Journal of
Clinical & Cellular Immunology

Mukonyora, J Clin Cell Immunol 2015, 6:5
DOI: 10.4172/2155-9899.1000358

Review Open Access

A Review of Important Discontinuous B-Cell Epitope Prediction Tools

Michelle Mukonyora[1,2*]

[1]*Biotechnology Platform, Agricultural Research Council, Private Bag X05, Onderstepoort, 0110, South Africa*

[2]*Department of Life Sciences, College of Agriculture and Environmental Sciences, University of South Africa, Florida, 1710, South Africa*

[*]**Corresponding author:** Michelle Mukonyora, Biotechnology Platform, Agricultural Research Council, Private Bag X05, Onderstepoort, 0110, South Africa, Tel: +27125299121; E-mail: MukonyoraM@arc.agric.za

Received date: August 17, 2015; **Accepted date:** October 02, 2015; **Published date:** October 09, 2015

Abstract

The identification of B-cell epitopes is imperative for the rational design of vaccines, diagnostics and immunotherapeutics. Several bioinformatics resources are freely available for the prediction of B-cell epitopes, however despite advances in recent years, they still possess limited predictive capabilities. The aim of this review is to highlight and describe the algorithms of the most widely used free B-cell epitope prediction resources. The reasons behind the limited predictive powers of these algorithms are also discussed.

Keywords: B-cell epitopes; Epitope prediction; BEPro; Discotope; Ellipro; Epitopia; Seppa; Algorithms

Introduction

A B-cell epitope is a collection of distinct amino acid residues on an antigen that antibodies recognize and specifically bind to, thereby activating a protective immune response [1-3]. B-cell epitopes are classified according to their orientation in space as being either linear or discontinuous. Linear B-cell epitopes are composed of contiguous residues in the primary structure [1]. On the other hand, discontinuous B-cell epitopes comprise residues remotely located in the primary structure that are brought into close proximity due to the folding of the protein [1]. Only 10% of B-cell epitopes are linear and 90% are discontinuous [3]. Since linear B-cell epitopes do in fact adopt a conformation, the distinction between linear and discontinuous B-cell epitopes is a grey area [2,4].

The identification of B-cell epitopes is key to designing more effective vaccines [5,6]. Recombinant vaccines containing either a single or multiple B-cell epitopes from different serotypes can be rationally designed in a cost, time-effective and safe manner [7]. Also, protein subunits with the structural and immunogenic properties of their whole antigens may be designed and used as therapeutic and diagnostic tools [8-10].

Limitations of Experimental B-Cell Epitope Determination Methods

Experimental methods of elucidating B-cell epitopes include monoclonal antibody (MAb)-resistant variant studies, also known as virus neutralization tests [11], peptide scanning [8,12,13] and MAb-antigen contact studies [11]. Despite the successes of experimental B-cell epitope determination methods, they are laborious and not feasible when searching for epitopes on a large scale [1]. MAb-antigen contact studies, which is deemed the most reliable of the B-cell epitope determination strategies, is curbed by the limited availability of X-ray crystal structures of MAb-antigen complexes [14]. Computational B-cell epitope prediction methods have therefore been proposed as a cost and time-effective alternative to the laborious and resource-intensive classical experimental methods [15].

Computational B-Cell Epitope Prediction Methods

High-performance computers are able to execute algorithms of increasing complexity at decreasing costs and timespans. Consequently, computational methods reduce epitope prediction time by as much as 95% [16] and also have the potential to predict B-cell epitopes on a genome-wide scale [1].

Computational B-cell epitope prediction methods exploit the inherent physicochemical properties of B-cell epitopes in their algorithms [17]. B-cell epitopes tend to be more exposed to solvent than their surrounding surface-exposed residues and it is this high surface-exposure of antigenic regions that makes them highly flexible [18-20]. A high-flexibility is necessary in order to accommodate the conformational changes that take place upon B-cell epitopes binding with Abs [19]. Furthermore, it would be reasonable to assume that flexibility is a prerequisite of antigenic sites when one takes into consideration the plasticity of the complementarity determination regions (CDRs) of Abs.

