A hybrid method of propensity scales and support vector machine in a linear epitope prediction

Hsin-Wei Wang, Ya-Chi Lin, Tun-Wen Pai*
Department of Computer Science and Engineering,
National Taiwan Ocean University,
Keelung, Taiwan, R.O.C.
e-mail: twp@mail.ntou.edu

Pei-Wen Tsai
Institute of Molecular and Cell Biology,
National Tsing Hua University
Hsinchu, Taiwan R.O.C.

Hao-Teng Chang*
Graduate Institute of Molecular Systems Biomedicine,
China Medical University,
Taichung, Taiwan, R.O.C.
e-mail:htchang@mail.cmu.edu.tw

***Abstract*—An epitope activates B cells to amplify and induce antibodies which can neutralize the foreign molecules, particles and pathogens. It also plays a crucial role in developing synthetic peptides for vaccination. Identification of epitopes using biological screening approaches is time consuming and high cost. Therefore, bioinformatics approaches are developed to enhance the speed of identifying the epitopes and conserve time. Herein, a combinatorial methodology based on physico-chemical properties and SVM (Support Vector Machine) techniques was proposed to address the aim of this study. Datasets of epitope and non epitope segments with 2, 3 and 4 residues in length were trained and applied as statistical features of SVM. After training, three datasets including one curated and two public ones were employed to evaluate the performance of the proposed system which was also compared with four existing LE predictors, BepiPred, ABCpred, BCPred and FBCPred. Our proposed system has presented better specificity, accuracy, and positive prediction value (PPV) in most testing cases. High specificity and PPV of a linear epitope prediction can lead to an efficient and effective design on biological experiments.**

Keywords: linear epitope; support vector machine; physico-chemical property;antibody-antigen; amino acid segment

I. INTRODUCTION

Antigenic epitopes on protein surface can elicit immune response with specific antibodies. Prediction of B-cell linear epitopes provides pre-analysis for biologists prior to their immunobiological experiments and vaccine design. Hence, an accurate epitope prediction provides an important role in prevention and control of diseases. There are two major types of epitopes: linear epitope (LE) and conformational epitope (CE) [1]. Refer to Figure 1, circles with slash lines in the left side of Figure 1 represent a LE composed of contiguous stretches of amino acid residues, while the circles with slash lines in the right side of Figure 1 are illustrated as a CE which is composed of non-contiguous segments constructed by folding of the polypeptide chain. Although it has been estimated that approximately more than 90% of B-cell epitopes belong to discontinuous types [2], most of antigen structures are not yet resolved and verified by structural biologists. Due to insufficient spatial information of CEs and strong dependency upon LEs, it is still important to predict LEs for fundamental immunology and general applications. In this study, we have emphasized the prediction methodology based on the combination of propensity scale methods with SVM techniques to improve the performance of LE prediction.

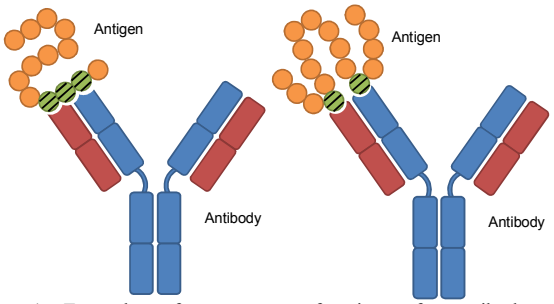

Figure 1. Examples of two types of epitope for antibody-antigen combination. Left side represents a Linear Epitope (LE) composed by a continuous segment and right side displays a Conformational Epitope (CE) with discontinuous stretches.

Most of published epitope prediction methods were based on physico-chemical properties of amino acids, such as flexibility [3], surface accessibility [4], hydrophilicity [5], secondary structure [6], antigenicity [7], and *etc.*. However, in 2005, Blythe and Flower [8] employed 484 amino acid propensity scales to thoroughly investigate the relationship between LE and global peaks of propensity profiles, and the results showed that the performance of the best set of scales was only slightly better than a random model. Accordingly, a study based on analyzing local peaks of propensity profiles as LE candidates was proposed to enhance the prediction performance [9]. Furthermore, several studies

978-0-7695-4373-4/11 $26.00 © 2011 IEEE
DOI 10.1109/CISIS.2011.89

attempted to improve the accuracy of LE prediction by featuring the machine learning approaches. For example, BepiPred [10] proposed the combination of Hidden Markov model (HMM) with Parker hydrophilicity scale [5] to predict LEs, and the performance was improved within an obvious scale.

