W24–W29 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* *Published online 2 May 2017*
doi: 10.1093/nar/gkx346

BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes

Martin Closter Jespersen[1], Bjoern Peters[2], Morten Nielsen[1,3,*] and Paolo Marcatili[1,*]

[1]Department of Bio and Health Informatics, Technical University of Denmark, Kgs. Lyngby 2800, Denmark, [2]La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037, USA and [3]Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Buenos Aires, Argentina

Received February 02, 2017; Revised April 04, 2017; Editorial Decision April 16, 2017; Accepted April 20, 2017

ABSTRACT

Antibodies have become an indispensable tool for many biotechnological and clinical applications. They bind their molecular target (antigen) by recognizing a portion of its structure (epitope) in a highly specific manner. The ability to predict epitopes from antigen sequences alone is a complex task. Despite substantial effort, limited advancement has been achieved over the last decade in the accuracy of epitope prediction methods, especially for those that rely on the sequence of the antigen only. Here, we present BepiPred-2.0 (http://www.cbs.dtu.dk/services/BepiPred/), a web server for predicting B-cell epitopes from antigen sequences. BepiPred-2.0 is based on a random forest algorithm trained on epitopes annotated from antibody-antigen protein structures. This new method was found to outperform other available tools for sequence-based epitope prediction both on epitope data derived from solved 3D structures, and on a large collection of linear epitopes downloaded from the IEDB database. The method displays results in a user-friendly and informative way, both for computer-savvy and non-expert users. We believe that BepiPred-2.0 will be a valuable tool for the bioinformatics and immunology community.

INTRODUCTION

B-cells are considered a core component of the adaptive immune system, as they have the ability to recognize and provide long-term protection against infectious pathogens or cancerous cells. They perform these functions by producing antibodies, proteins that are either secreted or expressed on the B-cell surface, and that recognize their molecular target (called antigen) by binding to a part of it (called epitope) in a highly selective manner. This recognition process is exploited in vaccines to provide a long-term protection toward desired pathogens, using different methods, such as attenuated and subunit vaccines.

B-cell epitopes can be divided into two groups. Linear epitopes are formed by linear stretches of residues in the antigen protein sequence. In contrast, discontinuous (conformational) epitopes are formed by residues far apart in the antigen sequence that are brought together in space by its folding. Even though the majority of epitopes are conformational, most contain one or few linear stretches (1).

Reliable B-cell epitope prediction tools are of primary importance in many clinical and biotechnological applications such as vaccine design and therapeutic antibody development, and for our general understanding of the immune system (2–4).

Several structure-based tools have been developed and can be used to predict and analyse epitopes when the antigen structure is known (5–9). However, structural information is only available for a very small proportion of antigens, and in the vast majority of cases one is left with analyzing the primary sequence only. The accuracy of such sequence-based predictors is generally poor, and little improvements have been achieved over the past years. The training of current methods is in most cases based on of peptides experimentally validated to bind antibodies (10–13) and are generally associated with low performance of prediction tools (4), which could be due to the starting data being poorly annotated and noisy.

Here, we present BepiPred-2.0, a web server for sequence-based B-cell epitope prediction. Unlike the BepiPred-1.0, BepiPred-2.0 is trained only on epitope data derived from crystal structures, which is presumed to be of higher quality and indeed resulted in a significantly improved predictive power when compared to other available tools (10,11).

MATERIALS AND METHODS

We describe briefly the dataset and method used for training BepiPred-2.0, and the validations we have performed. More details on the material and methods can be found in Supplementary Materials.

