DOCUMENT of B_pred

B_MM, BPBB and JBFB: Three Highly Efficient Linear B-cell Epitope Prediction Tools

INTRODUCTION

B-cell epitopes are antigen deterministic parts of a protein that antibodies or B-cell receptors can bind. Identification of B-cell epitopes greatly facilitates peptide vaccine design, antibody generation, and molecular diagnosis. The experimental approaches are often laborious and expensive, and consequently, *in silico* prediction models appear quite attractive, which can guide the experimental identification processes more efficient, and less costly. Most of the B-cell epitopes (90%) consist of amino acids sequentially sequestered but spatially brought into proximity by protein folding. These epitopes are called discontinuous or conformational epitopes. The rest are called linear epitopes, which are mainly composed of a contiguous stretch of amino acids. Though the number is small, linear B-cell epitopes have received much research interest since they are easier to predict and more convenient for application.

Initial prediction of B-cell epitopes was based on amino acid propensity scale (Hopp and Woods, 1981). It endued each amino acid a propensity value (e.g., hydrophilicity) and recursively calculated the average score over a sliding window of a fixed length. The location

of an epitope was delineated according to the average score of a local region. This propensity scale strategy was widely adopted by different prediction models (Kolaskar and Tongaonkar, 1990; Alix, 1999; Odorico and Pellequer, 2003). A subsequent study compared the performance of prediction models based on 484 single amino acid propensity scales, and demonstrated that these models didn't perform better than a random model significantly, however (Blythe and Flower, 2005). New scales, e.g., AAPs, were explored and used for development of new B-cell epitope prediction tools (Chen et al., 2007). Besides, the performance has been greatly improved for the machine learning models developed recently (Saha and Raghava, 2006; Chen et al., 2007; EL-Manzalawy et al., 2008; Sweredoski and Baldi, 2009; Wee et al., 2010). Not like previous propensity scale methods, machine learning techniques can take full advantage of information from known epitopes or non-epitopes.

While most machine learning methods learned amino acid propensity or sequence-derived features, Wee et al presented a unique model, BayesB, which adopted a support vector machine (SVM) to learn position-specific amino acid composition (Aac) features of both epitopes and non-epitopes (Wee et al., 2010). The method was shown to outperform other similar applications, such as BCPred (EL-Manzalawy et al., 2008), COBEpro (Sweredoski and Baldi, 2009) and Chen et al's method (Chen et al., 2007), on independent datasets. In practice, however, the tool seems to have little use because it predicts too many positive epitopes within a query antigen; and, for each searching peptide, it simply classifies it as an epitope or non-epitope, without any score, rank or probabilistic confidence. In addition, the model only learns Aac features without integration of peptide secondary structure (Sse) or solvent accessibility (Acc), which could represent important properties of an epitope. In this research, we developed JBFB, a Joint Bayesian Features-based B-cell

epitope prediction tool, which integrated all Aac, Sse and Acc features in a Bi-Profile Baysian model (Sweredoski and Baldi, 2009). As a complement, we also set up a high-effective Markov chain-based model, B_MM, regardless of any position or sequence length related information, which merely took the features of Aac conditional on sequentially adjacent amino acids. Both JBFB and B_MM apparently outperformed BayesB and other software in terms of accuracy, sensitivity, specificity and other performance assessment items. Both of them can also assign each peptide a prediction score or probability to be an epitope.