According to Chen's study [11], it showed that the occurrence frequencies of some amino acid pairs (AAPs) in epitope datasets are significantly higher than in non-epitope datasets, or vice versa. This statistical feature is worth of noticing and can be applied to enhance the performance of a LE prediction system. Hence, both the advantages of featuring statistical distribution of verified epitopes and preserving the antigenic characteristics of candidate peptides are considered simultaneously in this study. Here, amino acid segments with 2 ($AASs^2$), 3 ($AASs^3$) and 4 ($AASs^4$) residues in length of both epitopes and non-epitopes were initially and exclusively collected and considered as the statistical features for LIBSVM (A Library for Support Vector Machines) [12]. A system combining AASs, SVM and a physico-chemical based LE prediction system, LEPD [9], was designed to improve the overall performance of LE prediction.

II. IMPLEMENTATION

A. System architecture

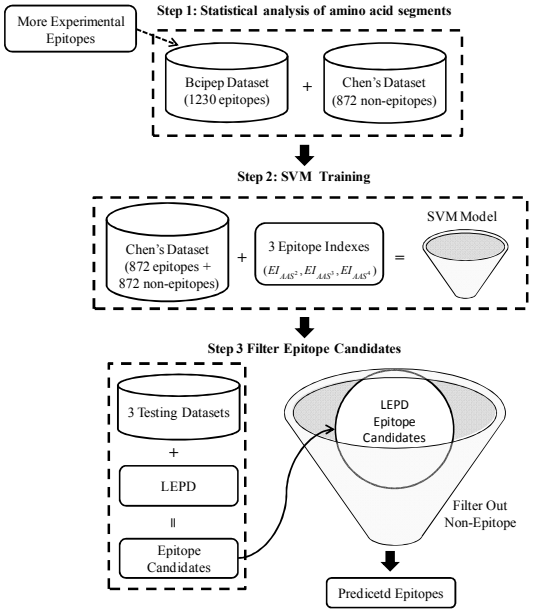

Figure 2. System flow chart of the proposed method.

The proposed system is divided into three main steps and shown in Figure 2. The first step evaluates the statistical characteristics of occurrence frequencies of AASs with various lengths from the independently collected epitope and non-epitope datasets respectively. The second step performs the self-training of an SVM classifier on the 1744 (872+872) verified segments to establish a determination model for epitope candidates. For the third step, the proposed SVM classifier filters epitope candidates directly from the predicted results of LEPD system. The following sections describe the details of each step.

B. Datasets

The trained B-cell epitopes were taken from Bcipep database [13] which comprises experimentally verified B-cell epitopes. The Bcipep database contains 1230 non-redundant LEs whose lengths ranging from 3 to 56 residues within more than 1000 antigen proteins. This dataset was applied for analyzing statistical properties of epitopes with respect to the occurrence frequencies of AASs with lengths from 2 to 4 amino acids.

More specifically, 872 epitopes and 872 non-epitopes from Chen *et al.* [11] were employed as the training data in this study. Lengths of all selected segments within the dataset were restricted to a uniform length of 20 amino acids. Here, the original source of the 872 epitopes were also adopted from the Bcipep database. Due to the training processes from any learning machine require fixed length training data, and the facts of B-cell epitopes from Bcipep database are various in lengths. Therefore, Chen *et al.* applied ‘‘truncation-extension treatment’’, that is, if the length of a B-cell epitope is longer than 20 residues, then truncates superfluous residues from both N- and C-terminals equally to retain the central 20 amino acids only; if the length of a B-cell epitope is shorter than 20, then extends adjacent residues to both N- and C- terminals equally until it reaches 20 residues long. After truncation-extension treatment and removing duplicated segments, Chen *et al.* acquired 872 unique B-cell epitopes with an identical length of 20 residues. On the other hand, the 872 non-epitopes were created by random selection from the Swiss-Prot database [14]. The set of generated non-epitope segments were all with 20 residues in length and contained different segments comparing to verified epitopes. These 872 non-epitopes are used for analyzing AASs statistics of non-epitopes.

