International Journal of Proteomics & Bioinformatics

Short Communication

T-Cell Epitope Prediction by Integration of Pattern Recognition and Motif Searching: Case Study on HLA-A*02, HLA-A*11 and HLA-A*24 - ∂

Yee Ying Lim, Cheh Tat Law and Yee Siew Choong*

Institute for Research in Molecular Medicine (INFORMM), Universiti Sains Malaysia, USM, Pulau Pinang, Malaysia.

***Address for Correspondence:** Yee Siew Choong, Institute for Research in Molecular Medicine (INFORMM), Universiti Sains Malaysia,11800, USM, Pulau Pinang, Malaysia, Tel: +604-653-4801; Fax: +604-653-4803; E-mail: yeesiew@usm.my

Submitted: 01 September 2017; **Approved:** 12 October 2017; **Published:** 13 October 2017

Cite this article: Lim YY, Law CT, Choong YS. T-Cell Epitope Prediction by Integration of Pattern Recognition and Motif Searching: Case Study on HLA-A*02, HLA-A*11 and HLA-A*24. Int J Proteom Bioinform. 2017;2(1): 027-030.

ABSTRACT

T-cell epitopes have a huge potential in the development of vaccine, disease prevention and diagnostics as well as therapeutics. As a numbers of epitopes have been identified experimentally in the past 20 years, databases focusing on different kinds of epitopes are now available. However, for vaccine development particularly, biochemical and immunological experiments are costly, time consuming, with low immunogenicity and reversible. Narrowing down the epitopes of interest could reduce the number of wet laboratory experiments and made vaccine designs cheaper and faster. In this works, we utilized support vector machine for pattern recognition in T-cell epitopes and further ranked the results by the present of anchor residues. We used the 9-mers peptide sequences obtained from Immune Epitope Database (IEDB) for HLA-A*02, HLA-A*11 and HLA-A*24 at that are related to Asian population for this works. The results showed that this two steps approach for T-cell epitope prediction is able to provide reasonable output. Therefore, the integration of pattern recognition and ranking by the present of anchor residue could be useful for future development of more alleles with various peptide lengths.

Keywords: T-Cell epitope prediction; Pattern recognition; Support Vector machine; Anchor residues

INTRODUCTION

Epitope-based vaccines make use of short, antigen-derived peptides (which corresponding to epitopes) that are administrated to trigger a protective humoral (B cell epitopes) and/or cellular (T cell epitopes) immune response. T cell epitopes are presented to T cells in association with Major Histocompatibility Complex (MHC) proteins. While cytotoxic T cell recognize intracellular peptides displayed by MHC class I molecules (CD8+ T cell epitopes), T helper cells recognize peptides that are taken up from the extracellular space and displayed by MHC class II molecules (CD4+ T cell epitopes). The peptide-MHC complex (pMHC) interacts with the T cell receptor, leading to its activation and subsequent induction of a cellular immune response. Epitope-based vaccines offer several potential benefits over traditional vaccines, including the precise control over the immune response activation, the ability to focus on the most relevant antigen regions, as well as production and biosafety advantages due to their chemically simple and well-characterized composition. The peptide T cell epitopes, specifically, can be used for the accurately monitoring the immune responses which activation by Major Histocompatibility Complex (MHC) and rationally designing vaccines [1-5]. Therefore, accurate prediction of T cell epitopes is crucial for this epitope-based vaccine development and clinical immunology.

As the molecular basis of immune recognition and the immune response, epitopes provide valuable information that is useful for disease prevention, diagnosis and therapeutic [6-8]. As a large number of epitopes have been identified since 1990s, various epitope databases (e.g. IEDB [9], Antigen [10,11], Bcipep [12], Epitome [13], MHCPEP[14], SYFPEITHI [15], MHCBN [16], FIMM [17] and EPIMHC [18]) are therefore being developed. However, the experimental identifications of epitopes from an antigen (e.g. phage display library, overlapping peptides, ELISA, immunofluorescence, immunohistochemistry, radioimmunoassay, Western blotting, X-ray crystallography and NMR studies on antibody-antigen structure, attenuation of the wild type pathogens by random mutations and serial passages, etc) are very expensive, time consuming, with low antigenicity and reversible. Therefore, predictive methods and software focusing on different types of epitopes have been witnessed (e.g. ABCpred [19], BCEPred [20], BepiPred [21], CED [22], Discotope [21], EMT [23]). With the aid of the epitope predictive software and databases, the list for the proteins of interest are now can be narrowed down, thus drastically reduce the number of laboratory experiments [24].

