Computational Methods in Linear B-cell Epitope Prediction

KavithaK V
Dept of Computer Science
College of Engineering
Thiruvananthapuram, India

Saritha R
Dept of Computer Science
College of Engineering
Thiruvananthapuram, India

Vinod Chandra S S
Computer Centre
University of Kerala
Thiruvananthapuram, India

ABSTRACT

Immune systems protect the body from foreign molecules known as antigens. It has great pattern recognition capability that may be used to distinguish between foreign cells entering the body (non- self or antigen) and the body cells (self). Any substance like proteins, polysaccharides, lipoproteins, polypeptides, nucleoproteins and nucleic acids that can induce the immune system to produce a corresponding antibody is called an antigen. This ability of antigen is called antigenicity. That portion of the antigen which can bind with the antigen binding site of the antibody is called B-cell epitope or antigenic determinant. B-cell epitopes can be linear or conformational. These epitopes play a vital role in the development of peptide vaccines, in diagnosis of diseases, immune based cancer therapies and also for allergy research. Since experimental methods of identifying epitopes are costly and time consuming, computational methods for prediction are desirable. This paper reviews various approaches like amino acid scale based methods and machine learning methods used for the prediction of linear B-cell epitopes.

General Terms

Machine Learning, Supervised Learning, Data Mining, Artificial Intelligence

Keywords

Immunity, B-cell epitopes, SVM, Aminoacid scale, Antigenicity

1. INTRODUCTION

Life is a battle field in which human beings are like soldiers, attacked from all sides by dreadful organisms such as bacteria, viruses, fungi etc. Such disease causing microbial agents are called pathogens.To protect us from the hazardous effects of these organisms, human body is equipped with a defence mechanism in the form of immune system. Thus the human body resists the invasion of pathogens and their toxic products. The study of immune system is called immunology. The foreign material or pathogen which enters the body and is capable of stimulating the immune system is called an antigen. It is also called immunogen. They are large-sized proteins or polysaccharides, present on the walls of bacteria and on the coats of viruses. An antigen stimulates the body to produce a specific antibody. They are also called antibody generating things. Epitopes also called immunogenic determinants are the portion of antigens that can bind with the antigen binding site of immunoglobulin [1].

The protective molecules produced by the body in response to antigens are called antibodies. They are globular proteins and are also called immunoglobulins (Ig). Antibodies are always antigen specific. Antigen and antibody have complementary reactive sites that fit together like lock and key. Antibodies react with antigens and make them inactive or harmless. The B-lymphocytes are responsible for the production of antibodies in response to pathogens [2]. The immunoglobulins fight against the bacteria chiefly by three mechanisms like agglutination (binding with bacteria or viruses); opsonisation (form a coat on microbes to facilitate phagocytosis by cells) and neutralization (neutralize toxins from microbes).

2. IMMUNITY

Immunity is the inborn or acquired resistance of living organisms to infection of microorganisms.

2.1 Types of Immunity

Immunity is broadly classified into innate immunity and acquired immunity [3]. Innate immunity is the natural or inborn resistance of the body against infections. Resistance that an individual acquires during his life is known as acquired or adaptive immunity. In this type of immunity, specific antibodies are produced in response to specific antigens. Adaptive immunity is further classified into active and passive immunity. The long-lasting immunity, developed by the antibodies produced by the organism’s own cells, is called active immunity. The immunity developed in bodies by the inoculation of readymade antibodies produced in the plasma of an animal or human is called passive immunity. It is less efficient and inferior than active immunity.

2.2 Types of Immune Response

There are two types of immune response. They are humoral immunity and cell-mediated immunity. The immune reaction mediated by B-lymphocytes is called humoral immunity or antibody mediated immune system (AMIS). It involves the production of specific antibodies by plasma cells in response to specific antigen that enters in to the body. The immune reaction mediated by T-lymphocytes is called cell mediated immunity. It is usually involved in the destruction of infected cells and cancer cells and in graft rejection. When the immune system encounters a foreign molecule for the first time, it generates an immune response to eliminate the invader. This is called primary immune response. At the same time, immune system produces memory cells too. When same invader is encountered for the second time, memory cells immediately produce a heightened secondary immune response. This is called immunological memory [4].

