Journal of Computational Intelligence in Bioinformatics.
ISSN: 0973-385X Volume 1 Number 1 (2008) pp. 45–53

http://www.ripublication.com/jcib.htm

Benchmarking the Propensity Scales for the Prediction of Linear B-Cell Epitopes

Satarudra Prakash Singh[1], Sanjay Tyagi[2], Feroz Khan[3] and B.N Mishra[4]*

[1]Amity Institute of Biotechnology, Amity University, Lucknow, Uttar Pradesh
[2, 3, 4] Department of Biotechnology, Institute of Engineering & Technology, Lucknow –226021, Uttar Pradesh, India
**Corresponding author email: sprakashsingh@gmail.com*

Abstract

Subunit vaccine designing is an integral part of vaccine design strategy which requires the identification of T and B-cell epitopes in an antigenic protein sequence. Although conformational aspects of antibody binding complicates the problem of experimental B-cell epitope identification, the prediction of linear B-cell epitopes is useful in the development of peptide based vaccine and immunodiagnostics. Available computational methods for prediction of linear- B cell epitope are bound to use for a fixed window size (epitope length) and they are not evaluated comprehensively on the dataset of experimental B cell epitopes.

The present work is related to analyze length of B cell epitopes found in Bcipep database which ranges from 5-20 amino acids. In this study effort has been made to benchmark propensity scales of six physico-chemical properties viz. hydrophilicity, flexibility, polarity, turns, exposed surface and antigenic propensity, which are being used to predict linear B cell epitopes in protein sequence. The prediction accuracy of the individual propensity scale for these six physico-chemical properties varies from 51.00% to 58.37%. Their performances are also evaluated by ROC curve for the thresholds ranging from -3.0 to + 3.0. Out of six physico-chemical properties, flexibility is the most accurate scale because of its maximum Aroc value (0.60). In order to improve the prediction accuracy a combined scale is used taking in to account the flexibility (Karplus et.al) with hydrophilicity (Parker et. al.), Polarity (Grantham) and Exposed surface (Jenin) which shows better accuracy (59.74%) at a threshold 0.65 to the other possible combinations.

Keywords: Peptide vaccine, epitope, B-cell, ROC.

Introduction

In the era of genomics and proteomics, peptide based vaccine designing and immunodiagnosis is the most effective for diseases ranging from malaria to cancer (De Groot et al. 2002). It does critically require identification of regions in the pathogen native protein sequences, which are recognized by either B-cell or T-cell receptors (Schirle et al. 2001). The antigenic regions of protein recognized by the binding sites of immunoglobulin molecules are called B-cell epitopes (Van Regenmortel 1993). B-cell epitopes can be classified into two categories; i)conformational/discontinuous epitope, where residues are distantly separated in the sequence and brought into physical proximity by protein folding and ii) linear/continuous epitope, comprised of a single continuous stretch of amino acids (a.a) within a protein sequence that can react with anti-protein anti-bodies (Barlow,D.J. et. al. 1986). Most of the B-cell epitopes were thought to be discontinuous. However, in late 1980s it was shown that this conformational restriction is not a necessary condition for the production of protein-reactive anti-peptide antibodies (Walter G, 1986-88). The designing of the conformational epitopes is difficult and so experimental B-cell epitopes largely include linear epitopes. These linear epitopes can be exploited in the development of synthetic vaccines or disease diagnosis. These epitopes are also important for allergy research and in determining cross-reactivity of IgE-type epitopes of allergens (Selo et al. 1999). A number of vaccines based on B-cell epitopes are currently under clinical phase trials against viruses (El Kasmi K.C. and Muller, C.P 2001), bacteria (Sabhanini et al. 2003) and cancer (Kieber-Emmons et al. 1999). The experimental identification of epitopes binding specifically to anti-peptide antibodies requires the binding assay of each peptide in an antigenic protein sequence which are very laborious and time consuming .