METHODS

1. Datasets

EL-Manzalawy et al annotated a non-homologous 28-mer dataset containing 637 B-cell epitopes and 637 non-epitopes. The epitopes were originally retrieved from Bcipep database while the non-epitopes were prepared from the protein sequences stored in SwissProt database. The 28-mer dataset rather than the more frequently cited 20-mer dataset was used for training the models in this research, since the Sse of terminal 3 amino acids at each side of a peptide is often inaccurately predicted by software. The original 28-mer peptides were predicted for their Sse and Acc with Sspro 4, and the primary sequences, Sse and Acc of 4-mer regions at each side were truncated thereafter so as to obtain the sequences, Sse and Acc of central 20-mer fragments. These 20-mer peptides, Sse and Acc comprised the training datasets. The training datasets of for models of different peptide length (8-, 10-, 12-, 14-, 16-, or 18-mer) were directly derived from these 20-mer datasets, truncating equal number of amino acids, Sse or Acc elements at each side from sequences in corresponding datasets. The Sse of each peptide was represented as a sequence of H, E or C, which represents ‘helix’, ‘strand’ or ‘coil’ respectively, while Acc was represented as a combination of B or E, representing ‘buried’ or ‘exposed’ respectively.

EL-Manzalawy’s 20-mer dataset, Chen’s dataset, and Saha and Raghava’s dataset were also adopted, used as independent datasets for model performance comparison.

2. Position-specific Aac and joint features

The Bi-profile Bayesian Aac features were extracted from positive and negative peptide datasets, according to Shao et al., 2009, Wee et al., 2010 and Wang et al., 2011. The position-specific Aac features of epitopes and non-epitopes were furthermore learned with a SVM.

Extraction of joint position-specific features can refer to Wang et al., 2013a. Briefly, let $S = \{s_1,s_2\ldots,s_{i-1},s_i,\ldots,s_n\}$ represents a sequence of peptide (epitope or non-epitope), where s_i represents amino acid at position i and n represents total length of the sequence. Similarly, let $A = \{a_1,a_2\ldots,a_{i-1},a_i,\ldots,a_n\}$ and $B = \{b_1,b_2\ldots,b_{i-1},b_i,\ldots,b_n\}$ represent Sse and Acc of S respectively, while a_i and b_i represent the Sse and Acc element at position i respectively. For each position of the epitopes or non-epitopes, the joint probability of Aac, Sse and Acc can be represented as a probability vector, $\mathbf{P}i = \{s_i, a_i, b_i\,|C\}$, in which C is epitope or non-epitope, the s_i, a_i and b_i could be each type of amino acid, Sse element and Acc element, respectively. A $120 \times n$ probability profile matrix was obtained for total n positions of either type of peptides (epitopes or non-epitopes). The joint probabilities were approximated with a maximum likelihood method. Each training sequence was further represented as a vector of joint probabilities according to epitope and non-epitope joint probability profiles. Therefore, a sequence with length n was finally represented as a vector with $2n$ probability elements. The total m training sequences led to an $m \times 2n$ probability matrix. A SVM was used to train these position-specific joint features.

Radial basis kernel function $K(S_i,S_j) = \exp(-\gamma\|S_i - S_j\|^2)$ was selected for SVM prediction. SVM parameter γ and penalty parameter C were optimized using grid search based on 10-fold cross-validation.

3. Aac conditional on preceding adjacent position and Markov model

Given a peptide sequence $S = \{s_1,s_2\ldots,s_{i-1},s_i,\ldots,s_n\}$. For any k ($1< k \leq n$), $P\{s_1,s_2\ldots,s_{k-1},s_k\} = P\{s_1,s_2\ldots,s_{k-1}\} \times P\{s_i|s_1,s_2\ldots,s_{k-1}\}$. Assume Aac at each position only depends on its immediately preceding Aac, which means $P\{s_k|s_1,s_2\ldots,s_{k-1}\} = P\{s_k|s_{k-1}\}$. Therefore, P(S) can be represented as a one-order Markov chain, $P\{s_1,s_2\ldots,s_{i-1},s_i,\ldots,s_n\} = \prod P\{s_k|s_{k-1}\} \times P\{s_1\}$, where $1< k \leq n$. There are two categories of sequences, epitopes (C_1) and non-epitopes (C_{-1}). For either category C, $P(S|C) = P\{s_1,s_2\ldots,s_{i-1},s_i,\ldots,s_n|C\} = \prod P\{s_k|s_{k-1}, C\} \times P\{s_1|C\}$, where $1< k \leq n$. Each conditional probability and $P\{s_1|C\}$ can be estimated using a maximum likelihood strategy. Both $P(S|C_1)$ and $P(S|C_2)$ were calculated for each positive or negative training sequence. An R value was calculated using the formula: $R = \log\{P(S|C_1)/ P(S|C_2)\}$, where the logarithm base was 2. A Gaussian distribution was approximated for the R values of positive and negative training sequences, respectively, using the method described in Wang et al., 2013b. The decision function was also adopted to discriminate B-cell epitope and non-epitopes (Wang et al., 2013b).

