artificial neural networks (ANNs), and support vector machines (SVMs).

One of the major problems faced indeveloping machine learning devices for B-cell epitope prediction is their variable length, mostly in the range of 6-25, of epitopes. Since machine learning techniques require fixed length ofpeptide, Akcesme, et. al. (2017) trained 20 different RANNs for each peptide length. To improve the accuracy, indeed they trained eleven experts for each length, 220 RANNs in total. Using majority voting in the committee, they achieved an accuracy of 97.20% in distinguishing

epitopes from non-epitopes. In this article we compared the success of ANNs in Akcesme, et. al. (2017), with the existing online methods that rely on residue properties based on hydrophilicity, flexibility, charge, mass, etc. Summing up the scores of residues from propensity vectors related to these physiochemical properties, finally they selectthe regions in an antigen protein, which have peaked scores and assign them as B-cell epitopes. Thesemethods are subjective in nature because they heavily depend on the thresholds used, and one can never be sure about the boundaries of epitopes.

The performance of the method that relies on committees of RNNs in that work is tested on five antigens with known linear b-epitopes, together with five well known online b-epitope predictors. It is seen that RNNs over scored four of other predictors, while there is no significant difference between successes of RNNs and the online b-epitope predictor ABCpred.

APPENDICES

Wet lab reported b-epitopes of five antigens.

A1.Mycobacterium Tuberculosis

MTEQQWNFAGIEAAA
NFAGIEAAASAIQGN
ASAIQGNVTSIHSLL
NVTSIHSLLDEGKQS
SLLDEGKQSLTKLAA
KQSLTKLAAAWGGSG
AAWGGSGSEAYQGVQ
GSEAYQGVQQKWDAT
QQKWDATATELNNAL
TATELNNALQNLART
ALQNLARTISEAGQA
TISEAGQAMASTEGN
QAMASTEGNVTGMFA

A2.Plasmodium Falciparum:

AEENVEENVEEVEENVEENV
AEENVEH
DDEHVEEPTVA
DDEHVEEPTVADDEHVEEPTVA
EENVEENV
EENVEENVEENVEENV
EENVEENVEENVEENVEENVEENV
EENVEHDA
EENVEHDAEENVEENV
EENVEHDAEENVEHDA
EENVEHDAEENVEHDAEENVEENV
EENVEHDAEENVEHDAEENVEHDA
EENVEHDAEENVEHDAEENVEHDAEENVEHDA
EKVDNLGRSGGDIIK
EKVDNLGRSGGDIIKKMQTLWDEIMDINKRK

IVGYIMHGISTINTEMK
LFDYNEKVDNLGRSG
LGRSGGDIIKKMQTL
MQTLWDEIMDINKRK
NADMNEITERYFKLAENYY
PTVAEEH
SLRWIFKHVAKTHLK
TVAEEHV
TVAEEHVEEPTVAEE
VAEEHVE
VEENVEE

A3. Meningitis

VDNQKQQHGALRNQGSRFHIKATHNFGD,
ARTRTTETGKGVKTEKEKSVGVGLRVYF,
FGDGFYAQGYLETRFVTKASENGSDNFGD,
FGDITSKYAYVTLGNKAFGEVKLGRAKT,
GEKTREQAVLFGVDHKLHKQLL,
GVLATLGYRFSDLGLLVSLDSGYAKT,
LSIIAEQSNSTVDNQK,
YAKTKNYKIKHEKRYFVSPGFQYEL,
YELMEDTNVYGNFKYERTSVDQGEKTR

A4. Human Polio Virus

ALTLSLPKQQDSLPDTKA,
ATNPLAPSDTVQTRHVVQ,
DNEQPTTRAQKLFAM,
KEVPALTAVETGATNPLA,
KHVRVWCPRPPRAVPYYG,
NGHALNQVYQIMYIPPGA,
QKLFAMWRITYKDTV,
QPTTRAQKLFAMWRI,
TRAQKLFAMWRITYK,
TRAQKLFAMWRITYKDTV,
TRHVVQRRSRSESTIESF,
TSKVRIYMKPKHVRVW,
VAIIEVDNEQPTTRAQKL,
VRVVNDHNPTKVTSKVRI,
YIPPGAPTPKSWDDYTWQ,
AAPARISVPYVGLA,
AHSKEVPALTAVET,
ARISVPYVGLANAY,
AYSHFYDGFAKVPL,
DFGVLAVRVVNDHN,
DTVQLRRKLEFFTY,
FAMWRITYKDTVQL,
HFYDGFAKVPLKTD,
HNPTKVTSKVRIYM,
IFYTYGAAPARISV,
LANAYSHFYDGFAK,
NPSIFYTYGAAPAR,
PKHVRVWCPRPPRA,
PSDTVQTRHVVQRR,
QKLFAMWRITYKDT,
QLRRKLEFFTYSRF,
QPTTRAQKLFAMWR,
RPPRAVPYYGPGVD,