A bioinformatics approach to predict linear B cell epitope in a protein sequence can be the best alternative to reduce the number of peptides to be synthesized for wet lab experimentation. In the past, numbers of computational methods and programs have been developed for predicting linear B-cell epitopes, which are based on hydrophilicity, accessibility, flexibility, or secondary structure propensities scales of the 20 natural amino acids (Pellequer JL and Westhof, 1993; Alix AJ, 1999). Recently BEPITOPE program has been developed to predict continuous B cell epitopes and searching for patterns in either a single protein or a complete translated genome (Odorico M. and Pellequer JL, 2003). An Alternative strategy for predicting linear B-cell epitope is ABCpred (http://www.imtech.res.in/raghava/abcpred/) which uses a neural network trained and tested on the Bcipep, B-cell epitope database (GPS Raghava et al., 2004).

Presently, it is difficult to analyze that particularly which propensity scale or method is better than the other in order to predict B cell epitope, because there is no benchmarking of existing methods on the same dataset and normalized scales. Raghava GPS et al. (2004) evaluated the various residue properties like hydrophilicity, flexibility, polarity, exposed surface, turns, antigenic propensity and surface accessibility which are commonly used in these existing methods and a web server BcePred has been developed for predicting B cell epitopes in an antigenic

sequence (http://www.imtech.res.in/raghava/bcepred/)**.** Blythe and Flower (2005) also performed an exhaustive assessment of amino acid propensity scale using B-cell epitope database AntiJen. Practically it is not possible to evaluate all these methods and programs in their original form because many of these programs are not freely available, while some provides only qualitative information (visualization etc.) rather than quantitative. Most of these programs are not automatic server where you can give the query sequence and get the predicted epitopes.

The aim of present work is to provide a detail analysis of B epitope length in the database Bcipep (GPS Raghava et al., 2004) and benchmark the six propensity scales of two different methods using threshold independent parameter area under ROC curve.

Materials and Methods

Dataset of B-cell epitopes and non epitopes

The dataset of experimental linear B-cell epitopes were downloaded from the Bcipep database for epitopes length analysis and evaluation of scales/methods used in the study [http://www.imtech.res.in/raghava/bcipep/].The non B-cell epitope datasets were randomly generated from Swiss-Port protein sequence database (Bairoch, A and Apweiler, R; 2000).

Normalization of Protscale scales

The physico-chemical properties of amino acid scale were taken from Protscale, which allow the computational representation of the motif produced by any amino acid scale on a selected protein [http://www.expasy.org/tools/protscale.html]. The profile consists of 20 values assigned to each of amino acids on the basis of their relative propensity, as described by the scale. The original values of each scale were set between +3 to -3 using the formula;

Normalization Score = [(score-min)/ (max-min) – 0.5]*6

Where,

Score- value of residue in the given scale,

Max - maximum value

Min - minimum value

Algorithm: The following steps were used to predict the linear B-cell epitopes in amino acid sequence of protein,

1. The input amino acid sequence of a protein/peptide(s) was parsed into overlapping peptide fragments of selected window size (5-20). For example, if window size is seven, then the input sequence would be parsed into (n-7+1) number of overlapping peptides, where n is the number of a. a in the input sequence.
2. Scoring of each peptide fragment using specific profile score of scale/ method and if score of a peptide is greater than the selected threshold value, then the

peptide is predicted as a linear B-cell epitope .

3. In order to benchmark the selected scales/methods followings threshold dependent parameters were used:

Sensitivity (SE): Sensitivity is defined as the fraction of experimentally identified epitopes that are correctly predicted as an epitope. Higher sensitivity means that almost all of the potential epitopes shall be included in the predicted result.

$$SE= TP/ (TP+FN)$$

Where, TP and FN refer to the true positives and false negatives respectively. **Specificity (SP):** Specificity is defined as the fraction of experimentally identified non-epitopes that are correctly predicted as a non-epitopes.

$$SP=TN/ (TN+FP)$$

Where, TN and FP refers to true negatives and false positives respectively. **Accuracy:** A commonly used parameter to measure the prediction performance is accuracy, defined as

$$Acc= (TP+TN)/ (AP+AN)$$

Where, AP and AN refers to actual positives and actual negatives respectively.

The major problem with the threshold dependent parameter is that they only measure the performance at a given threshold and it is also found that the prediction results varies with thresholds, so a ROC curve was plotted between sensitivity (true positive fraction) on the y-axis against (1- specificity) values (false positive fraction) on the x-axis for all available thresholds. A single threshold independent parameter, area under the ROC curve (A_{roc}) measures the discrimination ability of a method to correctly classify epitopes and non-epitopes in a given dataset. As a condition, if $A_{roc}> 0.5$, the method/scale has the ability to predict the epitopes.