Computational B-cell epitope prediction methods are broadly divided into sequence and structure-based methods as well as into linear and discontinuous epitope prediction methods. What all these methods essentially have in common is that they provide a way of correlating the physico-chemical properties of the respective amino acids to their probable location in the protein structure [1,21].

Prediction Tools for Linear B-Cell Epitopes

Propensity scale methods are the most common way by which linear B-cell epitopes are predicted and they are entirely dependent on the primary structure of the proteins [22,23]. The original propensity scale methods make use of hydrophilicity [24], secondary structure [25,26] and side-chain solvent accessibility [27] in their algorithms [23]. Modern linear epitope algorithms make use of a combination of propensity scale methods, but have been shown to only be marginally better at predicting linear epitopes [1,28]. In a similar manner to the experimental peptide scanning methods, propensity scales are not highly successful for discontinuous B-cell epitope prediction unless a given reading frame contains the amino acids that are the major determinants of the conformation of the B-cell epitope [29]. Alternatively, structure-based methods are more ideally suited for the prediction of discontinuous B-cell epitopes [22].

Prediction Tools for Discontinuous B-Cell Epitopes

Discontinuous B-cell epitope prediction methods employ various algorithms that mostly exploit a combination of structural and propensity scale-based information [22]. Some examples of discontinuous epitope prediction programmes that are a combination of structure and propensity scale-based methods are Discotope2.0, BEPro and SEPPA [4,30,31]. It has been shown that the most successful integrated methods consider amino acid composition, secondary structure and surface exposure in their algorithms [30]. There are however, some purely structure or propensity scale-based discontinuous prediction programmes that perform as well as integrated ones, namely Ellipro and Epitopia respectively [9,32].

Discontinuous B-cell epitope prediction methods require the 3-D structure of the antigen as input [33]. In cases where no structure is available, some of the programmes build homology models of the antigens and then proceed to predict B-cell epitopes from the models [32].

Devising a Discontinuous Epitope Prediction Method

Training dataset construction

The first step to devising any B-cell epitope prediction algorithm is the definition of a training dataset. Discontinuous B-cell epitope prediction methods use X-ray crystallographic information of MAb-antigen complexes to train their algorithms [22]. Redundancy is removed from the training datasets by generally allowing protein families to have equal representation [18]. To avoid over-fitting the algorithm, different parts of the dataset are used for training and evaluation [18].

B-cell epitope definition and benchmark dataset annotation

In order to train the prediction algorithms, a B-cell epitope needs to be defined [22] and the various prediction methods describe B-cell epitopes differently. In the *Discotope2.0* [30] dataset, B-cell epitopes are defined as those antigen amino acids that are a distance of at most 4Å from any of the Ab atoms [18]. In the *BEPro* training dataset, a B-cell epitope is any antigen residue that is no further than 6Å from the CDRs of the Ab chains, thereby excluding incidental contacts [4,34].

Surface exposure is another measure incorporated in the B-cell epitope prediction algorithms in order to aid in the definition of epitopes. In *Epitopia*, a surface amino acid is defined as any residue on a 3-D structure with a relative accessible surface area (relative ASA) greater than 0.05 [17]. For *SEPPA*, a residue was defined as surface exposed if it had at least 1Å^2 of ASA [31]. Furthermore, a surface exposed residue was a B-cell epitope if it lost at least 1Å^2 of ASA upon binding with its Ab [31]. In *Discotope2.0* the upper half-sphere neighbour count measure [35] was used as a measure of surface exposure (Figure 1) [30,35].

