Conference paper

Li Cen Lim, Yee Ying Lim and Yee Siew Choong*

Data curation to improve the pattern recognition performance of B-cell epitope prediction by support vector machine

https://doi.org/10.1515/pac-2020-1107

Abstract: B-cell epitope will be recognized and attached to the surface of receptors in B-lymphocytes to trigger immune response, thus are the vital elements in the field of epitope-based vaccine design, antibody production and therapeutic development. However, the experimental approaches in mapping epitopes are time consuming and costly. Computational prediction could offer an unbiased preliminary selection to reduce the number of epitopes for experimental validation. The deposited B-cell epitopes in the databases are those with experimentally determined positive/negative peptides and some are ambiguous resulted from different experimental methods. Prior to the development of B-cell epitope prediction module, the available dataset need to be handled with care. In this work, we first pre-processed the B-cell epitope dataset prior to B-cell epitopes prediction based on pattern recognition using support vector machine (SVM). By using only the absolute epitopes and non-epitopes, the datasets were classified into five categories of pathogen and worked on the 6-mers peptide sequences. The pre-processing of the datasets have improved the B-cell epitope prediction performance up to 99.1 % accuracy and showed significant improvement in cross validation results. It could be useful when incorporated with physicochemical propensity ranking in the future for the development of B-cell epitope prediction module.

Keywords: B-cell epitopes; chemistry and its applications; data pre-processing; pattern recognition; support vector machine; VCCA-2020.

Introduction

Adaptive immune system is activated to concede and demolish invading pathogens in the vertebrate or higher order organisms after the innate immunity activation [1]. Adaptive immunity is segmented by B-cells which were responsible for the humoral immunity and the cell-mediated immunity, T-cells [1]. The epitope, also known as antigenic determinant, is the region of an antigen molecule that is recognized by an antibody [2]. B-cell epitopes are the regions that are recognized by B-cell receptor and bind onto the B-lymphocytes surface to produce antibodies. On the other hand, T-cell epitopes are short peptides within an antigen which presented on major histocompatibility complex (MHC) molecules that permit T-cell receptor to recognize and thus stimulate the T-cells production.

Article note: A collection of invited papers based on presentations at the Virtual Conference on Chemistry and its Applications (VCCA-2020) held on-line, 1–31 August 2020.

***Corresponding author: Yee Siew Choong**, Institute for Research in Molecular Medicine (INFORMM), Universiti Sains Malaysia, Minden, Penang, Malaysia, e-mail: yeesiew@usm.my
Li Cen Lim and Yee Ying Lim, Institute for Research in Molecular Medicine (INFORMM), Universiti Sains Malaysia, Minden, Penang, Malaysia

© 2021 IUPAC & De Gruyter. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. For more information, please visit: http://creativecommons.org/licenses/by-nc-nd/4.0/

B-cell epitope has been practiced extensively in the field of epitope-based vaccine design, antibody production and therapeutic development [3]. Several immunological experiments such as phage display library [4–6], overlapping peptides [7–9], ELISA [10, 11], Western blotting [12, 13], immunofluorescence [14, 15], X-ray crystallography [16, 17] and NMR [17] studies were used for identification of epitopes from an antigen. However, these experimental approaches are exorbitant and time consuming. Thus, in silico or computational bioinformatics tools make available for a more cost effective approach for B-cell epitopes prediction prior to experimental validation.

Most B-cell epitope prediction studies are focused on the sequence bases and correlated with physicochemical properties such as flexibility [18], hydrophobicity [19], solvent accessibility [20], antigenicity [21] and many others. BEPITOPE [22], PEOPLE [23] and BcePred [24] are some of the available B-cell epitope prediction programs that study a given query protein sequence with a sliding window and various propensity scale. Lately, the machine learning based approaches, such as hidden Markov model (HMM), artificial neural network (ANN), k-nearest neighbors (KNN) and support vector machine (SVM) were implemented to improve the prediction performance. For instance, BepiPred [25] used the two propensity scales of Parker's hydrophilicity scale and Levitt's secondary structure scale [26] combined with HMM. ABCPred with an accuracy of 65.93 % uses ANN to predict the B-cell epitopes using fixed length pattern [27]. Söllner and Mayer used the decision tree and KNN with the molecular operating environment [28]. BCPred [29], LEPS [30], LBtope [3] and SVMTriP [31] used SVM to predict the linear epitopes in different combination.