*To whom correspondence should be addressed. Tel: +45 4525 2489; Fax: +45 4593 1585; Email paolo.marcatili@gmail.com
Correspondence may also be addressed to Morten Nielsen. Email: mniel@cbs.dtu.dk

Structural dataset

A dataset consisting of 649 antigen-antibody crystal structures was obtained from the Protein Data Bank (PDB) (14). In each complex, we identified the antibody molecules using HMM models developed elsewhere, and for each antibody we define its antigens as all the non-antibody protein chains that have at least one atom in a 4 Å radius from its Complementarity Determining Region (CDR) atom (15). We removed complexes in which the antigen sequence was >70% identical to any other sequence in our dataset, thus obtaining 160 structures. We randomly selected five structures published after 2014 as a final evaluation dataset and used the remaining 155, split into five equally-sized partition for cross-validation, to create our training dataset. The epitope residues were defined as those in a 4 Å radius of any antibody residue's heavy atom. Also, if multiple identical antigen chains bind to the same antibody, the epitope was defined as the union of the epitope residues on all the chains, thus resulting in a positive dataset of 3542 residues. All 36 785 non-epitopes were defined as negatives. All the positive and negative residues were used when evaluating the methods' performance, but for training the negative dataset was downsized by random sampling to the same size of the positive one (see Supplementary Materials for more details).

Training a random forest prediction model

To predict the probability that a given antigen residue is part of an epitope, a Random Forest Regression (RF) algorithm was trained using a 5-fold cross validation approach. Each residue was encoded using its computed volume (16), hydrophobicity (17), polarity (18), together with the relative surface accessibility (RSA) and secondary structure (SS) as predicted by NetSurfP (19) of all the residues in a window of size 9 centered on the residue itself. Also, the overall volume of the antigen obtained by summing the individual volumes of all the antigen's residues was used, for a total of 46 variables. A rolling average of window 9 was then performed on the RF output to obtain the final BepiPred-2.0 predictions. More details on the parameter optimization can be found in the Supplementary text and Supplementary Figures S1 and S2.

Evaluation measurements

We evaluated the performance for each antigen in terms of the area under the receiver operation curve (AUC), the area under the first 10% of the receiver operation curve normalized by multiplying by 10 (AUC10%), the positive predictive rate (PPR) and the true positive rate (TPR) of the top 60 predictions (20).

When comparing the performance of two models, a paired *t*-test was calculated on their performances on individual antigens. A confidence interval of 95% was used to define a significant difference between two compared models.

Evaluation on a linear epitopes dataset

A set of known linear peptides that were tested for immune recognition and were found to be epitopes (positive assay results) or non-epitopes (negative assay results) were downloaded from the Immune Epitope Database (IEDB) (21). Peptides shorter than five or larger than 25 amino acids were removed, as B cell epitopes rarely are outside these boundaries (1). Only peptides confirmed as positives in two or more separate experiments were included in the positive dataset, and only peptides seen as negative in two or more separate experiments and never observed as positives in any experiment were included in the negative dataset. This resulted in 11 834 positives and 18 722 negative peptides. Each peptide was mapped back on its original protein sequence, and this was used to calculate the output prediction. This dataset is available for download on the BepiPred web page (http://www.cbs.dtu.dk/services/BepiPred/download.php).

The evaluation was only performed on the residues within the positive and negative peptides. In this case, an AUC was calculated only on the pooled positive and negative residues and not per antigen sequence.

WEB INTERFACE

In order to use BepiPred-2.0 (http://www.cbs.dtu.dk/services/BepiPred/) the user only needs the sequences of the protein of interest in fasta format. All the predictions are done on the fly, and in seconds to minutes, depending mainly on the size of the input data, the user will be redirected to the result page. Here, the predicted epitopes are indicated in the input protein sequences. All the most common browsers such as Chrome, Firefox, Microsoft Edge and Safari are supported, but some graphical features are not available on Internet Explorer. In the following paragraphs, the input page and the output pages will be described in further detail. Some tips and tricks and a more detailed description of the web server can be found on BepiPred-2.0 Instructions/Help page.

Input page

The user can submit up to 50 protein sequences in fasta format either by pasting them into the textbox or by uploading them as a single file. Nucleic acid sequences are not supported and protein sequences should be longer than 10 amino acids and shorter than 6000. Example sequences are available when clicking the button 'Example Antigens'. When clicking 'Submit' the user will be redirected to a job queue page which is updated every 20 s. When the predictions are completed, the user will be automatically redirected to the output page. Optionally, the user can provide an email address and the result page link will be emailed when the job is completed.