SAMTVDDFGVLAVR,
SEVAQGALTLSLPK,
SVPYVGLANAYSHF,
TKVTSKVRIYMKPK,
TRAQKLFAMWRITY,
TSKVRIYMKPKHVR,
TSSNPSIFYTYGAA,
TVDDFGVLAVRVVN,
TVQTRHVVQRRSRS,
TWQTSSNPSIFYTY,
TYGAAPARISVPYV,
TYKDTVQLRRKLEF,
VLAVRVVNDHNPTK,
VNDHNPTKVTSKVR,
VRIYMKPKHVRVWC,
VRVVNDHNPTKVTS,
VRVWCPRPPRAVPY,
WCPRPPRAVPYYGP,
WRITYKDTVQLRRK,
YMKPKHVRVWCPRP,
YVGLANAYSHFYDG,
AYAPPGAQPPTSRK,
CGSMMATGKILVAY,
FLFCGSMMATGKIL,
HQGALGVFAIPEYC,
ILVAYAPPGAQPPT,
KFTFLFCGSMMATG,
PHQIINLRTNNSAT,
PPGAQPPTSRKEAM,
RAVPYYGPGVDYRN

A5. Plasmodium Vivax

AYFLLGPVVKTLFNK
EGGSEFSERIGNSLS
EVIGNELADNIANEIVSSLQK
FDVKTQLKATAKKVL
FNKVEDVLHKPIPDT
KVLVEALKAALEPTE
LALFCFVNVLSLRGK
LEEEEAEDEFSDELLD
LKAALEPTEKIVAST
LKATAKKVLVEALKA
LQKDSASFLQSGFDV
MHLFNKPPKGKMNKV
NEIVSSLQKDSASFL
NKVNRVSIICAFLALFCFVNV
PDTIWEYESKGSLEE
PPKGKMNKVNRVSII
PTEKIVASTIKPPRVSEDAYFLLGPVV
PVVKTLFNKVEDVLH
SERIGNSLSSFLSES
SESASLEVIGNELAD
SFLQSGFDVKTQLKA
SLSSFLSESASLEVI
TASSSLEGGSEFSER
VLHKPIPDTIWEYES
VNVLSLRGKSGSTAS

YESKGSLEEEEAEDE

REFERENCES

Adams HP, Koziol JA. Prediction of binding to MHC class I molecules. J Immunol Methods;185:181; 1995.

Akcesme, B. Akcesme, F. Adilovic, M., and Can, M. (2017) Recurrent Neural Networks for Linear B-Epitope Prediction in Antigens, 2 B. Southeast Europe Journal of Soft Computing Vol.6 No.2 (1-7).

Alix AJ. (1999) Predictive estimation of protein linear epitopes by using the program PEOPLE. Vaccine;18:311–314.

Balkin, S.D. (1997) Using recurrent neural networks for time series forecasting, Working Paper Series number 97–11, International Symposium on Forecasting, Barbados.

Bhasin M, Raghava GPS. (2003) SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. Bioinformatics;20:421.

Bhasin, M., and Raghava, G.P.S. (2004) Prediction of CTL epitopes using QM, SVM and ANN techniques, Vaccine 22, 3195–3204.

Blythe MJ, Flower DR. (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. ProtSci;14:246–248.

Brusic V, Rudy G, Harrison LC. (1994) Prediction of MHC binding peptides by using artificial neural networks. In: Complex mechanism of adaptation. Amsterdam: IOS Press. p. 253–60.

Caro-Aguilar, I., Rodríguez, A., Calvo-Calle, JM., Guzmán, F., De la Vega, P., Patarroyo, ME., Galinski, MR., and Moreno, A. (2002) Plasmodium vivax Promiscuous T-Helper Epitopes Defined and Evaluated as Linear Peptide Chimera Immunogens, Infection And Immunity, p. 3479–3492 Vol. 70, No. 7

Chandra, S., and Singh, TR., (2012) Linear B cell epitope prediction for epitope vaccine design against meningococcal disease and their computational validations through physicochemical properties, Netw Model Anal Health Inform Bioinforma 1:153–159.

Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33: 423–428.

Cornette JL, Margalit H, DeLisi C, Berzofsky JA. (1993) The amphipathic helix as a structural feature involved in T cell recognition. In: Epand RM, editor. The amphipathic helix. Boca Raton: CRC Press.

DeLisi C, Berzofsky JA. (1985) T-cell antigenic sites tend to be amphipathic structures. Proc Natl Acad Sci USA;82:7048.