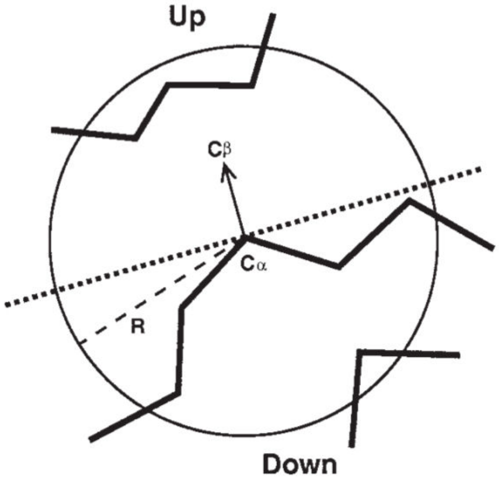

Figure 1: Half sphere exposure. The half sphere exposure (HSE) is a measure of the degree of surface exposure of an amino acid on a protein. It counts the number of C∝ neighbours within a half sphere of a given radius. The fewer the number of C∝ neighbours the more exposed an amino acid is. The dotted line represents the position of the plane that divides the sphere with a central C∝ atom and radius 'R'. The thick black lines represent part of the C∝ backbone of the protein. Adapted from Hamelryck [35].

An additional B-cell epitope definition that is part of *Ellipro*'s algorithm is that of the protrusion index [32]. The protrusion index provides a simplistic way of detecting those parts of the protein that protrude from the protein's surface. Residues with high protrusion index values are often associated with antigenic sites [20].

Discontinuous Epitope Prediction Machine-Learning Algorithms

Five discontinuous B-cell epitope prediction algorithms are discussed in this review, namely *Discotope* (versions 1.0 and 2.) [18,30], *BEPro* [4], *Ellipro* [32], *Epitopia* [9] and *Seppa* [31] (Table 1). These are among the most widely used and freely available discontinuous B-cell epitope prediction algorithms to date, as well as the ones suitable for the analysis of multimeric structures such as virus capsid proteins [36].

Discotope

Discotope (versions 1.0 and 2.0) integrates amino acids statistics expressed as log-odds ratios, spatial information and surface exposure in its algorithm [18,30]. It is notable in that it was the first B-cell epitope prediction method (as *Discotope1.0*) to make use of both propensity scale scores and structural information in its algorithm [18]. During execution of the *Discotope* algorithm, a 10Å radial sphere around each residue along the antigen chain is explored for intra-molecular contact residues (Figure 1). The total number of residues within the sphere is subtracted from the sum of propensity scores of those 'contact' residues' [30]. *Discotope1.0* is available as a standalone version, while *Discotope2.0* is available as an online server (Table 1).

Epitope Prediction Method	Online Server	Brief Notes (Features used in prediction method)
Discotope1.0 *Discotope2.0*	www.cbs.dtu.dk/services/DiscoTope/	Amino acid statistics, spatial context, surface accessibility
BEPro	http://pepito.proteomics.ics.uci.edu	Half sphere exposure values
Ellipro	http://tools.immuneepitope.org/ellipro	Protrusion index
Epitopia	http://epitopia.tau.ac.il	Based on Naïve Bayes classifier with physicochemical and structural geometrical properties
SEPPA	http://lifecenter.sgst.cn/seppa/	Unit patch of residue triangle

Table 1: A summary of the features of some widely used B-cell epitope prediction methods. Adapted from Sun et al. [33].

BEPro

BEPro, formerly known as *PEPITO*, was initially conceived as an alternative method to *Discotope*. *BEPro*, like *Discotope* combines propensity scales with surface exposure information, namely the upper half sphere neighbour count measure (Figure 1). *BEPro* utilizes the *Discotope* amino acid propensity scale [18] in its algorithm, side-chain orientation as well as solvent accessibility data [4]. *BEPro* is available online as a part of the SCRATCH suite of programmes [37] (Table 1).