Output page

The BepiPred-2.0 output page contains a navigation bar with various tabs. The 'Summary' tab shows each individual sequence result in a horizontal and vertical scrollable table. The default output format shows the BepiPred-2.0 predictions and epitope classification for each sequence. The BepiPred-2.0 predictions are used to set the background color of the protein sequences. All predictions greater than a user-defined threshold (by default 0.5) are marked as 'E' in

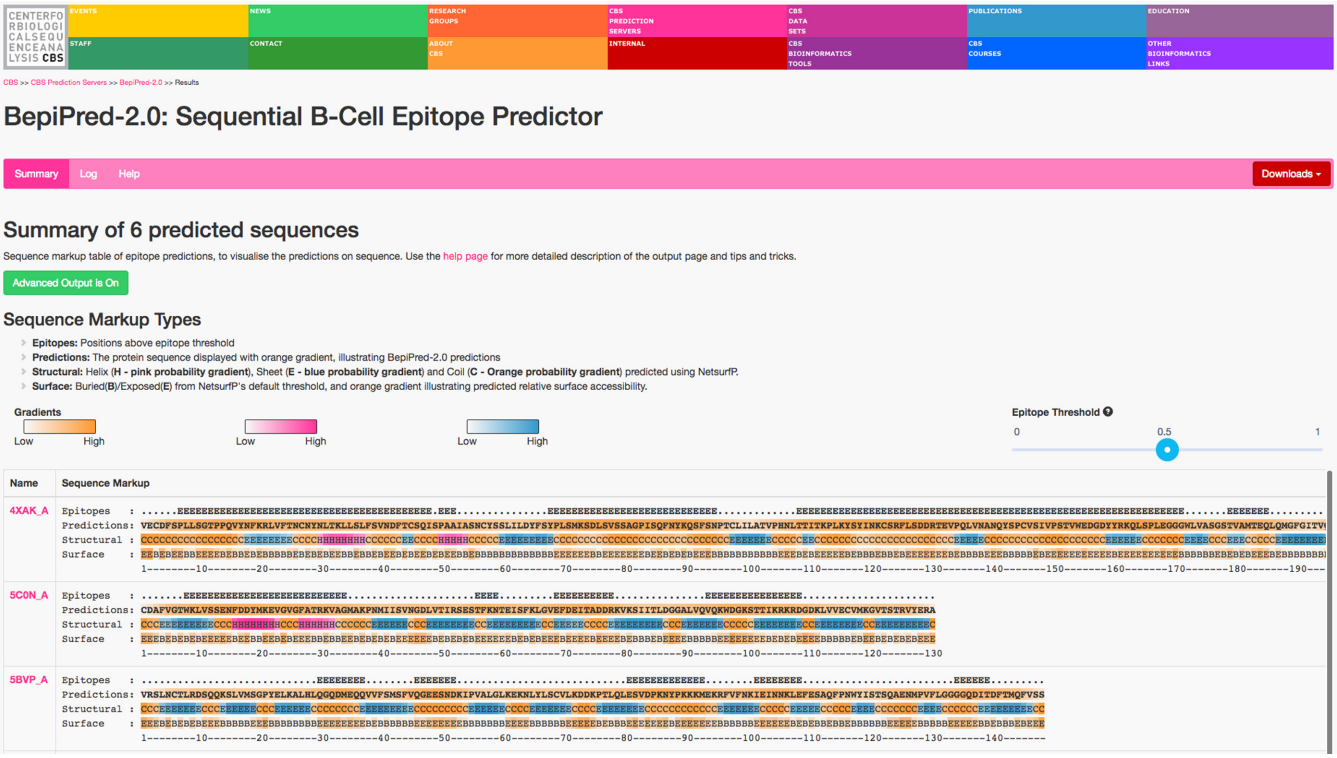

Figure 1. The Summary output page in Advanced Output mode, showing BepiPred-2.0 and NetSurfP predictions for each query sequence.