In this work, two steps of data processing for T-cell epitope prediction were proposed. Firstly, the dataset was grouped trained for pattern recognition by the length of the amino acids. Then, predicted results were further ranked by the present of anchor residues. We utilized HLA-A*02, HLA-A*11 and HLA-A*24 9-mers dataset that are more related to Asian population in this study. The integration of both pattern recognition and ranked the results by anchor residues showed that the approach is relatively reasonable and able to narrow down the predicted results. This approach sees the possibility to extend to other alleles and other length of peptide sequences.

METHODOLOGY

The methodology of this work is showed in figure 1. The T-cell epitopes (both positive and negative dataset) on HLA-A*02, HLA-A*11 and HLA-A*24 9-mers were first obtained from Immune Epitope Database (IEDB) [25]. The positive dataset includes all positive epitope sequences that have been verified by experimental means while negative dataset includes those epitope sequences that did not give immunological T-cell responses. Table 1 shows the total number of epitope sequences retrieved from IEDB that were used in this study. The training was performed on 80% of the dataset and the remaining 20% was used as testing dataset for each alleles. The training and testing was performed using Support Vector Machine (SVM) technique implemented in MATLAB R2015a. All dataset was first converted to ASCII format prior to training and testing by SVM Fine Gaussian module.

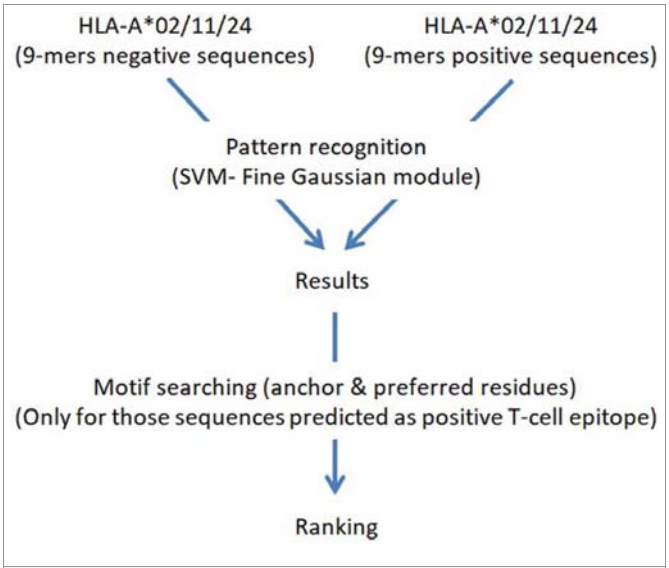

Figure 1: The methodology flowchart of the current work on T-cell epitope prediction module development using pattern recognition and motif searching approaches.

The module where then exported and implemented in a local host. We then employed a ranking system for the input of a query sequence. The query sequence will first be truncated to 9-mers window. The "positive" sequences will then be ranked according to the present of anchor residues. Table 2 shows the location and anchor residues in each allele [26-28]. The present of the anchor residue in the predicted positive sequence was given 3 marks, preferred residue gave 2 marks while others were given 1 mark. Therefore, a predicted positive sequence will have a maximum of 6 marks and minimum of 2 marks. The ranking of the predicted positive sequences was then arranged from the maximum to minimum marks.

RESULTS AND DISCUSSION

SVM has been widely used in solving a wide variety of biological problems. Current machine learning techniques involves Artificial Neural Network (ANN), Hidden Markov Model (HMM) and etc. However, the SVM algorithm which was developed recently has produced improve performances [29]. On the other hand, vaccine design could be benefit from T-cell epitope prediction specifically when the cost for experimental work is limited [30].

Results showed that the accuracy is 72.7%, 99.1% and 65.6% for HLA-A*02, HLA-A*11 and HLA-A*24, respectively. The differences in the accuracy percentage could be due to the number of sequences in each allele. The positive dataset for HLA-A*24 was nearly half of the negative dataset, the accuracy was therefore the lowest among the three alleles. The highest accuracy from HLA-A*11 might be due to the negative dataset is only less than 40% of the positive dataset, thus a higher prediction on the "positive" results occurred.

We also randomly picked a recently published experimental derived data [31] as the query sequence to the local host that we have integrated the pattern recognition and ranking system to see the accuracy of the output. We combined all the five sequences (FMGDIHQPL, LLSTAALPV, FLQLLLVTL, REANLSHYV and LEATYASTL) for CD8 specific epitopes tabulated by the authors where these sequences covered more than 80% of the world's population as vaccine candidate for human papillomavirus [31]. The results from our prediction showed that the five sequences were within the top predicted positive T-cell epitope (Figure 2). Therefore, we believed that the pattern recognition by SVM and results ranking by anchor residues could be a useful to predict the T-cell epitopes.