2.3 Cells of the Immune System

Lymphocytes are the main cells involved in the immune system. There are two kinds of Lymphocytes. They are called T–cells and B-cells. [5] The lymphocytes which are differentiated in the thymus are called T-cells. They are responsible for cellular immunity. The lymphocytes which are differentiated in the gut-associated bursal lymphoid tissues are called B-cells. B-cells produce antibodies and inactivate the antigens. They are responsible for humoral immunity.

2.4 How B-cells Respond to Antigens

There are many types of B-cells in the body. Each B-cell is antigen specific and on their plasma membrane, there are receptors for specific antigen. T-cells are also antigen specific. Fig 1 shows how the immune system defends the body. When T-cells come into contact with a specific antigen, the receptors on their membrane recognize the antigen. The T-cells stimulated by the contact of the antigen divide rapidly to give rise to a group of T-cells called a clone of T-cells. The Helper T-cells can in turn stimulate the B-cells. When B-cell is stimulated by the antigen, it gives rise to a clone of plasma cells. These plasma cells produce antibodies at the rate of 2000molecules/second. These antibodies circulate in the body fluids and inactivate the antigen by binding to it. The inactivated antigen-antibody complex is engulfed by phagocytes. B-cells are short lived and are constantly replaced by new cells from the bone marrow.

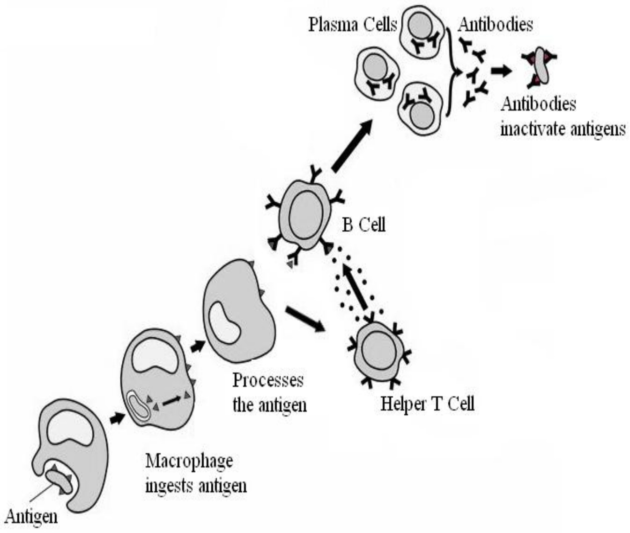

Fig1: How the immune system defends the body

3. B-CELL EPITOPE PREDICTION

B-cell epitope prediction is important in various research areas like vaccine design, immunodiagnostics and allergy research. Experimental methods can be used for prediction of epitopes but such methods are costly and timely process to find the antigen-antibody reactive sites. Computational methods accelerate reliable prediction of epitopes for experimental design. It is also a critical challenge in immunoinformatics and computational biology. Such methods are cost effective and less time consuming. B-cell epitopes can be linear or conformational by its structure. Linear epitopes consists of linear sequence of amino acids that can be recognized by antibodies. Conformational epitopes consists of amino acid residues that are distantly separated in the sequence but are brought into physical proximity via protein folding [6]. Even though ninety percent of epitopes are conformational, researchers are interested in linear epitopes.

4. RELATED WORK

There are different methods for predicting linear B-cell epitopes. Amino acid scale based methods and machine learning approaches are the computational methods for predicting linear epitopes.

4.1 Amino Acid Scale Based Methods

In this method, the location of B-cell epitopes can be identified based on correlation between physico-chemical properties and antigenic determinants in protein sequence [7]. Amino acid scale-based methods apply amino acid scales to compute the scores of a residue i in a given protein sequence. The i - (n - 1)/2 neighbouring residues on each side of residue i is used to calculate the score for residue i in a window of size n. The final score for residue i is the average of the scale values for n amino acids in the window. These values are used as the basis for predicting whether the given amino acid residue is likely to be a part of linear B-Cell Epitope. The epitope predictions are based on the scale values for each of the 20 amino acids.