the 'Epitopes' line above the protein sequence itself. Using the 'Epitope Threshold' slider the epitope classifications can be modified as desired. By pressing the '?' button next to the threshold slider, a plot of the expected sensitivity and specificity for each threshold value will be displayed. Hovering the mouse over the sequences shows the prediction values of the specific residue and hovering over the protein name will reveal the description of the protein from the fasta header. Clicking on the 'Advanced Output is Off' button will switch to the advanced visualization mode, in which structural predictions from NetSurfP are added, as shown in Figure 1. This allows experienced users to display detailed information and achieve a better interpretation of the results. The 'Log' tab will show a log of the computations and possible errors that have occurred and the 'Help' tab contains tips and tricks and a detailed description of the output page. The predictions can be downloaded as JSON or CSV format by using the dropdown tab 'Downloads' and a short description of the files can be found by clicking 'All Downloads'.

RESULTS

Here, we demonstrate the functionality of Bepipred-2.0, a web server constructed from a large set of structurally defined B cell epitopes, and show how this updated method significantly outperform BepiPred-1.0 (10) on both structural and linear epitope validation datasets.

Cross-validation results

We used a 5-fold cross validation approach to estimate the performance of the Bepipred-2.0 method. The dataset consisted of epitopes derived from 165 solved structures, in which no two antigens shared >70% sequence similarity. The final RF model achieved an AUC of 0.62 and an AUC10% of 0.121 on this test set of structural epitopes.

Figure 2 displays the Gini importance, describing the importance of each feature, for each variable and cross-fold partition (22). The Gini importance values for the 5 RF models are highly consistent, confirming the robustness of the proposed model. Moreover, the figure shows that besides the residue type, the predicted RSA is one of the most important features contributing to the predictive power of our method.

On this dataset, BepiPred-2.0 outperforms the two other tested methods, namely BepiPred-1.0 (10), LBtope (11), both among the most used methods for linear epitope prediction. Likewise the method outperforms a baseline predictor solely based on the RSA values provided by NetSurfP. The AUC, AUC10% and corresponding p values are displayed in Table 1.

In many real-case scenarios, the users are only interested in analyzing the top-scoring predictions, as they are using the predictions to prioritize a few candidates for experimental testing. To assess this case, the average positive predictive value (PPV) and true positive rate (TPR) on the 60 top-scoring residues per protein were calculated. As evident from Figure 3, BepiPred-2.0 achieves a significantly better PPV and a marginally better TPR when compared to the other methods.

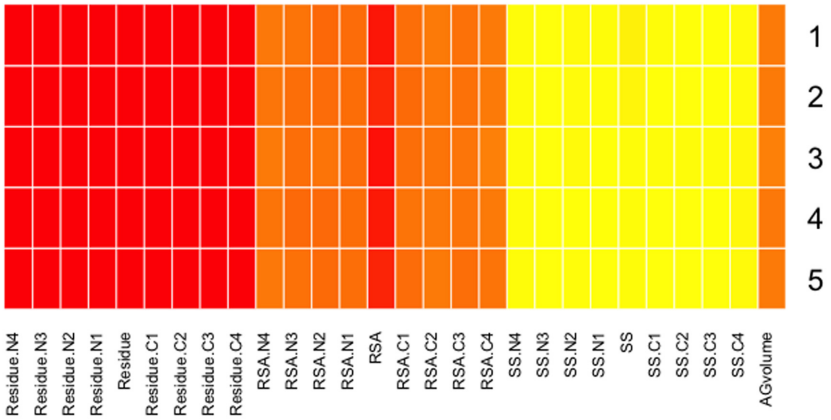

Figure 2. A heat map of each feature's impact on the 5 models generated from the 5-fold cross validation. Ranging from yellow to red, where yellow is low impact and red high impact. Each row specifies one cross-fold RF model and the columns specify the feature. The '.Nx' and '.Cx' suffix specifies positions relative to the investigated residue towards N and C-terminals, respectively.