Results and Discussions

The Bcipep database analysis shows that the maximum numbers of epitopes are found in the range 5-20 amino acids (figure1.0). In the benchmark analysis of propensity scales to predict the linear B-cell epitopes we have used twelve scales of six physico-chemical properties (Hydrophilicity, flexibility, polarity, turns, exposed surface and antigenic propensity), i.e. two methods for each physico-chemical property scales. The sensitivity and specificity of different scales/methods were calculated on the B cell epitope datasets of Bcipep database (data not shown) and it is also found that the prediction results varies with thresholds, so it is very difficult to know which particular scale/method is better for the prediction of linear B-cell epitopes. From the ROC curves plotted (not shown) for the selected scales/methods, it was observed that Parker et al. method for hydrophilicity scale, Karplus and Schulz method for

flexibility scale, Grantham method for polarity scale, Levitt method for turns scale, Jenin method for exposed surface scale and Kolaskar method for antigenic propensity scale are better suited for the prediction of linear B-cell epitopes.

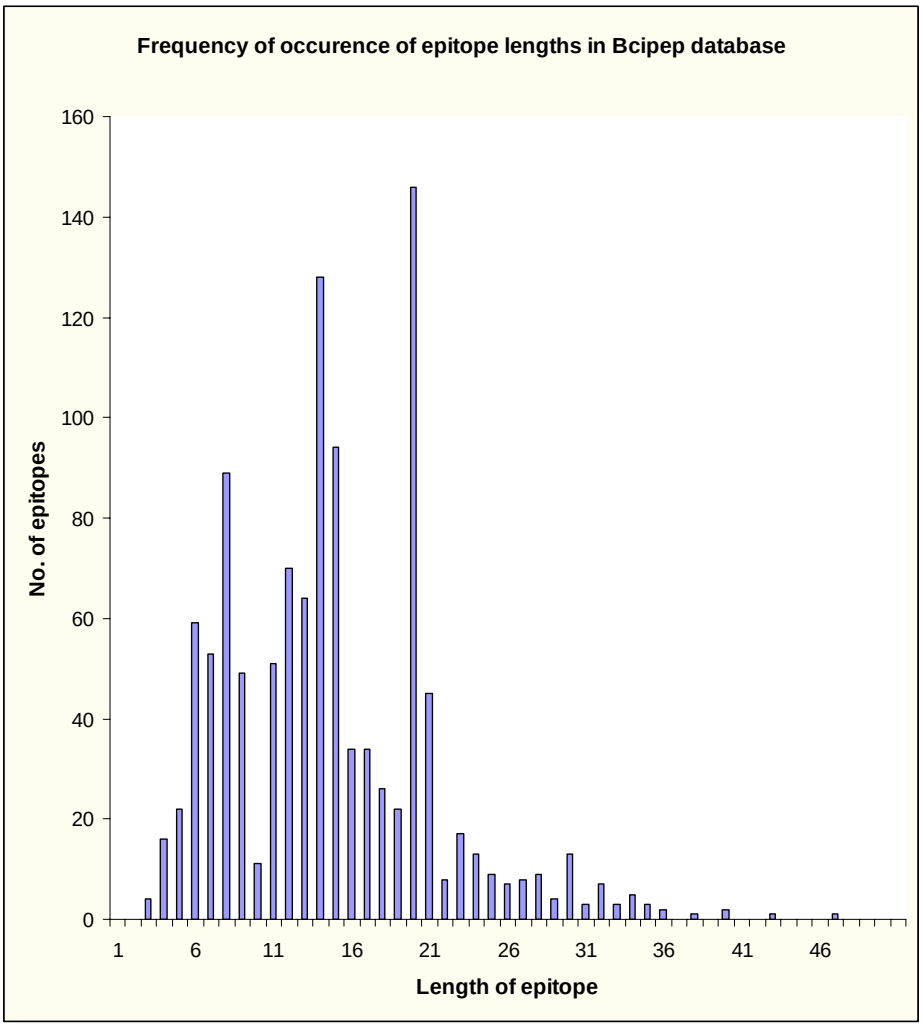

Figure 1.0 Frequency of occurrence of epitope lengths in Bcipep database

In order to improve the prediction accuracy, different combinations of scales have been used in group and it is found that combination of scales flexibility(Karplus et.al) with hydrophilicity(Parker et. al.), Polarity(Grantham) and Exposed surface(Jenin) shown better accuracy (59.74%) than the other combinations at a threshold of 0.65. The predictive performances of individual as well as combined method/scale are shown in Table 1.0. A ROC curve area is plotted (figure 2.0) for Karplus method of flexibility scale (Aroc=0.60) which has comparatively much better prediction than others scale/method.