C. Statistical analysis of amino acid segments

The proposed filtering mechanism is adopted from Chen’s idea, AASs with 2 residues ($AASs^2$), as statistical features for LE verification. Furthermore, the proposed method also extended the 2-residue segments to 3 and 4 residues. In the following sections, we describe the statistical analysis of three proposed features. Table 1 defines the required 7 variables in statistical analysis for query AASs.

Since there are 20 different amino acids, a total of 400 possible combinations of residue pairs are analyzed for their occurence frequencies within the epitope and non-epitope datasets. A proportion distribution diagram of $AASs^2$ is shown in Figure 3. Each point in Figure 3 represents an AAS^2, and the X-axis represents the occurrence frequencies of AAS^2 from the training dataset (X_{AAS^2}), and the Y-axis

represents the occurrence times from epitope dataset of the query AAS^2 divided by the total occurrence times of the AAS^2 from epitope and non-epitope datasets (Y_{AAS^2}). The calculations of feature values for X_{AAS^2} and Y_{AAS^2} are shown in Equation (1) and (2).

Table 1. Required variables in statistical analysis for amino acid segments.

Variables	Description
$N^+_{AAS^l}$	The occurrence times of an *l*-residue amino acid segment in the epitope dataset.
$N^-_{AAS^l}$	The occurrence times of an *l*-residue amino acid segment in the non-epitope dataset.
$f^+_{AAS^l}$	The occurrence frequencies of an *l*-residue amino acid segment in the epitope dataset.
$f^-_{AAS^l}$	The occurrence frequencies of an *l*-residue amino acid segment in the non-epitope dataset.
$Total^+_{AASs^l}$	The total occurrence times of all *l*-residue amino acid segments in the epitope dataset.
$Total^-_{AASs^l}$	The total occurrence times of all *l*-residue amino acid segments in the non-epitope dataset.
EI_{AAS^l}	Epitope Index of an *l*-residue amino acid segment.

$$X_{AAS^2} = \frac{N^+_{AAS^2} + N^-_{AAS^2}}{Total^+_{AASs^2} + Total^-_{AASs^2}} \quad (1)$$

$$Y_{AAS^2} = \frac{N^+_{AAS^2}}{N^+_{AAS^2} + N^-_{AAS^2}} \quad (2)$$

In Figure 3, important information of statistical characteristics of $AASs^2$ associated with epitopes is provided. First, if X_{AAS^2} of a specified AAS^2 is large, it represents that AAS^2 occurred with relatively high frequencies in the training dataset. Secondly, if Y_{AAS^2} is close to 0.5, it reflects that the AAS^2 possesses low identifiable ability during predicting epitopes. According to the statistics, we found that there are about 38.25% of $AASs^2$ possessing values of Y_{AAS^2} more than or equal to 0.6 or less than or equal to 0.4. Generally speaking, if a particular AAS^2 is frequently appeared in protein sequences (large X_{AAS^2}) and possesses a high proportional rate within the epitope dataset (large Y_{AAS^2}) which means the AAS^2 possessing high probability to be an epitope candidate. Similarly, if an AAS^2 is with relatively low appearance rate in protein sequences (small X_{AAS^2}) and the AAS^2 is mostly found in the epitope dataset (large Y_{AAS^2}), this AAS^2 can also be considered as a good candidate for LE prediction. In order to explore these characteristics, the possibility of an AAS^2 being considered as a partial epitope segment is represented by its Epitope Index through the following calculation.