Table 1. AUC and AUC10% for BepiPred-2.0, BepiPred-1.0, LBtope and NetsurfP from the cross-validation data

PDB ID	AUC	*P* VALUE	AUC10%	*P* VALUE
BepiPred-2.0	0.62	1	0.121	1
BepiPred-1.0	0.57	$<1 \times 10^{-6}$	0.093	0.02
LBtope	0.54	$<1 \times 10^{-6}$	0.075	$<1 \times 10^{-6}$
NETSURFP	0.60	0.01	0.07	$<1 \times 10^{-6}$

Paired t-tests with BepiPred-2.0 results were used to obtain the *P* values.

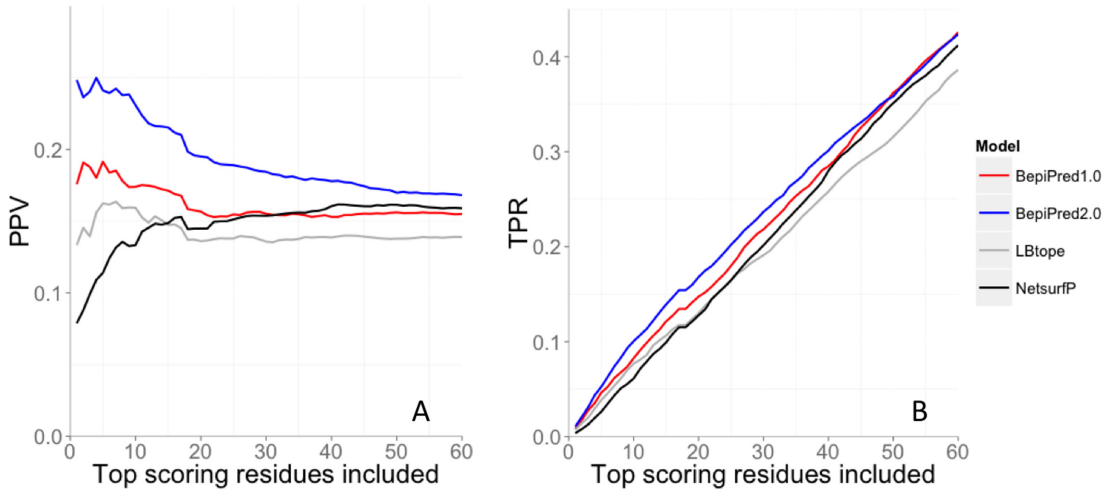

Figure 3. The average predictive positive value (PPV) (**A**) and true positive rate (TPR) (**B**) across all antigen sequences in test set using different number of top scoring residues. Four different methods are evaluated: BepiPred 1.0 (red), NetSurfP (black), LBtope (gray) and BepiPred 2.0 (blue).

Independent structural epitope benchmark

Table 2 displays the performance of BepiPred-2.0 compared to other methods, on the set of five structures in the evaluation dataset. These results confirm that BepiPred-2.0 has a higher performance than BepiPred-1.0 and LBtope. In particular, as AUC10% focuses on the highest predictions, the gap between BepiPred-1.0 and BepiPred-2.0 for this measure underlines the higher specificity of 2.0 compared to 1.0 for high scoring residues.

Evaluation of linear epitope predictive power

To perform a fair comparison, as BepiPred-1.0 was trained on linear epitopes, we tested the performance of Bepipred-1.0 and Bepipred-2.0 on a dataset consisting of 11.839 positive and 18.722 negative validated peptides obtained from the immune epitope database (IEDB, see Materials and Methods). The results of this benchmark are reported in Table 3, showing that also on this dataset BepiPred-2.0 outperforms BepiPred-1.0.