4. Performance assessment

The performance of models was compared based on five-fold cross-validation results. The parameters for performance assessment include Accuracy (*A*), Specificity (*Sp*), Sensitivity (*Sn*), Receiver Operating Characteristic (*ROC*) curve, the area under *ROC* curve (*AUC*) and Matthews Correlation Coefficient (*MCC*). The definition of these parameters refers to Wang et al, 2011.

EVALUATION REPORT

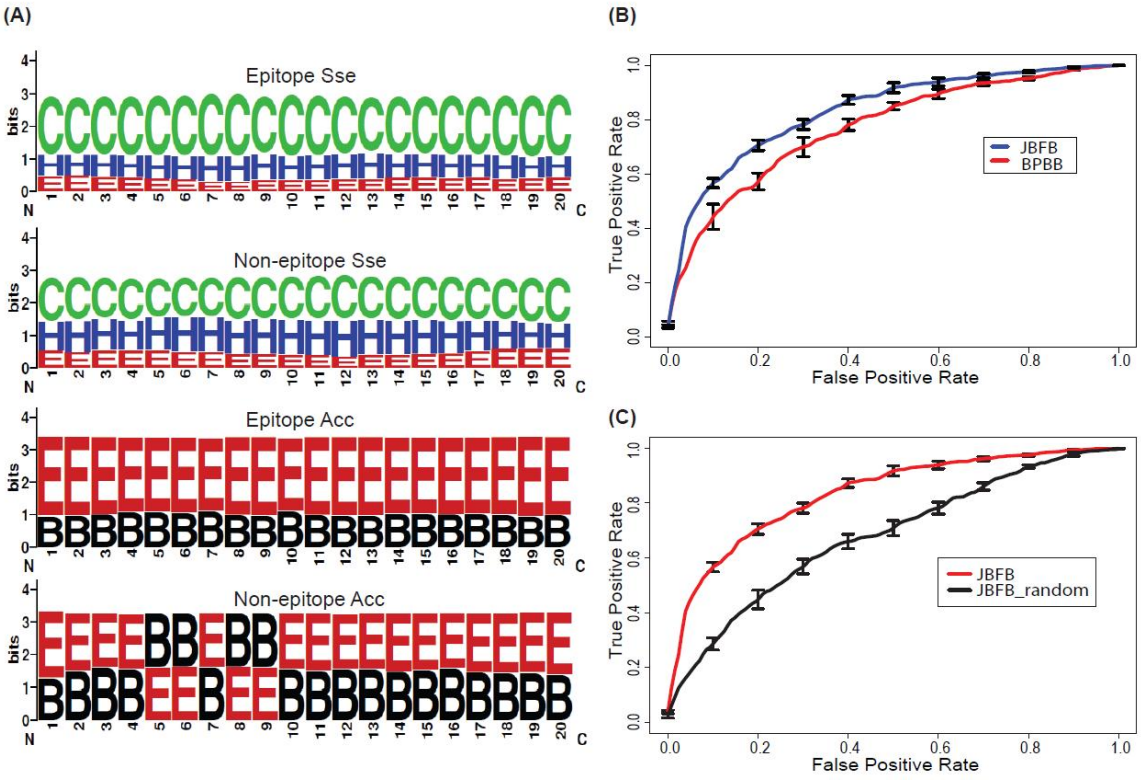
1. Improvement of B-cell epitope prediction performance by integration of Sse and Acc features