The Epitope Index (EI_{AAS^2}) of an AAS^2 is first obtained by Equation (3)-(5), which calculates the frequency of occurrence of a particular AAS^2 in the epitope dataset divided by the frequency of occurrence of the same AAS^2 in the non-epitope dataset, and then takes log values. The next step normalizes the EI_{AAS^2} as in the Equation (6). The value determined from the previous step is normalized into the range of [0, 1].

$$f^+_{AAS^l} = \frac{N^+_{AAS^l}}{Total^+_{AASs^l}} \quad (3)$$

$$f^-_{AAS^l} = \frac{N^-_{AAS^l}}{Total^-_{AASs^l}} \quad (4)$$

$$EI_{AAS^2} = \log\left(\frac{f^+_{AAS^2}}{f^-_{AAS^2}}\right) \quad (5)$$

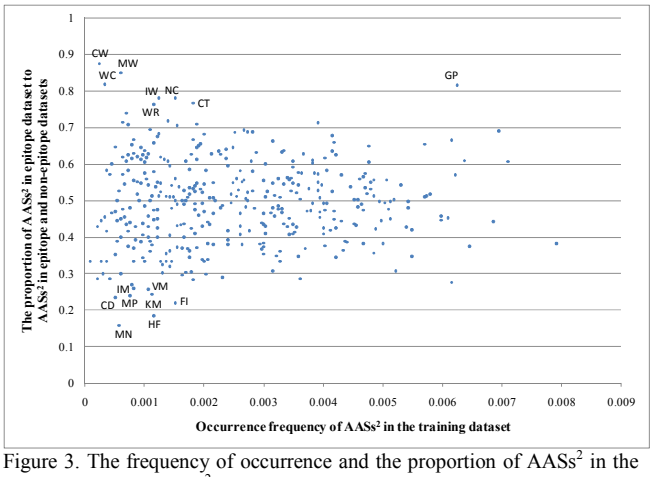

Figure 3. The frequency of occurrence and the proportion of $AASs^2$ in the epitope dataset to $AASs^2$ in the epitope and non-epitope datasets. The noted $AASs^2$ of CW, MW, WC, GP, IW, NC, CT, WR locating in upper part possessing higher possibility to be considered as epitope pairs, while $AASs^2$ of IM, VM, KM, MP, CD, FI, HF, MN in lower part for potential non-epitope pairs.

$$normalize(EI_{AAS^2}) = \frac{EI_{AAS^2} - \min(EI_{AASs^2})}{\max(EI_{AASs^2}) - \min(EI_{AASs^2})} \quad (6)$$

The functions $\max(EI_{AASs^2})$ and $\min(EI_{AASs^2})$ represent the maximum and minimum values of $AASs^2$ before normalization. The normalization is designed to avoid the dominance of an individual feature in the classifier learning processes. Since the length of all training epitopes and non-epitopes are all 20-mer peptides, each peptide is decomposed into 19 $AASs^2$. The individual Epitope Index of each AAS^2 were summed up and divided by 19 to obtain a corresponding Epitope Index of a 20-mer peptide. According to the calculation, we could obtain the Epitope Index of each peptide and employ it as the first feature of SVM. Similarly, we apply the same analysis on $AASs^3$ and $AASs^4$. However, there are totally 8,000 possible combinations for 3-residue segments and 160,000 combinations for 4-residue segments. Since a large portion of $AASs^3$ or $AASs^4$ might not appear within the non-epitope dataset, it will cause the problems of dividing by zero. Hence, the calculation of Epitope Indices of $AASs^3$ and $AASs^4$ are different from $AASs^2$. The calculations of

Epitope Indices of AASs^3 (EI_{AAS^3}) and AASs^4 (EI_{AAS^4}) are shown in Equation (7) and (8). It can be obtained by taking the occurrence times of 3- and 4-residue segments ($N^+_{AAS^l}$) divided by the total occurrence times of all segments with 3 and 4 residues ($Total^+_{AASs^l}$) from the epitope dataset respectively, and followed by normalizing the results into the range of [0, 1].

$$EI_{AAS^l} = f^+_{AAS^l} \tag{7}$$

$$normalize(EI_{AAS^l}) = \frac{EI_{AAS^l} - \min(EI_{AASs^l})}{\max(EI_{AASs^l}) - \min(EI_{AASs^l})} \tag{8}$$

where l=3 and 4.