In addition, these predictive programs were developed using different datasets. BcePred [24] was developed using a dataset of 1029 B-cell epitopes and the same number of non-epitopes. ABCPred [27] was tested on a dataset of 700 B-cell epitopes and 700 non-epitopes. BCPred used 701 linear B-cell epitopes and 701 non-epitopes as the datasets. The above-mentioned programs obtained their unique experimentally validated linear B-cell epitopes from BCIPEP [32] and random peptides non-epitopes from SWISS-PROT [33]. Meanwhile, the LEPS approach evaluated on four datasets, included AntiJen [34], HIV [35], PC [30] and AHP [30]. SVMTriP [31] used linear B-cell epitopes from Immune Epitope Database (IEDB) with the final dataset construct of 4925 epitope and extracted the same number of non-epitopes in the corresponding antigen sequences. For LBtope [3], five different datasets which have retrieved experimentally validated B-cell epitopes and non-epitope from IEDB. But only LBtope_Fixed (12 063 epitopes and 20 589 non-epitopes with fixed length of 20-residues), LBtope_Variable (14 879 epitopes and 23 321 non-epitopes with variable length) and LBtope_Confirm (1042 epitopes and 1795 non-epitopes which reported in at least two studies) were used to implement three different models in the their server.

In this work, we focused on the datasets pre-processing prior to B-cell epitope prediction training. We obtained the datasets for five categories of pathogen (namely bacteria, fungi, multicellular, unicellular and viruses) from IEDB. We only trained the epitopes and non-epitopes datasets based on sequence pattern prediction method using SVM. Results showed that the pre-processing of the B-cell epitope dataset can improve the pattern recognition performance of the B-cell epitope prediction

Methodology

We first obtained the datasets of five pathogen categories (bacteria, fungi, multicellular, unicellular and viruses) from IEDB [36]. The datasets were then classified into three types: "epitope" (positive sequence), "non-epitope" (negative sequence) and "ambiguous" (Fig. 1). An "epitope" is referred to a peptide that showed as a positive epitope from experiment validated results while "non-epitope" is referred to a peptide that showed as a negative epitope from the experimental results. The "ambiguous" peptide is referred to a peptide that showed both positive and negative results from different experimental methods. Therefore, each category dataset was manually curated into these three different types. Only the experimentally validated epitopes and non-epitopes were used in this study and the ambiguous sequences were excluded in this study.

All the epitope and non-epitope datasets were then truncated into 6-mers peptide. The identical/repeated sequences were removed. All datasets were then converted into ASCII format. In this study, the prediction was

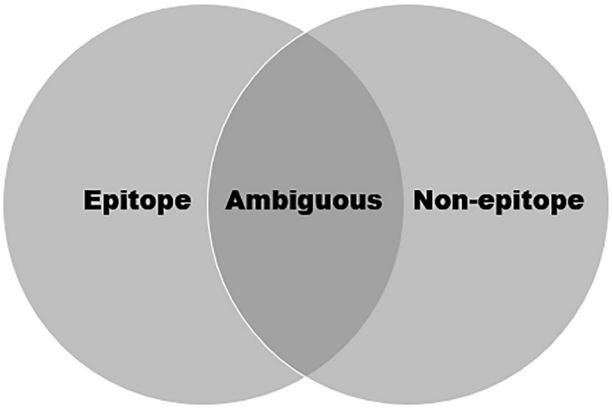

Fig. 1: The classifications of (IEDB) datasets: epitope (positive sequence from experiment results), non-epitope (negative sequence from experiment results) and ambiguous (both positive and negative from different experiment results).

made by the sequence-based patterns. SVM technique implemented by MATLAB 2018b was used to perform the training and testing with all the classifiers. A five-fold cross validation was used whereby the dataset was divided randomly into five equal subsets-four subsets were used for training and one subset was used for testing. The process was repeated five times in order to obtain the best subset combinations for the prediction.