Hopp and Woods introduced the first amino acid scale based method for linear B-cell epitope prediction [8]. They utilized Levitt hydrophilicity scale to assign a scale value to each amino acid. This method is based on the statement that hydrophilic regions in the protein are predominantly located on the surface and are potentially antigenic. Subsequently, several other amino acid scales have been proposed for linear B-cell epitope prediction. The locations of continuous epitopes have been correlated with various parameters. Hydrophilicity, flexibility, accessibility, turns, exposed surface, polarity, and antigenic propensities of polypeptide chain are physico-chemical properties with which the locations of linear epitopes have been correlated. Parker's hydrophilicity scale [9] was determined experimentally using high-performance liquid chromatography (HPLC) on a set of 20 synthetic peptides accounting for each of the 20 amino acids. Karplus and Schulz's flexibility scale [10] was constructed on the basis of protein segments derived from known temperature B factors of α Carbons of 31 proteins of known structure. Emini's solvent accessibility scores [11] were calculated based on surface accessibility scale. This scale has been determined by Janin and Wodak and reflected surface exposure probabilities for amino acids are computed on X-ray structures of 28 proteins [12]. A surface probability

(Sn) at a sequence position n is defined as the product of fractional surface probabilities for amino acids at positions from n - 2 to n + 3. Surface probability for a random hexapeptide is equal to 1.0 with probabilities greater than 1.0 indicating an increased probability for being found on the surface. Chou and Fasman's method [13] is based on calculating the probability of a stretch of residues to be a part of secondary structure β turn.

Peptides also possess the antigenic character and are antibody responsive based on which epitopes can be predicted. Antigenicity prediction was carried out using Kolaskar and Tongaonkar antigenicity scale [14]. This prediction is based on a semi-empirical approach, developed on physicochemical properties of amino acid residues i.e. hydrophilicity, accessibility and flexibility and their frequencies of occurrence in 156 experimentally determined epitopes from 34 different proteins. This approach has the efficiency to detect antigenic peptides with about 75% accuracy.

Based on the combinations of various physico-chemical properties, prediction methods like PREDITOP [15], PEOPLE [16], BEPITOPE [17] and BcePred [18] were designed to predict linear B-cell epitopes. In all these methods, a common feature they have is calculating the average amino acid scale value over a sliding window along a query protein sequence. The peak of the resulting profile is considered to correspond to the location of the B-cell epitope for the protein concerned. The values of the area under the receiver operating characteristic curve (AROC) for these methods did not exceed 0.60.

To study the correlation between location of linear epitopes and amino acid scale based profiles, Blythe and Flower have performed an extensive estimation of 484 amino acid propensity scales in a dataset of 50 proteins [19]. Their study found that by combining propensity scales, prediction accuracy could not be improved to a great extent. Only some best combinations of scale values produced a better result. The performance of propensity scale based methods is also optimistic. This is due to the small size of the datasets on which the methods had been evaluated. Therefore more sophisticated machine learning approaches for predicting linear B-cell epitopes need to be developed. Careful evaluation of methods should also be done in order to progress the state-of-the-art in linear B-cell epitope prediction.

4.2 Machine Learning Approaches

Machine learning approaches are used to improve the accuracy of linear B-cell epitope prediction based on the availability of experimentally identified linear B-cell epitopes. BepiPred [20] combines two amino acid propensity scales Parker hydrophilicity scale and Levitt secondary structure and a Hidden Markov Model (HMM) for epitope prediction. Bepipred was trained on linear epitopes. But there was only a slight improvement in prediction accuracy comparative to techniques that rely on study of amino acid physicochemical properties.