D. SVM training and filtering non-epitope candidates

After determining these three statistical features associated with occurrence frequencies, the proposed system applied all corresponding features for SVM training, and the best model was constructed for the following filtering processes and epitope identification. In this study, a query protein sequence was firstly submitted to the LEPD system, and followed by filtering the LEPD predicted epitope candidates through the designed SVM verification model. Accordingly, the proposed system removed non-epitope candidates through SVM classification and improved the deficiencies of too many false positive predicted epitopes generated by the LEPD system, i.e., reduced the number of false alarm rates.

E. Performance measurement

The sensitivity, specificity, accuracy, positive predictive value (PPV) and Matthews correlation coefficient (MCC) are five criteria for evaluating the performance of the proposed methods. Refer to Equations from (9) to (13), sensitivity represents the percentage of actual epitopes which are correctly identified as epitopes; specificity represents the percentage of non-epitopes which are correctly identified as non-epitopes; accuracy represents the percentage of epitopes and non-epitopes which are correctly identified simultaneously; PPV also called as precision rate, represents the percentage of amino acids predicted as epitopes and which are correctly identified as epitopes; MCC is a measure of predictive performance that incorporates both sensitivity and specificity into a single value between -1 and +1.

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{11}$$

$$PPV = \frac{TP}{TP + FP} \tag{12}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{13}$$

where TP is the number of true positive, FP is the number of false positive, TN is the number of true negative, and FN is the number of false negative.

III. Results

The performances of the proposed methodology on three datasets including HIV (Human Immunodeficiency Virus) [15], AntiJen [16] and PC [9] were analyzed, and which are consisted of 10, 171, and 12 proteins respectively. For comparison, the performances of the previously developed system which adopted antigenicity features only (LEPD system [9]) and the novel proposed system of combining additional machine learning techniques (LEPD+SVM) were evaluated. The LEPD+SVM approach was achieved by selecting the same default window sizes and SVM parameters for all predicting processes on the three testing datasets. The default window sizes in LEPD were set as 7 for calculating antigenicity scale and 11 for local peaks extraction, and the prediction results were improved unanimously in all aspects except sensitivity measurement. The averagely increasing results were 0.3567 in specificity, 0.1589 in accuracy, 0.1039 in PPV, and 0.0744 in MMC, while the sensitivity was decreased by 0.2988 in average, and the details of superior performance of adding SVM techniques are shown in Table 2. In this table, the quantitative data with gray background represent the best results between two approaches regarding to different datasets. It is clear that the combination of propensity scale methods with SVM techniques could improve the performance of LE prediction.

Table 2. Improved performance on 3 datasets

Dataset (#)	Method	Sensitivity	Specificity	Accuracy	PPV	MCC
HIV	LEPD	0.6074	0.5468	0.5549	0.6111	0.1308
(10)	LEPD+SVM	0.4395	0.7865	0.6162	0.7388	0.2330
AntiJen	LEPD	0.5667	0.4519	0.4803	0.2072	0.0216
(171)	LEPD+SVM	0.2416	0.8594	0.7409	0.2907	0.1019
PC	LEPD	0.5503	0.4549	0.4822	0.3392	0.0046
(12)	LEPD+SVM	0.1469	0.8778	0.6371	0.4398	0.0453
Average Increase		-0.2988	0.3567	0.1589	0.1039	0.0744

The symbol '#' denotes the protein number of the datasets.