Six indicators were used to measure the prediction performance. These indicators were the (1) accuracy – the proportion of correctly predicted peptides; (2) sensitivity – the likelihood that an epitope is correctly predicted as an epitope; (3) specificity – the likelihood that a non-epitope is correctly predicted as a non-epitope; (4) positive predictive value – the probability that subjects with a positive results are true epitopes in the dataset; (5) Matthews correlation coefficient (MCC); and (6) area under the ROC curve (AUC). The six indicators were calculated using the following equations:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100\,\% \tag{1}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\,\% \tag{2}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\,\% \tag{3}$$

$$\text{Positive predicitve value} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\,\% \tag{4}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{[(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})]}} \tag{5}$$

$$\text{AUC} = \frac{\text{Sensitivity}}{1 - \text{Specificity}} \tag{6}$$

where true positive (TP) represents the number of correctly predicted as epitope in the dataset is in fact an epitope; true negative (TN) represents the number of correctly predicted as non-epitope in the dataset is in fact a non-epitope; false positive (FP) represents the number of incorrectly predicted as non-epitope in the dataset is in fact an epitope; and false negative (FN) represents the number of incorrectly predicted as epitope in the dataset is in fact a non-epitope. The operating characteristics are illustrated in Table 1.

Results

In this work, only the “epitopes” (positive sequences) and “non-epitopes” (negative sequences) were used. In order to improve the data quality with for the best accuracy, the ambiguous and repetitive data

Table 1: The classification table of operating characteristics.

Actual class	Predicted class	
	Negative	Positive
Negative	True negative (TN)	False positive (FP)
Positive	False negative (FN)	True positive (TP)

were not included. Table 2 shows the total number of B-cell epitope sequences for five categories of pathogen retrieved from IEDB and the number of epitope, non-epitope and ambiguous for each category. The peptide sequences in both positive and negative datasets were truncated into length of six (6-mers) and the ambiguous data were again being removed. We tried with all SVM classifiers to ascertain the settings that can produce the best prediction results with the data. In addition, the variation in the advanced options, box constraint level as well as kernel scale were tested. Results showed that the Fine Gaussian classifier with Gaussian kernel 0.61 and 1 box constraint settings provided the best prediction results for all datasets.

Table 3 shows the five-fold cross validation results obtained by the Fine Gaussian classifier for five categories of pathogen. The accuracies are in the range from 71.3 to 99.1 %. The accuracy of performance could be due to the number of datasets in both epitopes and non-epitopes. Unicellular was the highest among the five categories of pathogen in term of accuracy, specificity and positive predictive value. These might due to the dataset contains nearly 99 % of the non-epitope sequences and only 1 % of the dataset was the epitope sequences, hence higher specificity but lower sensitivity. However, multicellular obtained the highest MCC and AUC scores instead of unicellular. Both MCC and AUC metrics are sensitive to the imbalance data since they take account in all four classes (accuracy, sensitivity, specificity and positive predictive value). AUC score near one is considered a good prediction to distinguish the positive and negative where AUC score of 0.5 showed a prediction performing poorly in prediction that are almost random. All five categories of pathogen have the AUC score between 0.77 and 0.90 indicating the predictions are able to perform well in the prediction. The lowest accuracy obtained from viruses compared with others might be due to the epitopes and non-epitopes dataset are having almost 50 % repetitive data.

We then compared the performance of our prediction with other existing B-cell epitope prediction programs. Table 4 shows the results produced using only the epitopes and non-epitopes datasets did improve the prediction performance in term of accuracy, positive predictive value, MCC and AUC. The lowest accuracy for available prediction approaches was 58.48 % by LBtope and the results in this work has the lowest accuracy of 71.30 %. Our results also has the positive predictive values and MCC of 71.11 % and 0.402 compared with 32.07 % and 0.104 from LEPS. We did not compare the sensitivity and specificity with the available prediction programs as these results depend on the positive and negative datasets used in the training.