The length of B-cell epitopes varies from 5 to 30. For prediction fixed length epitopes are appropriate. So a truncation-extension treatment was adopted. According to such an approach, to create 20 length pattern there is a need to equally truncate the surplus residues at both N- and C-terminals, if the epitope length is longer than 20 amino acids. If the epitope length is less than 20 amino acids, then the length is increased by equally extending the peptide segments toward both the N- and C-terminals along the protein chain until it reach 20 residues long.

Artificial neural network was used in ABCPred [21] for predicting linear B-cell epitopes. A non-redundant data set of 700 B-cell epitopes obtained from Bcipep database and 700 non-epitope peptides obtained randomly from Swiss Prot database was used for prediction. Both feed-forward and recurrent neural networks were evaluated on this dataset using 5-fold cross validation tests. Input sequence windows ranging from 10 to 20 amino acids, were tested and the best performance, 65.93% accuracy, was obtained using a recurrent neural network trained on peptides of length 16.

Two machine learning methods, decision trees and nearest-neighbour method were tested by Sollner et.al [22]. They combined these methods with feature selection on 1478 attributes extracted from a variety of propensity scales, neighbourhood matrices, and respective probability and likelihood values. The accuracy is 72% when tested on a dataset of 1211 B-cell epitopes and 1211 non epitopes using five-fold cross-validation.

Cheng proposed a new scale called amino acid pair (AAP) propensity scale [23] for predicting linear B-cell epitiopes. The B-cell epitope data set was taken from Bcipep database [24], which is a collection of experimentally determined B-cell epitopes. He developed an amino acid pair (AAP) antigenicity scale that assigns to each dipeptide a propensity value. AAPs are generated by decomposing the peptides of proteins continuously. Cheng's dataset consists of 872 positive epitopes and 872 negative non epitopes. He proved that using SVM (support vector machine) classifier, the AAP antigenicity scale approach has an accuracy of 71% when AAP propensity scale was only considered, but an accuracy of 72.5% when AAP scale is combined with turns, antigenicity, flexibility, hydrophilicity and accessibility. The relevant parameters used for SVM in the combination method were C =32 and $\sigma^2 = 2$.

AAT-fs [25] developed for predicting linear epitopes was based on the amino acid triplet (AAT) antigenicity. After using AAT scale to create input vectors, a Support Vector Machine (SVM) was developed for the classification which is trained utilizing Radial Basis Function kernel on homology reduced datasets with fivefold cross validation. The AAT-fs method gets the better performance with an accuracy of 74% than AAP scale, and other existing B-cell epitope prediction algorithms.

El-Manzalawy et. al introduced four kernel functions into SVM, including spectrum kernel, mismatch kernel, local alignment kernel and subsequence kernel. The model using subsequence kernel, named BCPred [26], gave out the best results. The data set they used is a homology-reduced data set of 701 linear B-cell epitopes, extracted from Bcipep database, and 701 non-epitopes, randomly extracted from SwissProt. BCPred predict 12, 14, 16, 18, and 20-mer long epitopes directly from sequence using a new type of string kernel-based SVM. The best accuracy 75.8% they got using the subsequence kernel. EL-Manzalawy also tried to construct flexible length B-cell epitopes prediction models in two ways. One way is using kernel functions which can deal with the flexible length epitopes directly. Four kernel functions were used for flexible length epitopes prediction as well, and their performances were evaluated. The other way is mapping flexible length sequences into fixed length feature vectors. Among all methods, the model based on the subsequence kernel named FBCPred [27] gave out the best results.

BPairwise [28] was developed to predict flexible length linear B-cell epitopes. Here an encoding scheme based on pair wise sequence similarity using Smith Waterman algorithm was adopted, which can transform the flexible length peptides into fixed-length feature vectors. Support vector machine (SVM) was then used as the classification engine to construct prediction models. This method gave an accuracy of 66%.

BayesB [29] is a Support Vector Machines (SVM) prediction model employing Bayes Feature Extraction to predict linear B-cell epitopes of diverse lengths .The length varies from 12 to 20. An accuracy of 74.50% was achieved in this method.