The Sse and Acc of central 20-mer fragments of training sequences were retrieved and compared. As shown in Fig 1A ('Epitope Sse' and 'Non-epitope Sse'), the epitopes showed higher coil and lower helix or strand preference than non-epitopes. The Acc analysis also demonstrated that the amino acids in epitopes preferred to be exposed (Fig 1A, 'Epitope Acc' and 'Non-epitope Acc'). Consistent with prior knowledge about antigenic regions, the results indicated that the B-cell epitopes tend to be more flexible and hydrophilic.

The position-specific Sse and Acc features were integrated with Aac profiles in a Bayesian model and trained with SVM. The classification performance of the generated model, JBFB, was compared with that of BPBB (Bi-Profile Bayes based B-cell prediction tool), a revised version of BayesB, which learned Aac features only. Supplemental Table S1 listed the optimized parameters of JBFB and BPBB models. Fig 1B showed the ROC curves of these two models classifying the training data based on five-fold cross-validation results. Apparently, JBFB outperformed BPBB, indicating the Sse and Acc features can be used to strengthen the power of distinguishing B-cell epitope and non-epitopes. Table 1 also gave the *Sn*, *Sp*, *A*, *AUC*, and *MCC*, for all of which JBFB showed higher values than BPBB.

Permutation was performed to the training datasets, and a random model was set up based on the permutated data. With the same Bi-Profile Bayesian feature extraction and

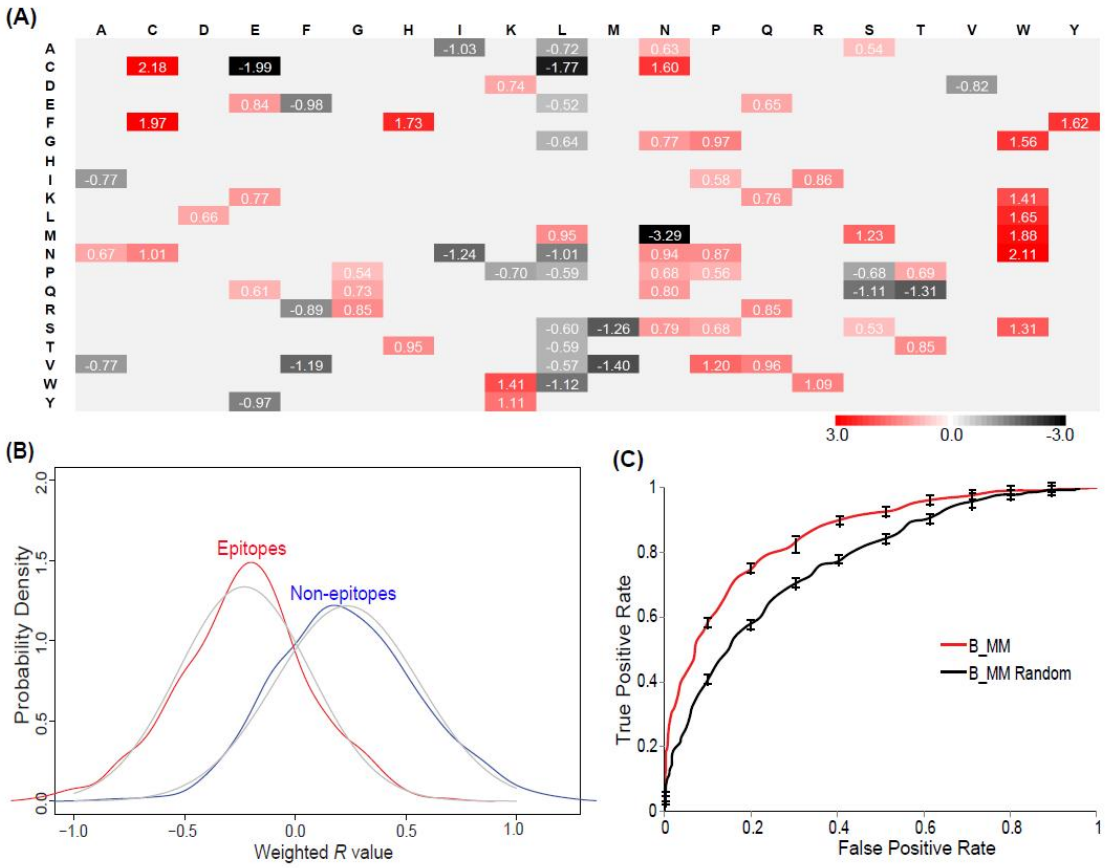
SVM training, the random model only distinguished epitopes and non-epitopes weakly (Fig 1C), demonstrating the good classification performance of JBFB is potentially due to the unique joint features of Sse, Acc and Aac in B-cell epitopes rather than overrepresentation of the features by complex SVM hyperplanes.