Table 1: The number of positive and negative epitope sequences for HLA-A*02, HLA-A*11 and HLA-A*24 obtained from IEDB T cell epitope database with the length of 9 amino acids.

Allele	Positive sequence	Negative sequence
*02	4379	4829
*11	1443	911
*24	1306	2129

Table 2: The anchor and preferred residues for 9-mers HLA-A*02, HLA-A*11 and HLA-A*24 [26-28].

Allele	Position # 2		Position # 9	
	Anchor	Preferred	Anchor	Preferred
*02	I, L, M, V	A, Q, T	I, V	A, L, M, T
*11	A, I, G, L, M, N, S, T, V	C, D, F	K	H, R, Y
*24	F, Y	I, L, M, T, V, W	F, I	L, M, W, Y

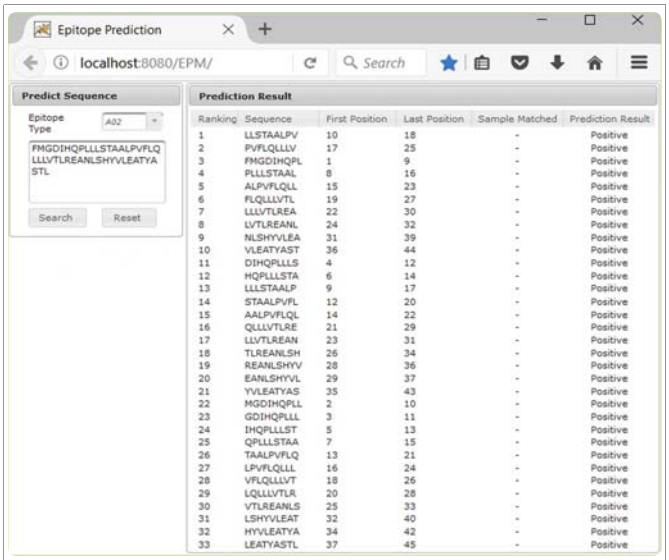

Figure 2: The predicted T-cell epitope results for the five sequences obtained from Vijayamahantesh, et al. [31] using pattern recognition and anchor residue ranking approach.

CONCLUSION

In this work, we integrated pattern recognition by SVM and further processed the predicted positive results by the present of anchor in the predicted sequences. The results are so far encouraging for at least the prediction for 9-mers HLA-A*02, HLA-A*11 and HLA-A*24. We see the possibility of applying similar approached for more alleles as well as with various length of peptide.

ACKNOWLEDGEMENTS

This work was supported by the Science Fund (305/CIPPM/613232) from Malaysian Ministry of Science, Technology and Innovations. Computational time was supported by Higher Institution Centre of Excellence (HICoE) Grant (311/CIPPM/44001005) from Malaysian Ministry of Education.

REFERENCES

1. Berzofsky JA. Structural features of protein antigenic sites recognized by helper T cells: what makes a site immunodominant? Year Immunol. 1986; 2: 28-38. https://goo.gl/z8r7QW
2. Berzofsky JA, Ahlers JD, Belyakov IM. Strategies for designing and optimizing new generation vaccines. Nat Rev Immunol. 2001; 1: 209-219. https://goo.gl/d1Faoe
3. De Groot AS, Scott DW. Immunogenicity of protein therapeutics. Trends Immunol. 2007; 28: 482-490. https://goo.gl/LQueMr
4. Sette A, Newman M, Livingston B, McKinney D, Sidney J, Ishioka G, et al. Optimizing vaccine design for cellular processing, MHC binding and TCR recognition. Tissue Antigens. 2002; 59: 443-451. https://goo.gl/DWg9HW
5. Sette A, Fikes J. Epitope-based vaccines: an update on epitope identification, vaccine design and delivery. Curr Opin Immunol. 2003; 15: 461-470. https://goo.gl/g7QG9K
6. Li S, Schmitz KR, Jeffrey PD, Wiltzius JJ, Kussie P, Ferguson KM. Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. Cancer Cell. 2005; 7: 301-311. https://goo.gl/FV5Zu2
7. Matsuo H, Kohno K, Niihara H, Morita E. Specific IgE determination to epitope peptides of omega-5 gliadin and high molecular weight glutenin subunit is a useful tool for diagnosis of wheat-dependent exercise-induced anaphylaxis. J Immunol. 2005; 175: 8116-8122. https://goo.gl/g7pHqk

8. Saxena AK, Singh K, Su HP, Klein MM, Stowers AW, Saul AJ, et al. The essential mosquito-stage P25 and P28 proteins from Plasmodium form tile-like triangular prisms. Nat Struct Mol Biol. 2006; 13: 90-91. https://goo.gl/G8URpz

9. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res. 2014; 43: 405-412. https://goo.gl/DyCLCq

10. Blythe MJ, Doytchinova IA, Flower DR. JenPep: a database of quantitative functional peptide data for immunology. Bioinformatics. 2002; 18: 434-439. https://goo.gl/jR7pcY

11. McSparron H, Blythe MJ, Zygouri C, Doytchinova IA, Flower DR. JenPep: a novel computational information resource for immunobiology and vaccinology. J Chem Inf Comput Sci. 2003; 43: 1276-1287. https://goo.gl/AxV6Qh

12. Saha S, Bhasin M, Raghava GP. Bcipep: a database of B-cell epitopes. BMC Genomics. 2005; 6: 79. https://goo.gl/AqC9UD

13. Schlessinger A, Ofran Y, Yachdav G, Rost B. Epitome: database of structure-inferred antigenic epitopes. Nucleic Acids Res. 2006; 34: 777-780. https://goo.gl/ZVQDLn

14. Brusic V, Rudy G, Harrison LC. MHCPEP: a database of MHC-binding peptides. Nucleic Acids Res. 1994; 22: 3663-3665. https://goo.gl/kQoNpD

15. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics. 1999; 50: 213-219. https://goo.gl/R9ZaJm

16. Bhasin M, Singh H, Raghava GP. MHCBN: a comprehensive database of MHC binding and non-binding peptides. Bioinformatics. 2003; 19: 665-656. https://goo.gl/Ba8Fw6

17. Schonbach C, Koh JL, Sheng X, Wong L, Brusic V. FIMM, a database of functional molecular immunology. Nucleic Acids Res. 2000; 28: 222-224. https://goo.gl/vUokXL

18. Reche PA, Zhang H, Glutting JP, Reinherz EL. EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. Bioinformatics. 2005; 21: 2140-2141. https://goo.gl/mgjuop

19. Saha S, Raghava GP. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins. 2006; 65: 40-48. https://goo.gl/JpwFeC

20. Chen J, Liu H, Yang J, Chou KC. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids. 2007; 33: 423-428. https://goo.gl/MX2AAw

21. Larsen JE, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. Immunome Res. 2006; 2: 2. https://goo.gl/UG9kdT

22. Huang J, Honda W. CED: a conformational epitope database. BMC Immunol. 2006; 7: 7. https://goo.gl/xYhsbR

23. Batori V, Friis EP, Nielsen H, Roggen EL. An in silico method using an epitope motif database for predicting the location of antigenic determinants on proteins in a structural context. J Mol Recognit. 2006; 19: 21-29. https://goo.gl/55GbAU

24. Yang X, Yu X. An introduction to epitope prediction methods and software. Rev Med Virol. 2009; 19: 77-96. https://goo.gl/NykVfy

25. Zhang Q, Wang P, Kim Y, Haste Andersen P, Beaver J, Bourne PE, et al. Immune epitope database analysis resource (IEDB-AR). Nucleic Acids Res. 2008; 36: 513-518. https://goo.gl/MEAJ2Z

26. Barouch D, Friede T, Stevanovic S, Tussey L, Smith K, Rowland Jones S, et al. HLA-A2 subtypes are functionally distinct in peptide binding and presentation. J Exp Med. 1995; 182: 1847-1856. https://goo.gl/gQnG1m

27. Falk K, Rotzschke O, Stevanovic S, Jung G, Rammensee HG. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature. 1991; 351: 290-296. https://goo.gl/Fc3sFj

28. Zhang QJ, Gavioli R, Klein G, Masucci MG. An HLA-A11-specific motif in nonamer peptides derived from viral and cellular proteins. Proc Natl Acad Sci U S A. 1993; 90: 2217-2221. https://goo.gl/dHxyMi

29. Pontil M, Verri A. Support vector machines for 3D object recognition. IEEE Trans Pattern Anal Mach Intell. 1998; 20: 637-646. https://goo.gl/q8Ttxf

30. De Groot AS. Immunomics: discovering new targets for vaccines and therapeutics. Drug Discov Today. 2006; 11: 203-209. https://goo.gl/YeyHR8

31. Vijayamahantesh, Amit A, Dikhit MR, Singh AK, Venkateshwaran T, Das VNR, et al. Immuno-informatics based approaches to identify CD8+ T cell epitopes within the Leishmania donovani 3-ectonucleotidase in cured visceral leishmaniasis subjects. Microbes Infect. 2017; 19: 358-369. https://goo.gl/a4NyHW