Linear Epitope Prediction System (LEPS) is a prediction model [30] based on physico-chemical properties and Support Vector Machine. Datasets of epitope and non epitope segments with 2, 3 and 4 residues in length were trained and applied as statistical features of SVM. AntiJen, HIV and PC database were used in this model and better specificity, accuracy, and positive prediction value (PPV) was achieved in most testing cases. High specificity and PPV of a linear epitope prediction can lead to an efficient and effective design on biological experiments.

A comparison of known linear epitope prediction models based on features used and machine learning techniques is summarized in Table 1. The machine learning techniques used in most of the methods are Hidden Markov Model (HMM), Artificial Neural Networks (ANN) and Support Vector Machines (SVM). Features commonly used in all these methods are physico chemical properties and antigenicity scales.

Table 1: Comparison of machine learning approaches in epitope prediction

Prediction Method	**Features used**	**Machine Learning Technique**
BepiPred	Parker hydrophilicity scale and Levitt secondary structure	Hidden Markov model
ABCPred	Hydrophilicity, accessibility, flexibility, turns, antigenicity, polarity	Feed Forward and recurrent Neural network
Cheng et.al method	Hydrophilicity, accessibility, flexibility, turns, antigenicity, Amino Acid Pair (AAP) antigenicity scale	Support Vector Machine(SVM)
BCPred	Hydrophilicity, accessibility, flexibility, turns, antigenicity, Amino Acid Pair(AAP) antigenicity scale	Subsequence kernel based SVM
AAT-fs	Amino acid triplet (AAT) antigenicity scale	Radial Basis Kernel based SVM
BayesB	Relative position specific amino acid propensity of a dipeptide	SVM employing Bayes Feature Extraction
LEPS	Hydropathy, accessibility, flexibility, turns, antigenicity, polarity, dipeptide, tripeptide and tetrapeptide antigenicity	Radial Basis Kernel based SVM

5. CONCLUSION

In this paper the various approaches for linear B-cell epitope prediction have been studied. Feature selection is important in epitope prediction. Different types of features and different machine learning approaches like HMM, SVM etc have been used in all these approaches. It is clear that performance accuracy varies based on features selected and learning techniques used. By combining the various features used in all these approaches and doing a Principal Component Analysis, several features which do not play a major role in epitope prediction can be filtered out. Thus a reliable computational model for predicting linear B-cell epitopes could be designed in future.

6. REFERENCES

[1] Parham P. The Immune System, 2nd edition, Garland Science Publishing, NewYork, NY, 2005.

[2] Li J, Y.A. Zhang, H. Boshra, A.E. Gelman, S. LaPatra,,L. Tort and J.O. Sunyer. "B lymphocytes from early vertebrates have potent phagocytic and microbicidal abilities". Nature Immunology vol 7, pp 1116–1124.

[3] Janeway, C. A., Jr.; Travers, P.; Walport, M.; and Shlomchik. Immunobiology, 5th edition, Garland Science Publishing, NewYork, NY, 2001.

[4] R.Ahmed and J.Sprent, "Immunological Memory", The Immunologist, vol 7, pp. 23-26, 1999.

[5] L .N. D. Castron and F. J.V. Zuben, "Artificial Immune Systems Part I: Basic theory and applications", Technical Report TR-DCA 01/99, Dec 1999.

[6] J. Greenbaum, P. Andersen, M. Blythe, H. Bui, R. Cachau,J. Crowe M. Davies, A. Kolaskar, O. Lund, S. Morrison, et al. " Towards a consensus on datasets and evaluation metric for developing B-cell epitope prediction tools". Journal of Molecular Recognition, vol 20, pp 75–82, 2007.

[7] Y. EL-Manzalawy and V.Honavar, "Recent advances in B-cell epitope prediction methods", Immunome Research, vol 6, Nov 2010.

[8] T.P. Hopp and YK.R.Woods, "Prediction of protein antigenic determinants from amino acid sequences" PNAS, vol 78, pp 3824-3828, 1981.