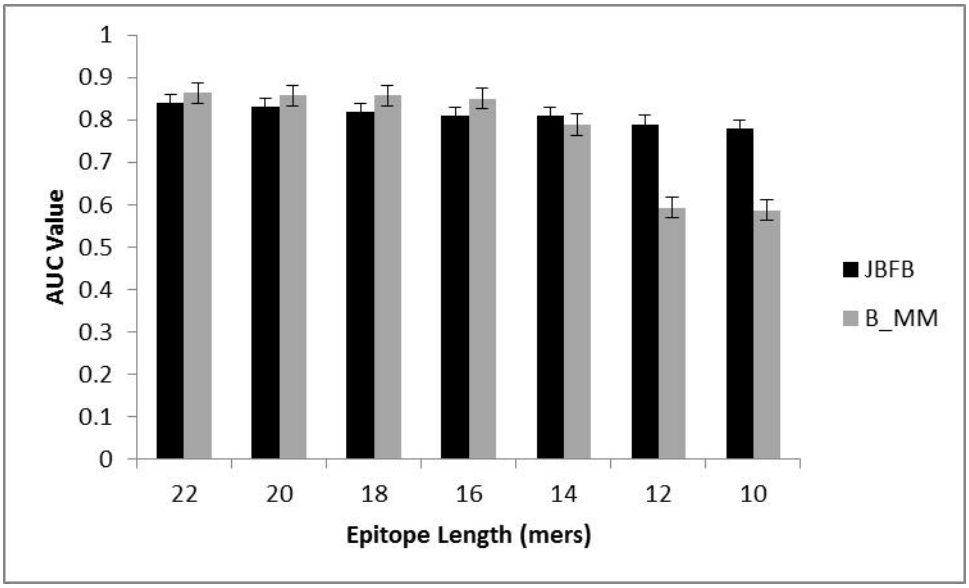
2. Discretion of B-cell epitopes and non-epitopes based on neighbor position-conditional Aac features and a Markov model

Besides the Sse and Acc preference in B-cell epitopes, we also observed the constraint of Aac posed by the amino acid species at its preceding neighbor position. For B-cell epitopes, amino acids were not evenly distributed when the amino acid was given at its preceding position. Fig 2A showed all the amino acids whose composition probability conditional on its preceding Aac was significantly different between epitopes and non-

epitopes (Bonferroni corrected binomial test, $P<0.05$). A Markov model was initiated and the likelihood ratios were calculated according to the first-order position-conditional Aac probability profiles of epitopes and non-epitopes. The likelihood ratios (R) of epitopes and non-epitopes were further approximated as two distinct Gaussian distributions (Fig 2B). A discretion model, B_MM, was set up based on the R distributions. A five-fold cross validation demonstrated the model had an excellent classification power with an AUC of 0.857, even better than JBFB (Fig 2C; Table 1). Permutation test also demonstrated that B_MM performed significantly better than random models (Fig 2C).