Table 2: The total number of peptide sequences retrieved from IEDB and the attribution of epitope, non-epitope and ambiguous for the five categories of pathogen.

Pathogen	No. of peptides	Epitope	Non-epitope	Ambiguous
Bacteria	40 949	4626	15 415	2102
Fungi	1346	258	278	77
Multicellular	3495	408	406	201
Unicellular	184 055	3666	174 474	544
Viruses	72 070	9446	12 515	4365

Table 3: The results obtained by the Fine Gaussian SVM classifier, kernel 0.61 and 1 box constraint for the five categories of pathogen.

Pathogen	Accuracy (%)	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	MCC	AUC
Bacteria	81.4	45.37	94.82	76.49	0.49	0.84
Fungi	81.7	**96.19**	59.71	78.40	0.62	0.88
Multicellular	81.5	89.55	71.63	79.41	**0.63**	**0.90**
Unicellular	**99.1**	17.18	**>99.0**	**95.25**	0.40	0.77
Viruses	71.3	62.45	78.96	71.11	0.42	0.78

MCC: Matthews correlation coefficient, AUC: Area under the ROC curve. Bold fonts indicate the best results in each category.

Discussion

Prior to B-cell epitope prediction module development, the B-cell epitope dataset obtained from IEDB were first pre-processed whereby only epitopes and non-epitopes were used. In addition, the data was also categorized into five pathogen types. Five-fold cross validation training and testing were then performed for the pattern recognition by SVM using 6-mers sequence length as according to Wang & Pai, the range of 6–30 residues are posing in almost 95 % of verified linear B-cell epitopes [37].

There were vast difference in the number of datasets between epitopes and non-epitopes for each category of pathogen. Therefore, indicating the MCC and AUC values could provide better information compared to other measurement, *e.g.* percentage of accuracy, in the performance evaluation of the developed module [38]. Besides, inaccurate values might be avoided by the measurements such as accuracy, sensitivity and sensitivity, are solely depending on the number of sequences for both epitopes and non-epitopes. For example, the dataset of viruses showed highest results in term of accuracy and specificity but not in MCC and AUC. Furthermore, the MCC and AUC values showed more consistent results compared to other available programs (Table 4).

The results in this work showed that the pattern recognition after data pre-processing has considerable improvement compared with other available B-cell epitope prediction programs (Table 4). However, these results were the indirect comparison since most of the programs were developed based on physicochemical properties. The improvement might be due to the data preparation and machine learning approach that were using in this study. The existing prediction methods were developed using different datasets and machine learning approaches. Most of them used smaller dataset with the experimentally verified epitopes from different databases (*e.g.*: BCIPEP, IEDB) and non-epitopes were randomly chosen from SWISS-PROT. However, in this study we retrieved both epitope and non-epitope datasets from the IEDB database. In addition, the

Table 4: The comparison between the existing B-cell epitope prediction approaches with the prediction results in this work.

Server	Accuracy (%)	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	MCC	AUC
ABCPred	65.93	67.14	67.71	65.61	0.32	NA
BcePred	58.70	56.00	61.00	NA	NA	NA
LEPS	72.52	26.97	84.22	32.07	0.10	NA
SVMTriP	NA	80.10	NA	55.20	NA	0.70
BcPred	67.90	72.61	63.20	NA	0.36	0.76
LBtope	58.48	67.92	53.10	NA	0.20	0.65
Bacteria	81.4	45.37	94.82	76.49	0.49	0.84
Fungi	81.7	**96.19**	59.71	78.40	0.62	0.88
Multicellular	81.5	89.55	71.63	79.41	**0.63**	**0.90**
Unicellular	**99.1**	17.18	**>99.0**	**95.25**	0.40	0.77
Viruses	71.3	62.45	78.96	71.11	0.18	0.78

MCC: Matthews correlation coefficient, AUC: Area under the ROC curve.

obtained datasets were further categorized into five types of pathogen to increase the accuracy and reliability of the prediction. Besides, the latest and vast experimental data were used in the prediction could also contributed to a more rational results.