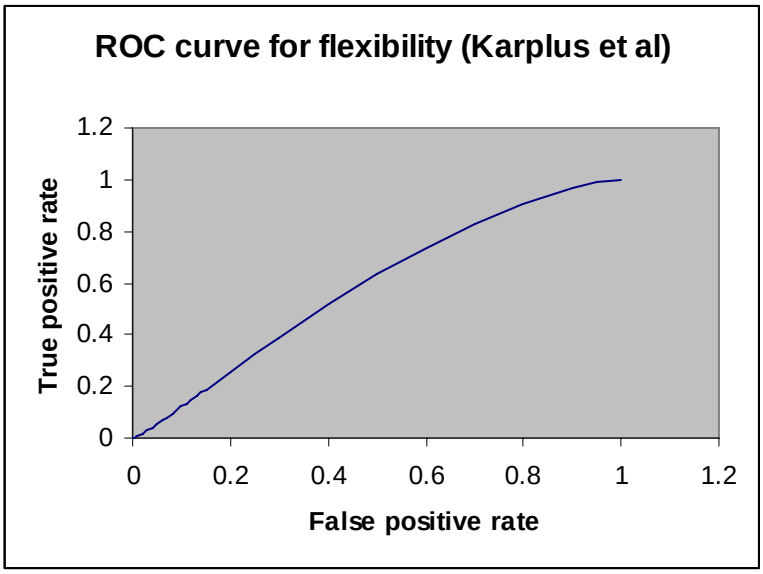

Figure 2.0 ROC curve for the best amino acid propensity scale

Table 1.0 The predictive performance of individual and combined propensity scales

Propensity scale	Method	Optimized Threshold	Sensitivity (%)	Specificity (%)	Accuracy (%)
Hydrophilicity	Parker et al (1)	0.75	36.07	75.06	55.58
	Hopp & Wood (1a)	0.65	36.00	74.21	55.11
Flexibility	Karplus et al (2)	0.65	46.82	69.91	58.37
	Ponnuswamy et al. (2a)	0.65	45.60	70.60	58.11
Polarity	Ponnuswamy et al. (3)	0.65	30.00	77.00	53.50
	Grantham (3a)	0.70	35.01	76.62	55.81
Exposed Surface	Janin J. (4)	0.75	35.00	74.00	54.50
	Radzicka et al. (4a)	0.70	30.05	72.00	51.00
Turns	Pellequer et al. (5)	0.70	26.62	78.00	52.31
	Levitt M. (5a)	0.70	30.52	76.00	53.26
Antigenic Propensity	Kolaskar et al. (6)	0.60	55.00	57.00	56.00
	Welling et al. (6a)	0.65	53.05	57.79	55.42
Combined methods/scales (2+1+3a+4)		0.65	55.52	62.52	59.74

Conclusions

The accurate prediction of linear B-cell epitopes is crucial in the peptide based vaccine and immunodiagnostics development for diseases ranging from malaria to cancer. Presently, it is difficult to analyze which scale/method is better than the others because there is no benchmarking of existing methods on the same dataset and normalized scale.

In order to benchmark, we have used two scales for six physico-chemical properties (Hydrophilicity, flexibility, polarity, turns, exposed surface and antigenic propensity) which have been frequently used for predicting continuous B-cell epitopes. The performances of propensity scales are tested on the same dataset of Bcipep database which accuracy is varying with 51.00% to 58.37% at optimized threshold. Out of these selected scales/methods, flexibility (Karplus et.al) is the most accurate because of maximum area under ROC curve (Aroc=0.60). The combined scales flexibility (Karplus et.al) with hydrophilicity (Parker et. al.), Polarity (Grantham) and Exposed surface (Jenin) shows better accuracy (59.74%) at a threshold 0.65 compared to the other combinations.