3. Epitope length and Performance comparison



It remains unclear about the optimized length of B-cell epitopes. Therefore, we also tested the performance of the models with different epitope length. Epitopes of 10-mer, 12-mer, 14-mer, 16-mer, 18-mer, 20-mer and 22-mer were constructed and trained with JBFB, BPBB and B_MM algorithms, respectively. For both algorithms, the performance deteriorated as the length being reduced, though it only slightly decreased for JBFB models but deteriorated strikingly for B_MM 12-mer and 10-mer models (Fig 3). For either of the algorithms, performance difference was extremely subtle for the 22-mer and 20-mer models (Fig 3); the optimized length of JBFB, BPBB and B_MM models was therefore set as 20-mer.

JBFB, BPBB and B_MM 20-mer models were further trained with other independent datasets, followed by a performance comparison with other software tools, including BayesB, BCPred, ABCPred and Chen's method. Generally, JBFB showed the best performance (Table 2).

Table 1. Five-fold cross-validation performance of JBFB, BPBB and B_MM

Model	S_n vs. S_p (%)	*A* (%)	*AUC*	*MCC*
BPBB	71.41 vs. 70.47	70.94	0.7699	0.4188
JBFB	74.69 vs. 75.16	74.92	0.8324	0.4984
B_MM	75.67 vs. 79.43	77.55	0.8570	0.5514

Table 2. Comparison of model performance with independent datasets

Dataset	Method[1]	*Sn*	*Sp*	*A*	*AUC*	*MCC*
BCPred_20	BCPred	0.726	0.632	0.679	0.758	0.36
	AAP	0.529	**0.752**	0.641	0.700	0.288
	BPBB	0.697	0.710	0.704	0.774	0.407
	JBFB	0.724	0.750	**0.737**	**0.807**	**0.474**
	B_MM	**0.732**	0.635	0.683	NA	0.368
ABCPred_20	ABCPred	0.571	0.716	0.644	NA	0.287
	BPBB	0.66	0.73	0.671	0.768	0.391
	JBFB	0.684	**0.769**	**0.718**	**0.792**	**0.454**
	B_MM	**0.759**	0.639	0.699	NA	0.400
AAP_20	AAP	0.609	**0.754**	**0.711**	NA	0.366
	BPBB	0.671	0.71	0.691	0.756	0.382
	JBFB	0.693	0.714	0.703	**0.770**	**0.407**
	B_MM	**0.750**	0.622	0.686	NA	0.375

[1] The performance of different models were evaluated based on the average 5-fold cross-validation results except B_MM, for which the performance was evaluated according to direct predictions on each test dataset with the original model trained on BCPred_28 centered datasets.

Table S1. Optimized parameters for BPBB and JBFB models with a 10-fold cross-validation grid search

Model	Features[a]	Kernel	Gamma	Cost
BPBB	Aac	RBF[b]	2^{-10}	8
JBFB	Joint Aac, Sse, Acc	RBF	2^{-11}	16

a. Position-specific features;

b. Radial basis kernel function.

APPLICATION

Both standalone version of the software tools and the webserver were developed to implement B_MM, BPBB and JBFB. The tools and the webserver could be available via the website: http://61.160.194.165:3080/B_pred/.

With the webserver, users could implement the three models simultaneously or independently. The decision cutoff is set as 0 for the models by default, meaning that a fragment with a value larger than 0 would be predicted as a positive B epitope. Users can also change the cutoff by themselves. For example, a larger cutoff (0.5 for example) is recommended if larger precision is desired.