Table 2. Benchmark of the five left out antigens, comparing the performances of BepiPred-1.0, BepiPred-2.0, LBtope, BCPreds (26) and CBTope (27)

PDB ID	BEPIPRED-1.0		BEPIPRED-2.0		LBTOPE		BCPREDS		CBTOPE	
	AUC	AUC10%	AUC	AUC10%	AUC	AUC10%	AUC	AUC10%	AUC	AUC10%
4WFF	0.660	0.000	0.738	0.033	0.438	0.000	0.442	0.026	0.920	0.650
4XAK	0.739	0.183	0.657	0.104	0.471	0.000	0.465	0.000	0.570	0.091
4Z5R	0.327	0.000	0.576	0.038	0.515	0.019	0.539	0.110	0.300	0.002
5BVP	0.525	0.082	0.569	0.228	0.493	0.099	0.411	0.010	0.600	0.130
5C0N	0.596	0.000	0.473	0.000	0.397	0.113	0.26	0.000	0.560	0.080
Average	0.573	0.055	0.596	0.080	0.467	0.046	0.423	0.029	0.590	0.194
St.Dev.	0.157	0.081	0.100	0.091	0.046	0.055	0.103	0.046	0.220	0.261

Table 3. A comparison of BepiPred-1.0 and BepiPred-2.0 on experimental validated linear epitopes and non-epitope peptides. The AUC and AUC10% are calculated for each antigen and averaged. The AUC and AUC10% *P* values are obtained with paired and non-paired t-tests, respectively

	AUC	AUC10%
BepiPred-1.0	0.548	0.074
bepipred-2.0	0.574	0.080
P value	$<1 \times 10^{-6}$	$<1 \times 10^{-6}$

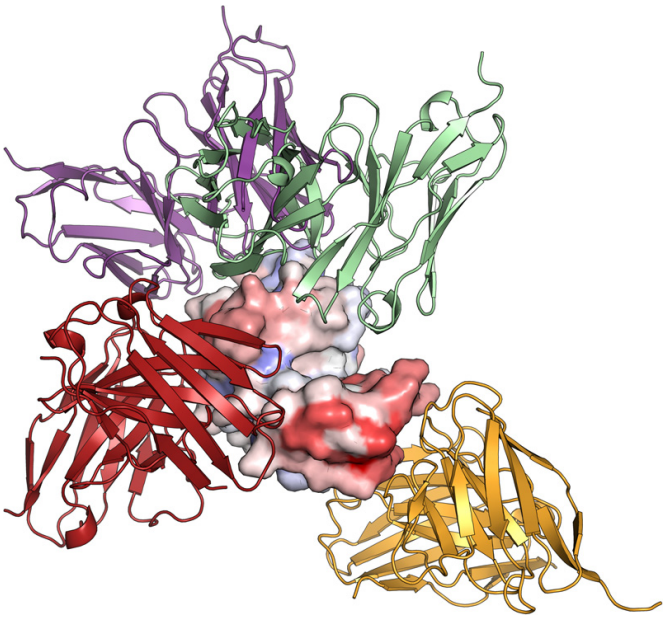

Figure 4. Lysozyme (displayed as surface, coloured from blue to red according to BepiPred-2.0 predictions) with four unique epitope regions obtained from antibodies 1BVK (purple), 1C08 (red), 1MLC (yellow) and 4TSB (green).

Case study: lysozyme epitope regions

Prototypical antibody targets, such as lysozymes have been crystallized in complex with different antibodies binding to different regions, whereas most proteins have only been crystallized with a single antibody. Using lysozyme as an example, we can see that four unique epitope regions are currently present in different solved structures (1BVK, 1C08, 1MLC, 4TSB), as shown in Figure 4, where the lysozyme is colored according to BepiPred-2.0 predictions. It is important to note that if we evaluate the performance of our method only on one of these four epitope at a time we get an average AUC of 0.593 ± 0.171 and an average AUC10% of 0.127 ± 0.211. If, on the other hand, all epitope regions are included in the evaluation, BepiPred-2.0 achieves an AUC of 0.713 and AUC10% of 0.304. This result confirms earlier findings that a possible major reason for the relative low predictive performance of B cell epitope predictions stems from the bias and incomplete annotation of currently available epitope benchmark data (9).