The major limitation on accuracy of the existing tool to predict linear B cell epitope is the lack of more B cell epitope dataset for training the algorithms. Given the weak predictive performance of the profiling (propensity scales) method demonstrated here, the development of more sophisticated approaches like machine learning (ANN, SVM etc.) algorithm is required to address this need and should be an obtainable goal comparable to T cell epitope prediction accuracy ~90%.

Acknowledgement

We acknowledge U.P Technical University, Lucknow and Amity University, Uttar Pradesh, Lucknow to support the research at the Institute of Engineering & Technology, Lucknow.

References

[1] Alix AJ., 1999 ‘Predictive estimation of protein linear epitopes by using the program PEOPLE’,*Vaccine*,Vol 18, pp 311-314.

[2] Bairoch,A. and Apweiler,R 2000 ‘The SWISS-PROT protein sequence database and its supplement TrEMBL’, Nucleic Acids Res.,Vol 28, pp45-48.

[3] Barlow,D.J., Edwards, M.S., and Thornton,J.M., 1986 ‘Continuous and discontinuous protein antigenic determinants’. Nature, Vol 322, pp747-748.

[4] Blythe M.J. and Flower D.R., 2005 ‘Benchmarking B cell epitope prediction: Underperformance of existing methods’ Protein Science,Vol 14, pp 246-248.

[5] Corcoran,A., Mahon,B.P. and Doyle,S., 2004 ‘B cell memory is directed toward conformational epitopes of parvovirus B19 capsid proteins and the unique region of VP1’, *J Infect Di*s. vol189, pp1873-1880

[6] Hopp,T.P. and Woods,R.K.,1981 ‘Predictions of protein antigenic determinants from amino acid sequences’, Proc. Natl. Acad. Sci. USA., Vol 78 , pp 3824-3828.

[7] Janin,J. and Wodak,S.,1978 ‘Conformation of amino acid side-chains in proteins’, J.Mol.Biol., Vol 125, pp 357-86.

[8] Karplus,P.A. and Schulz,G.E.,1985 ‘Prediction of chain flexibility in proteins: a tool for the selection of peptide antigen’, Naturwissenschaften, Vol 72 , pp 212-213.

[9] Kawashima, S. and Kanehisa, M., 2000 'AAindex: amino acid index database', Nucleic Acids Res., Vol 28, pp 374.

[10] Kieber-Emmons, T., Luo, P., Qiu, J., Chang, T.Y., Insung, O., Blaszczyk-Thurin, M. and Steplewski Z., 1999 'Vaccination with carbohydrate peptide mimotopes promotes anti-tumor responses' N*at Biotechno*l. vol 17, pp 660-665.

[11] Kleter G.A., and Peijnenburg A.A., 2002 'Screening of transgenic proteins expressed in transgenic food crops for the presence of short amino acid sequences identical to potential, IgE–binding linear epitopes of allergens', BMC Structural Biology.

[12] Kolaskar, A.S. and Tongaonkar,P.C., 1990 'A semi-emperical method for prediction of antigenic determinants on protein antigens', FEBS, Vol 276 , pp172-174.

[13] Korber, B., Brander, C., Haynes, B., Koup, R., Kuiken, C., Moore, J., Walker, B. and Watkins, D. 2002., 'HIV Monoclonal Antibodies in HIV Molecular Immunology 2001'. Theoretical Biology and Biophysics group ,pp 278.

[14] Langeveld, J.P., martinez-Torrecuadrada,J., boshuizen,R.S., Meloen,R.H., and Ignacio,C.J., 2001 'Characterisation of a protective linear B cell epitope against feline parvoviruses', *Vaccine*, Vol 19, pp2352-2360.

[15] Meng,J., Dai,X., Chang,J.C., Lopareva,E., Pillot,J., Fields,H.A. and Khudyakov,Y.E., 2001 'Identification and characterization of the neutralization epitope(s) of the hepatitis E virus'. *Virology*., vol 288, pp 203-211.

[16] Odorico, M. and Pellequer, J.L. 2003 'BEPITOPE: predicting the location of continuous epitope and patterns in proteins', *J Mol Recognit*., Vol 16, pp 20-22.