El-Manzalawy Y, Dobbs D, Honavar V (2008) Predicting linear B-cell epitopes using string kernels. Journal of Molecular Recognition 21: 243–255.

El-Manzalawy Y, Dobbs D, Honavar V (2008) Predicting flexible length linear B cell epitopes. ComputSyst Bioinformatics Conf, 7:121-132

El-Manzalawy Y, Dobbs D, Honavar V (2008) Predicting linear B-cell epitopes using string kernels. J MolRecognit 21: 243-255.

Emini EA, Hughes JV, Perlow DS, Boger J. (1985) Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J Virol;55:836–839.

Flower DR. (2003) Towards in silico prediction of immunogenic epitopes.Trends Immunol;24:667–674.

Gulukota K, Sidney J, Sette A, DeLisi C. (1997) Two complementary methods for predicting peptides binding major histocompatibility complex molecules. J MolBiol; 267:1258.

Isea, R. (2017) Quantitative Prediction of Linear B-Cell Epitopes. Biomedical Statistics and Informatics. Vol. 2, No.1,, pp.1-3.

Karplus PA, Schulz GE. (1985) Prediction of chain flexibility in proteins: a tool for the selection of peptide antigen. Naturwissenschaften;72:212,213.

Kolaskar, AS., and Tongaonkar, PC. (1990) A semi-empirical method for prediction of antigenic determinants on protein antigens, Volume 276, number 1,2, 172-174 FEBS 09210.

Larsen, J. Lund, O and Nielsen, M (2006) Improved method for predicting linear B-cell epitopes, Immunome Research, vol. 2, article 2, W24–W29

Nomoto A., Omata T., Toyoda H., Kuge S., Horie H., Kataoka Y., Genba Y., Nakano Y., Imura N. (1982) Complete nucleotide sequence of the attenuated poliovirus Sabin 1 strain genome." Proc. Natl. Acad. Sci. U.S.A. 79:5793-5797.

Odorico M, Pellequer JL. BEPITOPE: predicting the location of continuous epitope and patterns in proteins. J MolRecognit;16:20–22; 2003.

Parker KC, Bednarek MA, Coligan JE. (1994)Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J Immunol;152:163.

Parker JMD, Guo D, Hodges RS. (1986) New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry;25: 5425–5432.

Pellequer J-L, Westhof E, Regenmortel MHV. (1991) Predicting location of continuous epitopes in proteins from their primary structures. Methods Enzymol;203:176–201.

Pellequer J-L, Westhof E, Regenmortel MHV. (1993) Correlation between the location of antigenic sites and the prediction of turns in proteins. Immunol Lett;36:83–99.

Pellequer J-L, Westhof E. PREDITOP: A program for antigenicity prediction. J MolGraphics;11:204–210; 1993.

Rammensee HG, Friede T, Stevanoviic S. (1995) MHC ligands and peptide motifs: first listing. Immunogenetics 41:178 .

Saha.S and Raghava G.P.S. BcePred: (2004) Prediction of Continuous B-Cell Epitopes in Antigenic Sequences Using Physico-chemical Properties. In G.Nicosia, V.Cutello, P.J. Bentley and J.Timis (Eds.) ICARIS 2004, LNCS 3239, 197-204, Springer.

Saha, S., and Raghava G. P. S. (2006) Prediction of Continuous B-Cell Epitopes in an Antigen Using Recurrent Neural Network, PROTEINS: Structure, Function, and Bioinformatics 65:40–48.

Saha S, and Raghava GP. (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins. 65(1):40-8.

Singh H, and Raghava GP. (2001)ProPred: prediction of HLA-DR binding sites. Bioinformatics;17:1236.

Singh H, Raghava GPS. (2003) ProPred1: prediction of promiscuous MHC class I binding sites. Bioinformatics;19:1009.

Spouge JL, Guy HR, Cornette JL, Margalit H, Cease K, Berzofsky JA, (1987) Strong conformational propensities enhance T cell antigenicity. J Immunol;138:204.

Stille CJ, Thomas LJ, Reyes VE, Humphreys RE. (1987) Hydrophobic strip of helix algorithm for selection of T cell-presented peptides. Mol Immunol; 24:1021.

VanRegenmortel MHV, de Marcillac GD. (1988) An assessment of prediction methods for locating continuous epitopes in proteins. Immunol Lett;17:95–107.

Yang, Y., Chen, H., Liu, Z., Ma, H., Qin, L., Jin, J., Zheng, R., Feng, Y., Cui, Z., Wang, J., Liu, J., Hu,.Z, A (2013)Novel B-Cell Epitope Identified within Mycobacterium tuberculosis CFP10/ESAT-6 Protein, Published: https://doi.org/10.1371/