From Table 3 to Table 5, in comparison with four well-known LE prediction systems including BepiPred [10], ABCpred [17], BCPred [18], and FBCPred [19], our experimental results have shown that the proposed method enhanced the specificity, accuracy, PPV, and MCC for different datasets, but not the sensitivity. The indicators representing the best performance in all five different aspects are also shown with gray background. From these tables, we can observe that ABCpred$_{0.5}$ obtained the best performance in sensitivity among all systems for all three datasets, but the specificity of ABCpred$_{0.5}$ is immensely lower than other systems. It reflects that ABCpred predicts too many epitopes at a time so that high sensitivities could be achieved but with relatively low specificities. As a matter

of fact, the default threshold for ABCpred is defined as 0.5. The higher threshold settings the better specificities can be obtained, but with worse sensitivities correspondingly. It can be observed that our proposed system obtained the best performance in specificity, accuracy and PPV for all datasets except the accuracy for the HIV dataset. To maintain the performance of the proposed SVM mechanism, experimentally verified LEs should be considered as time goes by. Therefore, the proposed system will be designed with an extendible mechanism for updating the contents of experimental epitopes within the trained database, and guarantee that the trained dataset provide the best SVM classifier with respect to the up-to-date LE information.

Table 3. Comparative performance with four other existing systems on HIV dataset.

Method	Sensitivity	Specificity	Accuracy	PPV	MCC
LEPD+SVM	0.4395	0.7865	0.6162	0.7388	0.2330
BepiPred	0.5016	0.6085	0.5672	0.6122	0.0972
$ABCpred_{0.5}$	0.9609	0.0589	0.5584	0.5581	0.0536
$ABCpred_{0.6}$	0.9295	0.0725	0.5600	0.5557	0.0402
$ABCpred_{0.7}$	0.8797	0.1465	0.5659	0.5633	0.0564
$ABCpred_{0.8}$	0.6389	0.4413	0.5577	0.5769	0.0766
BCPred	0.8302	0.5164	0.6699	0.6496	0.3030
FBCPred	0.7696	0.5548	0.6810	0.6504	0.2967

The subscripts of ABCpred denote threshold values.

Table 4. Comparative performance with four other existing systems on AntiJen dataset.

Method	Sensitivity	Specificity	Accuracy	PPV	MCC
LEPD+SVM	0.2416	0.8594	0.7409	0.2907	0.1019
BepiPred	0.5179	0.5761	0.5552	0.2202	0.0604
$ABCpred_{0.5}$	0.9690	0.0473	0.2263	0.2021	0.0318
$ABCpred_{0.6}$	0.9530	0.0725	0.2409	0.2041	0.0398
$ABCpred_{0.7}$	0.8918	0.1450	0.2879	0.2054	0.0416
$ABCpred_{0.8}$	0.6733	0.4040	0.4470	0.2183	0.0546
BCPred	0.6132	0.5264	0.5255	0.2341	0.0917
FBCPred	0.6421	0.4777	0.4937	0.2227	0.0703

Table 5. Comparative performance with four other existing systems on PC dataset.

Method	Sensitivity	Specificity	Accuracy	PPV	MCC
LEPD+SVM	0.1469	0.8778	0.6371	0.4398	0.0453
BepiPred	0.4823	0.5972	0.5533	0.3819	0.0749
$ABCpred_{0.5}$	0.9796	0.0535	0.3660	0.3458	0.0731
$ABCpred_{0.6}$	0.9688	0.0860	0.3833	0.3510	0.0938
$ABCpred_{0.7}$	0.8974	0.1582	0.4075	0.3521	0.0711
$ABCpred_{0.8}$	0.6546	0.4026	0.4889	0.3621	0.0513
BCPred	0.5291	0.5143	0.5183	0.3568	0.0389
FBCPred	0.5431	0.4903	0.5100	0.3508	0.0294

IV. CONCLUSION

The proposed method applied verified epitope and non-epitope segments ranging from 2 to 4 amino acids as the features for SVM classifier and combined with the previously developed propensity scale based LEPD system to predict linear epitopes. With the combination of physico-chemical characteristics and an SVM classifier, the specificity, accuracy and PPV could be effectively improved. In comparison with four well-known prediction systems, experimental results have shown that our proposed method mostly outperforms the existing systems in terms of specificity, accuracy and PPV for different benchmark datasets including HIV, AntiJen and PC.

LE prediction is an important research topic for biological and medical researches such as vaccines design and disease diagnosis. To further improve the prediction performance of the proposed system, properly selecting crucial features is an important issue to overcome. Besides, with the design of extendable training mechanism, our proposed prediction system can include more experimentally verified epitopes to improve classification on epitope and non-epitope peptides. Under these filtration and enhancement, the accuracy of LE prediction can be guaranteed.