Ellipro

Ellipro differs from the other B-cell epitope prediction methods in that it does not require training [32]. It is based on the notion that residues that protrude from the protein surface are more accessible for Ab binding [38] and that these protruding residues can be identified by treating the protein as an ellipsoid [39]. *Ellipro* uses Thornton's method [20] in combination with a residue-clustering algorithm to predict B-cell epitopes [32]. *Ellipro* is available as a standalone version and as an online server, which is part of the Immune Epitope Database Analysis Resource (Table 1).

Epitopia

Epitopia applies two machine-learning based algorithms for the prediction of B-cell epitopes from either the tertiary structure of the antigen or directly from its sequence [9]. A total of 44 physico-chemical and structural-geometrical properties for structure-based prediction and 41 properties for sequence-based prediction were used to train the *Epitopia* algorithm [17]. The immunogenic properties used to predict B-cell epitopes from sequences naturally do not include some of the structural-geometrical properties used for structure-based prediction [17]. These properties included previously used as well as novel amino acid propensity scales. Epitopia may be used via the online server or it may be downloaded as a standalone version (Table 1).

SEPPA

SEPPA employs the concept of the 'unit patch of residue triangle' to describe the local spatial context of a protein's surface amino acids [31]. The novel concept of 'unit patch of residue' is used by *SEPPA* to give an improved description of the local spatial context on the antigen surface. The unit patch of residue triangle is made up of any three surface residues whose respective side-lengths is less than 4Å [31]. Those unit patches containing at least two B-cell epitopes were defined as epitope unit patches, and those containing less than two B-cell epitopes were defined as non-epitope patches [31]. Epitope propensity scores are summed up for all unit patches within a 15Å radius of each residue in the antigen [40].

Limitations of Computational B-Cell Epitope Prediction Methods

Despite significant advances made in devising computational B-cell epitope prediction methods, there are still limitations to the predictive powers of their algorithms. There are therefore continued efforts to improve their performances. One of the most widely used performance evaluators for machine-learning algorithms is the area under the receiver operating characteristic curve, also known as (AUC) or (ROC) curve [41-43]. The true positive rate (TPR) is plotted on the y-axis and the false positive rate (FPR) is plotted on the x-axis, thereby illustrating how the TPR depends on the FPR [43] The TPR is also called sensitivity or recall [43]. AUC values range between zero and one [41]. A method that scores 0.5 is deemed a random discriminator and one that scores a value of one has a perfect predictive capability [22,43]. Currently, the top performing B-cell epitope prediction methods have average AUC values ranging between 0.6 and 0.7, depending on the evaluation dataset used [4,9,30,31].

Improper benchmark annotation limits the predictive ability of b-cell epitope prediction algorithms and performance evaluations

One of the major limitations to the improved performance of computational B-cell epitope prediction methods is improper benchmark annotation. Most B-cell epitope prediction methods allow for the annotation of only one epitope per antigen in their training datasets [4,9,18]. This not only excludes a large portion of known epitopes but it does not take into consideration the fact that not all B-cell epitopes on any particular antigen have been experimentally identified [4,17,30].

Another form of improper benchmark annotation is that most of the X-ray crystal structures in the training datasets consist of Abs bound to single antigen chains, yet Abs *in vivo* are raised against whole biological units [30]. A negative consequence of this, is that several antigen contacts that are predicted as being available for binding to an Ab are in fact involved in long-range intra-molecular interactions [30].

Improper benchmark annotation therefore not only has a direct influence on the predictive abilities of the algorithms but also on the performance measures of the methods [4,22,30]. A limitation of the AUC for B-cell epitope prediction methods is that it underestimates the predictive power of the algorithms as long as the training datasets are under-annotated [9,30]. Otherwise good predictors consequently call a number of false negatives [9,30,44].

Conclusion

In spite of the limited predictive powers of the respective B-cell epitope prediction methods, using a consensus of the results of the top performing methods can ameliorate these limitations [36,45]. When predicting putative novel B-cell epitopes, consensus results reduce the likelihood of false positive results and increase confidence in positive results [36].