[9] J.M. Parker, D. Guo, R.S. Hodges, "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, vol 25, pp 5425-5432, 1986.

[10] P.A.Karplus, and G.E. Schulz, "Prediction of chain flexibility in proteins – a tool for the selection of peptide antigens", Naturwissenschaft, vol 72, pp 212-213, 1985.

[11] Yemini, J.Hughes, D.Perlow and J. Boger, "Induction of hepatitis A virus neutralizing antibody by a virus-specific synthetic peptide." Journal of Virology, vol55, pp 836-839, 1985.

[12] Janin, J. and Wodak, S.: Conformation of amino acid side-chains in proteins. Journal of Molecular Biology, vol 125, pp 357-86, 1978.

[13] P.Y.Chou and G.D.Fasman, "Prediction of secondary structure of proteins from amino acid sequences", Biochemistry, vol 47, pp 145–148, 1978.

[14] A.S. Kolaskar, P.C. Tongaonkar,,"A semi-empirical method for prediction of antigenic determinants on protein antigens", FEBS Lett,vol 276,pp172-174,1990

[15] J.Pellequer, E.Westhof and M.Van Regenmortel, "Correlation between the location of antigenic sites and the prediction of turns in proteins". Immunology, vol36, pp 83–99, 1993.

[16] A. Alix, "Predictive estimation of protein linear epitopes by using the program PEOPLE", Vaccine, vol 18, pp 311-14, 1999.

[17] M. Odorico, and J. Pellequer, "BEPITOPE: predicting the location of continuous epitopes and patterns in proteins." Journal of Molecular Recognition, vol 16, pp 20–22, 2003.

[18] S.Saha and G.Raghava, "BcePred: Prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties." International Conference on Artificial Immune System 2004, vol 3239, pp 197–204, 2004.

[19] M. J. Blythe and D.R. Flower, "Benchmarking B cellepitope prediction" Protein Science vol 14, pp 246-248, 2005.

[20] J.E.P.Larson, O.Lund and M.Neilsen, "Improved Method for predicting linear B-cell epitopes" Immunome Research. 2:2, 2006

[21] S. Saha and G. Raghava, "Prediction of continuous B-cell epitopes in an antigen using recurrent neural network." Proteins, vol 65: pp 40-48, 2006.

[22] J.Sollner and B. Mayer, "Machine learning approaches for prediction of linear B-cell epitopes on proteins". Journal of Molecular Recognition., vol 19, pp 200-208, 2006.

[23] J.Chen, H. Liu,J.Yang and K.Chou, "Prediction of linear B-cell epitopes using amino acid pair antigenicity scale"Amino Acids, vol 33, pp 423-428,2007.

[24] S. Saha and G. Raghava,"Bcipep: A database of B-cell epitopes", BMC Genomics, Vol 6, pp 79, 2005.

[25] L.Wang, J.Liu, S.Zhu and Y.Y.Gao, "Prediction of Linear B-cell epitopes using AAT scale" Third International Conference on Bioinformatics and Biomedical Engineering, ICBBE, pp 1-4, 2009.

[26] Y.EL-Manzalawy, D.Dobbs and V.Honavar, "Predicting linear B-cell epitopes using string kernels". J. Mol. Recognit., vol 21, pp 243-255, 2008.

[27] Y.EL-Manzalawy, D.Dobbs and V.Honavar, "Predicting flexible length linear Bcellepitopes"7th International Conference on Computational Systems Bioinformatics, pp 121-131, 2008.

[28] W.Zhang and Y.Niu,"Predicting flexible length linear B-cell epitopes using pair wise sequence similarity", Third International Conference on Biomedical engineering and Informatics, 2010.

[29] L. JK. Wee, D.Simarmata,,Y. Kam, F P. Lisa and J. C. Tong "SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction" BMC Genomics, vol 11,Dec 2010.

[30] H.W.Wang,Y.C.Lin and H.T.Chang," Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification"Journal of Biomedicine and Biotechnology, June 2011.