References

Ben-Yedidia T, Beignon AS, Partidos CD, Muller S, Arnon R. (2002). A retro-inverso peptide analogue of influenza virus hemagglutinin B-cell epitope 91–108 induces a strong mucosal and systemic immune response and confers protection in mice after intranasal immunization. Mol Immunol. 39(5-6):323-331.

Chen J, Liu H, Yang J, Chou KC. (2007). Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids. 33:423-428.

De Groot AS, Ardito M, Terry F, Levitz L, Ross T, Moise L, Martin W. (2013). Low immunogenicity predicted for emerging avian-origin H7N9: implication for influenza vaccine design. Hum Vaccin Immunother. 9(5):950-956.

EL-Manzalawy Y, Dobbs D, Honavar V. (2008). Predicting linear B-cell epitopes using string kernels. J Mol Recognit. 21(4):243-255.

Gao J, Faraggi E, Zhou Y, Ruan J, Kurgan L. (2012). BEST: improved prediction of B-cell epitopes from antigen sequences. PLoS One. 7(6):e40104.

Hopp TP, Woods KR. (1981). Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci U S A. 78(6):3824-3828.

Larsen JEP, Lund O, Nielsen M. (2006). Improved method for predicting linear B-cell epitopes. Immunome Res. 2:1-7.

Liang S, Zheng D, Standley DM, Yao B, Zacharias M, Zhang C. (2010). EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results. BMC Bioinformatics. 11:1-6.

Nielsen M, Lundegaard C, Worning P, Hvid CS, Lamberth K, Buus S, Lund O. (2004). Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. Bioinformatics. 20(9):1388-1397.

Ponomarenko J, Bui HH, Li W, Fusseder N, Bourne PE, Sette A, Peters B. (2008). ElliPro: a new structure-based tool for the prediction of antibody epitopes. BMC Bioinformatics. 9:1-8.

Ponomarenko JV, Bourne PE. (2007). Antibody-protein interactions: benchmark datasets and prediction tools evaluation. BMC Struct Biol. 7:1-19.

Saha S, Raghava GPS. (2006). Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins. 65(1):40-48.

Srivastava V, Yang Z, Hung IFN, Xu J, Zheng B, Zhang MY. (2013). Identification of dominant antibody-dependent cell-mediated cytotoxicity epitopes on the hemagglutinin antigen of pandemic H1N1 influenza virus. J Virol. 87(10):5831-5840.

Sweredoski MJ, Baldi P. (2008). PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. Bioinformatics. 24(12):1459-1460.

Sweredoski MJ, Baldi P. (2009). COBEpro: a novel system for predicting continuous B-cell epitopes. Protein Eng Des Sel. 22(3):113-120.

Terajima M, Babon JAB, Co MDT, Ennis FA. (2013). Cross-reactive human B cell and T cell epitopes between influenza A and B viruses. Virol J. 10:1-10.

Wang Y, Zhang Q, Sun MA, Guo D. (2011). High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. Bioinformatics. 27(6):777-84.

Wang Y, Sun M, Bao H, Zhang Q, Guo D. (2013a). Effective identification of bacterial type III secretion signals using joint element features. PLoS One. 8(4):e59754.

Wang Y, Sun M, Bao H, White AP. (2013b). T3_MM: a Markov model effectively classifies bacterial type III secretion signals. PLoS One. 8(3):e58173.

Wee LJ, Simarmata D, Kam YW, Ng LF, Tong JC. (2010). SVM-based prediction of linear B-cell epitopes using Bayes feature extraction. BMC Genomics. 11:1-9.

Yao B, Zheng D, Liang S, Zhang C. (2013). Conformational B-cell epitope prediction on antigen protein structures: a review of current algorithms and comparison with common binding site prediction methods. PLoS One. 8(4):e62249.

Yao B, Zhang L, Liang S, Zhang C. (2012). SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity.

Yasser EM, Dobbs D, Honavar V. (2008). Predicting flexible length linear B-cell epitopes. Comput Syst Bioinformatics. 7:121.