If B-cell epitope prediction methods are to improve, there needs to be constant efforts to update the training datasets of algorithms with current epitope experimental data. The Immune Epitope Database 3.0 [46] is a valuable resource in this regard, as it currently has curated experimental data of 120,000 B- and T-cell epitopes. This is representative of at least 95% of the published epitopes as of the end of 2012 and this data is free and available to the public [47].

References

1. Larsen JE, Lund O, Nielsen M (2006) Improved method for predicting linear B-cell epitopes. Immunome Res 2: 2.
2. Peters B, Sidney J, Bourne P, Bui HH, Buus S, et al. (2005) The immune epitope database and analysis resource: From vision to blueprint. PLoS Biol 3: e91.
3. Zhao L, Wong L, Lu L, Hoi SC, Li J (2012) B-cell epitope prediction through a graph model. BMC Bioinformatics 13: S20.
4. Sweredoski MJ, Baldi P (2008) PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. Bioinformatics 24: 1459-1460.
5. Aggarwal N, Barnett PV (2002) Antigenic sites of foot-and-mouth disease virus (FMDV): an analysis of the specificities of anti-FMDV antibodies after vaccination of naturally susceptible host species. J Gen Virol 83: 775-782.
6. Mateu MG, Martinez MA, Capucci L, Andreu D, Giralt E, et al. (1990) A single amino acid substitution affects multiple overlapping epitopes in the major antigenic site of foot-and-mouth disease virus of serotype C. J Gen Virol 71: 629-637.
7. Bergmann-Leitner ES, Chaudhury S, Steers NJ, Sabato M, Delvecchio V, et al. (2013) Computational and experimental validation of B and T-cell epitopes of the in vivo immune response to a novel malarial antigen. PLoS One 8: e71610.
8. Meloen RH, Puyk WC, Meijer DJ, Lankhof H, Posthumus WP, et al. (1987) Antigenicity and immunogenicity of synthetic peptides of foot-and-mouth disease virus. J Gen Virol 68 : 305-314.
9. Rubinstein ND, Mayrose I, Martz E, Pupko T (2009) Epitopia: a web-server for predicting B-cell epitopes. BMC Bioinformatics 10: 287.
10. Usherwood EJ, Nash AA (1995) Lymphocyte recognition of picornaviruses. J Gen Virol 76 : 499-508.
11. Aktas S, Samuel AR (2000) Identification of antigenic epitopes on the foot and mouth disease virus isolate O1/Manisa/Turkey/69 using monoclonal antibodies. Rev Sci Tech 19: 744-753.
12. Bittle JL, Houghten RA, Alexander H, Shinnick TM, Sutcliffe JG, et al. (1982) Protection against foot-and-mouth disease by immunization with a chemically synthesized peptide predicted from the viral nucleotide sequence. Nature 298: 30-33.
13. Thomas AA, Woortmeijer RJ, Puijk W, Barteling SJ (1988) Antigenic sites on foot-and-mouth disease virus type A10. J Virol 62: 2782-2789.
14. Kuroda D, Shirai H, Jacobson MP, Nakamura H (2012) Computer-aided antibody design. Protein Eng Des Sel 25: 507-521.
15. Hunter P (2006) Into the fold. Advances in technology and algorithms facilitate great strides in protein structure prediction. EMBO Rep 7: 249-252.
16. De Groot AS, Bosma A, Chinai N, Frost J, Jesdale BM, et al. (2001) From genome to vaccine: in silico predictions, ex vivo verification. Vaccine 19: 4385-4395.
17. Rubinstein ND, Mayrose I, Pupko T (2009) A machine-learning approach for predicting B-cell epitopes. Mol Immunol 46: 840-847.
18. Haste Andersen P, Nielsen M, Lund O (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci 15: 2558-2567.
19. Novotný J, Handschumacher M, Haber E, Bruccoleri RE, Carlson WB, et al. (1986) Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). Proc Natl Acad Sci U S A 83: 226-230.
20. Thornton JM, Edwards MS, Taylor WR, Barlow DJ (1986) Location of 'continuous' antigenic determinants in the protruding regions of proteins. EMBO J 5: 409-413.
21. Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci U S A 78: 3824-3828.
22. Greenbaum JA, Andersen PH, Blythe M, Bui HH, Cachau RE, et al. (2007) Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. J Mol Recognit 20: 75-82.
23. Lo Y, Pai T, Wu W, Chang H (2013) Prediction of conformational epitopes with the use of a knowledge-based energy function and geometrically related neighboring residue characteristics. BMC Bioinformatics 14: S3.
24. Parker JM, Guo D, Hodges RS (2013) 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 19: 5425-5432.
25. Chou PY, Fasman GD (1974) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 13: 211-222.
26. Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120: 97-120.
27. Janin J, Wodak S (1978) Conformation of amino acid side-chains in proteins. J Mol Biol 125: 357-386.
28. Pellequer JL, Westhof E, Van Regenmortel MH (1991) Predicting location of continuous epitopes in proteins from their primary structures. Methods Enzymol 203: 176-201.
29. Lin SY, Cheng CW, Su EC (2013) Prediction of B-cell epitopes using evolutionary information and propensity scales. BMC Bioinformatics 14 Suppl 2: S10.
30. Kringelum JV, Lundegaard C, Lund O, Nielsen M (2012) Reliable B cell epitope predictions: impacts of method development and improved benchmarking. PLoS Comput Biol 8: e1002829.
31. Sun J, Wu D, Xu T, Wang X, Xu X, et al. (2009) SEPPA: a computational server for spatial epitope prediction of protein antigens. Nucleic Acids Res 37: W612-616.
32. Ponomarenko J, Bui HH, Li W, Fusseder N, Bourne PE, et al. (2008) ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinformatics 9: 514.
33. Sun P, Ju H, Liu Z, Ning Q, Zhang J, et al. (2013) Bioinformatics resources and tools for conformational B-cell epitope prediction. Comput Math Methods Med 2013: 943636.
34. Schlessinger A, Ofran Y, Yachdav G, Rost B (2006) Epitome: database of structure-inferred antigenic epitopes. Nucleic Acids Res 34: D777-780.
35. Hamelryck T (2005) An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins 59: 38-48.
36. Borley DW, Mahapatra M, Paton DJ, Esnouf RM, Stuart DI, et al. (2013) Evaluation and use of in-silico structure-based epitope prediction with foot-and-mouth disease virus. PLoS One 8: e61122.
37. Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 33: W72-76.
38. Atassi MZ (1984) Antigenic structures of proteins. Their determination has revealed important aspects of immune recognition and generated strategies for synthetic mimicking of protein binding sites. Eur J Biochem 145: 1-20.

39. Taylor WR, Thornton JM, Turnell WG (1983) An ellipsoidal approximation of protein shape. J Mol Graph 1: 30-38.

40. Bublil EM, Freund NT, Mayrose I, Penn O, Roitburd-Berman A, et al. (2007) Stepwise prediction of conformational discontinuous B-cell epitopes using the Mapitope algorithm. Proteins 68: 294-304.

41. Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 240: 1285-1293.

42. Spackman KA (1989) Signal detection theory: Valuable tools for evaluating inductive learning. Sixth International Workshop on Machine Learning. Morgan Kaufman Publishers Inc., San Mateo, California, p. 160-163.

43. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27: 861-874.

44. Ponomarenko JV, Bourne PE (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation. BMC Struct Biol 7: 64.

45. Liang S, Zheng D, Standley DM, Yao B, Zacharias M, et al. (2010) EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results. Biomed Cent Bioinforma 11: 381-387.

46. IEDB (2015) Immune Epitope Database 3.0 [Internet]. [cited 2015 Jan 26].

47. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, et al. (2015) The immune epitope database (IEDB) 3.0. Nucleic Acids Res 43: 1-8.