Conclusions

SVM is commonly used in other available programs and it was also reported to be well performing in classification. In this work, Fine Gaussian classifier has the highest prediction speed for binary classification with medium memory usage and it also has high flexibility as written in the MATLAB documentations. A flexible method might be generalized well across the different training sets and perform better as complexity increases. In future, the physicochemical propensity scale prediction can be incorporated to make the prediction results more sensible prior to the development of a web-based B-cell epitope prediction server for open usage. In this study, we implemented the pattern sequence-based prediction by SVM on the fixed length epitopes (6-mers) and compared to the existing approaches. The pre-processing and categorizing the dataset showed an improvement in the prediction. In future, physicochemical properties may be included as additional prediction method as well as for results ranking and a web-based prediction server can be developed to benefit other researchers. Besides, the used datasets should also be constantly updated to improve accuracy of the performance.

Research funding: This work is supported by Fundamental Research Grant Scheme (FRGS/1/2018/STG05/USM/02/1; 203/CIPPM/6711680) from Malaysia Ministry of Higher Education.

References

[1] J. L. Sanchez-Trincado, M. Gomez-Perosanz, P. A. Reche. *J. Immunol. Res.* **2017**, 1 (2017).
[2] L. Potocnakova, M. Bhide, L. B. Pulzova. *J. Immunol. Res.* **2016**, 6760830 (2016).
[3] H. Singh, H. R. Ansari, G. P. S. Raghava. *PloS One* **8**, e62216 (2013).
[4] J. Liu, Q. Ma, F. Yang, R. Zhu, J. Gu, C. Sun, X. Feng, C. Du, P. R. Langford, W. Han, J. Yang, L. Lei. *Vet. Microbiol.* **205**, 14 (2017).
[5] J. Zhao, E. C. Sun, N. H. Liu, T. Yang, Q. Y. Xu, Y. L. Qin, Y. H. Yang, D. L. Wu. *Vet. Immunol. Immunopathol.* **148**, 364 (2012).
[6] E. C. Sun, J. Zhao, T. Yang, N. H. Liu, H. W. Geng, Y. L. Qin, L. F. Wang, Z. G. Bu, Y. H. Yang, R. A. Lunt, D. L. Wu. *Virol. J.* **8**, 100 (2011).
[7] M. W. Heuzenroeder, M. D. Barton, T. Vanniasinkam, T. Phumoonna. *Methods Mol. Biol.* **524**, 137 (2009).
[8] S. Gonzalez, L. Vina, C. Nazabal, G. Chinea, E. Caballero, A. Musacchio. *Biotechnol. Appl. Biochem.* **32**, 1 (2000).
[9] T. Vanniasinkam, M. D. Barton, T. P. Das, M. W. Heuzenroeder. *Methods Mol. Biol.* **1785**, 121 (2018).
[10] J. Ti, Z. Li, X. Li, Y. Lu, Y. Diao, F. Li. *PloS One* **12**, e0181177 (2017).
[11] J. M. Yang, H. J. Wang, L. Du, X. M. Han, Z. Y. Ye, Y. Fang, H. Q. Tao, Z. S. Zhao, Y. L. Zhou. *Cancer Immunol. Immunother.* **58**, 1387 (2009).
[12] A. Moming, D. Tuoken, X. Yue, W. Xu, R. Guo, D. Liu, Y. Li, Z. Hu, F. Deng, Y. Zhang, S. Sun. *PloS One* **13**, e0204264 (2018).
[13] X. Liu, Y. Li, Z. Li, X. Wei, Y. Ma, P. Cheng, R. Jiao, J. Fang, Y. Xing, J. Tang, M. Wang, T. Li. *Int. J. Biol. Macromol.* **112**, 537 (2018).
[14] T. Lagousi, J. Routsias, C. Piperi, A. Tsakris, G. Chrousos, M. Theodoridou, V. Spoulou. *J. Biol. Chem.* **290**, 27500 (2015).
[15] J. Ma, Y. Wei, L. Zhang, X. Wang, D. Yao, D. Liu, W. Liu, S. Yu, Y. Yu, Z. Wu, L. Yu, Z. Zhu, Y. Cui. *J. Med. Microbiol.* **67**, 423 (2018).
[16] G. Obmolova, A. Teplyakov, T. J. Malia, N. Wunderler, D. Kwok, L. Barone, R. Sweet, T. Ort, M. Scully, G. L. Gilliland. *Mol. Immunol.* **83**, 92 (2017).
[17] W. Ding, X. Huang, X. Yang, J. J. Dunn, B. J. Luft, S. Koide, C. L. Lawson. *J. Mol. Biol.* **302**, 1153 (2000).
[18] P. A. Karplus, G. E. Schulz. *Naturwissenschaften* **72**, 212 (1985).
[19] J. M. R. Parker, D. Guo, R. S. Hodges. *Biochemistry* **25**, 5425 (1986).
[20] E. A. Emini, J. V. Hughes, D. S. Perlow, J. Boger. *J. Virol.* **55**, 836 (1985).
[21] A. S. Kolaskar, P. C. Tongaonkar. *FEBS Lett.* **276**, 172 (1990).
[22] M. Odorico, J.-L. Pellequer. *J. Mol. Recogn.* **16**, 20 (2003).
[23] A. J. P. Alix. *Vaccine* **18**, 311 (1999).