[17] Parker,J.M.D., Guo,D. and Hodges,R.S., 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, Vol 25, pp 5425-5432.

[18] Pellequer,J.L., Westhof,E. and Regenmortel, M.H.V., 1991 'Predicting location of continuous epitopes in proteins from their primary structures', Methods in enzymology, Vol 203, pp 176-201.

[19] Pellequer,J-L., Westhof,E. and Regenmortel M.H.V., 1993 'Correlation between the location of antigenic sites and the prediction of turns in proteins', Immunol.Lett., Vol 36, pp 83-99.

[20] Pellequer.J.L and Wasthof., 1993 'PREDITOP: A program for antigenicity prediction', J. Mol. Graphics., Vol 11, pp 204-210.

[21] Ponnuswamy,P.K., Prabhakaran,M. and Manavalan,P., 1980 'Hydrophobic packing and spatial arrangements of amino acid residues in globular proteins', Biochim.Biophys.Acta., Vol 623, pp 301-316.

[22] Richard A. Goldsby, Thomas J.Kindt and Barbara A. Osborne., 1992 'Kuby Immunology', forth edition.

[23] Sabhanini, L., Manocha, M., Sridevi, K., Shashikiran, D., Rayanade, R., and Rao, D.N., 2003 'Developing subunit immunogens using B and T cell epitopes

and their constructs derived from F1 antigen of *Yersinia pestis* using novel delivery vehicles*', FEMS Immunol Med Microbio*l., vol 1579, pp 1-15.

[24] Saha,S., Bhasin,M. and Raghava,G.P.S., 2004 Bcipep: A database of B cell epitopes.

[25] Saha.S. and Raghava G.P.S., 2004 'BcePred: Prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties', In G.Nicosia, V.Cutello, P.J.Bentley and J.Timmis(Eds) ICARIS, LNCS 3239, pp.197-204.

[26] Sander, C. and Schneider, R.,1991 'Databases of homology-derived protein structures and the structural meaning of sequence alignment'. *Proteins* 9(1), pp 56-68.

[27] Selo,I., Clement,G., Bernard,H., Chatel,J., Creminon,C., Peltre,G., and Wal,J, 1999 'Allergy to bovine beta-lactoglobulin:specificity of human IgE to tryptic peptides', *Clin Exp Allerg*y., vol 29, pp 1055-1063.

[28] Ulbrandt,N.D., Cassatt,D.R., Patel,N.K., Roberts,W.C., Bachy,C.M., Fazenbaker,C.A. and Hanson,M.S. 2001 'Conformational nature of the Borrelia burgdorferi decorin binding protein A epitopes that elicit protective antibodies', *Infect Immu*n., vol 69, pp 4799-4807.

[29] Van Regenmortal et. al. 1993 'Anassessment of prediction methods for locating continuous epitope in proteins'. *Immunol Lett* , vol 17(2), pp 95-107.

[30] Van Regenmortel M.H,. 1993 'Synthetic peptides versus natural antigens in immunoassaya', *Ann Biol Clin*(Paris)., Vol 51, pp 39-41.

[31] Walter,G., 1986 'Production and use of antibodies against synthetic peptides', *J Immunol Methods,* Vol 88, pp 149-61.

[32] Westbrook,J., Feng,Z., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliland,G.L.,Bluhm,W.F., Weissig,H., Greer,D.S., Bourne,P.E. and Berman,H.M., 2002 'The Protein Data Bank: unifying the archive', *Nucleic Acids Res*., vol 30, pp 245-248.

[33] Wheller,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova, T.A., Wagner,L. and Rapp,B.A., 2002, 'Database resources of National Center for Biotechnology Information', *Nucleic Acids Re*s., vol 30, pp 13-16.

[34] Wu,H.C., Yeh,C.T., Huang,Y.L., Tarn,L.J. and Lung,C.C., 2001 'Characterization of neutralizing antibodies and identification of neutralizing epitope mimics on the Clostridium botulinum neurotoxin type A'. *Appl Environ Microbiol.*, vol 67, pp 3201-3207.

[35] Xiao,Y., Lu,Y. and Chen,Y.H., 2001 'Epitope-vaccine as a new strategy against HIV-1 mutation', *Immunol Let*t., vol 77, pp 3-6.