ACKNOWLEDGMENT

This work is supported by the National Science Council (NSC 96-2221-E-019-043 to T.-W. Pai), China Medical University (CMU97–289 to H.-T. Chang), and the Center of Excellence for Marine Bioenvironment and Biotechnology in National Taiwan Ocean University in Taiwan, R.O.C.

REFERENCES

[1] D.J. Barlow, et al., "Continuous and discontinuous protein antigenic determinants," *Nature*, vol. 322, no. 6081, 1986, pp. 747-748; DOI 10.1038/322747a0.

[2] G. Walter, "Production and use of antibodies against synthetic peptides," *J Immunol Methods*, vol. 88, no. 2, 1986, pp. 149-161.

[3] M. Vihinen, et al., "Accuracy of protein flexibility predictions," *Proteins*, vol. 19, no. 2, 1994, pp. 141-149; DOI 10.1002/prot.340190207.

[4] E.A. Emini, et al., "Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide," *J Virol*, vol. 55, no. 3, 1985, pp. 836-839.

[5] J.M. Parker, et al., "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, no. 19, 1986, pp. 5425-5432.

[6] L. Debelle, et al., "Predictions of the secondary structure and antigenicity of human and bovine tropoelastins," *Eur Biophys J*, vol. 21, no. 5, 1992, pp. 321-329.

[7] A.S. Kolaskar and P.C. Tongaonkar, "A semi-empirical method for prediction of antigenic determinants on protein antigens," *FEBS Lett*, vol. 276, no. 1-2, 1990, pp. 172-174; DOI 10.1016/0014-5793(90)80535-Q.

[8] M.J. Blythe and D.R. Flower, "Benchmarking B cell epitope prediction: underperformance of existing methods," *Protein Sci*, vol. 14, no. 1, 2005, pp. 246-248; DOI 10.1110/ps.041059505.

[9] H.T. Chang, et al., "Estimation and extraction of B-cell linear epitopes predicted by mathematical morphology approaches," *J Mol Recognit*, vol. 21, no. 6, 2008, pp. 431-441; DOI 10.1002/jmr.910.

[10] J.E. Larsen, et al., "Improved method for predicting linear B-cell epitopes," *Immunome Res*, vol. 2, 2006, pp. 2; DOI 10.1186/1745-7580-2-2.

[11] J. Chen, et al., "Prediction of linear B-cell epitopes using amino acid pair antigenicity scale," *Amino Acids*, vol. 33, no. 3, 2007, pp. 423-428; DOI 10.1007/s00726-006-0485-9.

[12] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001.

[13] S. Saha, et al., "Bcipep: a database of B-cell epitopes," *BMC Genomics*, vol. 6, no. 1, 2005, pp. 79; DOI 10.1186/1471-2164-6-79.

[14] A. Bairoch and R. Apweiler, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000," *Nucleic Acids Res*, vol. 28, no. 1, 2000, pp. 45-48; DOI 10.1093/nar/28.1.45.

[15] B.T.M. Kober, et al., "HIV Immunology and HIV/SIV Vaccine Databases 2003," 2003.

[16] H. McSparron, et al., "JenPep: a novel computational information resource for immunobiology and vaccinology," *J Chem Inf Comput Sci*, vol. 43, no. 4, 2003, pp. 1276-1287; DOI 10.1021/ci030461e.

[17] S. Saha and G.P. Raghava, "Prediction of continuous B-cell epitopes in an antigen using recurrent neural network," *Proteins*, vol. 65, no. 1, 2006, pp. 40-48; DOI 10.1002/prot.21078.

[18] Y. El-Manzalawy, et al., "Predicting linear B-cell epitopes using string kernels," *J Mol Recognit*, vol. 21, no. 4, 2008, pp. 243-255; DOI 10.1002/jmr.893.

[19] Y. El-Manzalawy, et al., "Predicting flexible length linear B-cell epitopes," *Comput Syst Bioinformatics Conf*, vol. 7, 2008, pp. 121-132.