[24] S. Saha, G. P. S. Raghava. BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties, in *Lecture Notes in Computer Science*, Nicosia G., Cutello V., Bentley P. J., Timmis J. (Eds.), p. 197, Springer, Berlin, Heidelberg (2004).
[25] J. E. Larsen, O. Lund, M. Nielsen. *Immunome Res.* **2**, 2 (2006).
[26] M. Levitt. *Biochemistry* **17**, 4277 (1978).
[27] S. Saha, G. P. S. Raghava. *Proteins: Struct. Funct. Bioinf.* **65**, 40 (2006).
[28] J. Söllner, B. Mayer. *J. Mol. Recogn.* **19**, 200 (2006).
[29] Y. EL-Manzalawy, D. Dobbs, V. Honavar. *J. Mol. Recogn.* **21**, 243 (2008).
[30] H.-W. Wang, Y.-C. Lin, T.-W. Pai, H.-T. Chang. *J. Biomed. Biotechnol.* **2011**, 1 (2011).
[31] B. Yao, L. Zhang, S. Liang, C. Zhang. *PloS One* **7**, e45152 (2012).
[32] S. Saha, M. Bhasin, G. P. S. Raghava. *BMC Genom.* **6**, 79 (2005).
[33] A. Bairoch, R. Apweiler. *Nucleic Acids Res.* **28**, 45 (2000).
[34] C. P. Toseland, D. J. Clayton, H. McSparron, S. L. Hemsley, M. J. Blythe, K. Paine, I. A. Doytchinova, P. Guan, C. K. Hattotuwagama, D. R. Flower. *Immunome Res.* **1**, 4 (2005).
[35] M. K. Gorny, S. Zolla-Pazner, Human monoclonal antibodies that neutralize HIV-1, in *HIV Immunology and HIV/SIV Vaccine Databases 2003*, B. T. M. Korber, C. Brander, B. F. Haynes, R. Koup, J. P. Moore, B. D. Walker, D. I. Watkins (Eds.), p. 37, Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos (2003).
[36] R. Vita, S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, B. Peters. *Nucleic Acids Res.* **47**, D339 (2019).
[37] H.-W. Wang, T.-W. Pai. *Immunoinformatics* **1184**, 217 (2014).
[38] D. Chicco. *BioData Min.* **10**, 35 (2017).