DISCUSSION

The BepiPred-2.0 web server provides a state-of-the-art B-cell epitope sequence-based prediction. We believe that the intuitive interface will aid researchers with limited computational knowledge to use and understand the results to their full extent. Additionally, the advanced option allows more experienced researchers to further interpret the output based on additionally predicted structural features.

Using crystallography derived structural epitope data for training and evaluation improved the performance significantly, compared to prior prediction tools trained on linear peptides tested for antibody recognition. Even when evaluated on the same type of data on which BepiPred-1.0 was trained on, BepiPred-2.0 achieved a significantly improved performance. We believe that this is a significant finding that will inform others on how B-cell epitope will be evaluated in the future. Furthermore, as illustrated in the lysozyme example, several regions can be recognised by different antibodies, raising the question on how to properly define epitopes (23,24). A possible solution to this, currently investigated by us and others (7,25), is to develop tools that can predict the epitope regions on an antigen specific for a single antibody or for an antibody library, thus increasing the specificity of the predictions. Currently epitope prediction tools can serve mostly as filters to discard regions unlikely to be epitopes from further experimental analysis, but with the increase in accuracy and specificity of these tools, we believe that they will allow for precise and targeted experiments.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We want to thank Kamilla Jensen, Vasileios Rantos, Mads Valdemar Anderson, Jens Vindahl Kringelum and Vanessa Jurtz for the useful discussions.

FUNDING

National Institutes of Health [HHSN272201200010C]. Funding for open access charge: NIH [HHSN2722012000 10C].
Conflict of interest statement. None declared.

REFERENCES

1. Kringelum,J.V., Nielsen,M., Padkjær,S.B. and Lund,O. (2013) Structural analysis of B-cell epitopes in antibody:protein complexes. *Mol. Immunol.*, **53**, 24–34.
2. Shirai,H., Prades,C., Vita,R., Marcatili,P., Popovic,B., Xu,J., Overington,J.P., Hirayama,K., Soga,S., Tsunoyama,K. *et al.* (2014) Antibody informatics for drug discovery. *Biochim. Biophys. Acta*, **1844**, 2002–2015.
3. El-Manzalawy,Y., Dobbs,D. and Honavar,V.G. (2017) In silico prediction of linear B-cell epitopes on proteins. *Methods Mol. Biol.*, **1484**, 255–264.
4. El-Manzalawy,Y. and Honavar,V. (2010) Recent advances in B-cell epitope prediction methods. *Immunome Res.*, **6**, S2.
5. Klausen,M.S., Anderson,M.V., Jespersen,M.C., Nielsen,M. and Marcatili,P. (2015) LYRA, a webserver for lymphocyte receptor structural modeling. *Nucleic Acids Res.*, **43**, W349–W355.
6. Olimpieri,P.P., Chailyan,A., Tramontano,A. and Marcatili,P. (2013) Prediction of site-specific interactions in antibody-antigen complexes: the proABC method and server. *Bioinformatics*, **29**, 2285–2291.
7. Sela-Culang,I., Benhnia,M.R.E.I., Matho,M.H., Kaever,T., Maybeno,M., Schlossman,A., Nimrod,G., Li,S., Xiang,Y., Zajonc,D. *et al.* (2014) Using a combined computational-experimental approach to predict antibody-specific B cell epitopes. *Structure*, **22**, 646–657.
8. Sircar,A. and Gray,J.J. (2010) SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput. Biol.*, **6**, e1000644.
9. Kringelum,J.V., Lundegaard,C., Lund,O. and Nielsen,M. (2012) Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput. Biol.*, **8**, e1002829.
10. Larsen,J.E.P., Lund,O. and Nielsen,M. (2006) Improved method for predicting linear B-cell epitopes. *Immunome Res.*, **2**, 2.
11. Singh,H., Ansari,H.R. and Raghava,G.P.S. (2013) Improved method for linear B-cell epitope prediction using antigen's primary sequence. *PLoS ONE*, **8**, 1–8.
12. El-Manzalawy,Y., Dobbs,D. and Honavar,V. (2008) Predicting linear B-cell epitopes using string kernels. *J. Mol. Recognit.*, **21**, 243–255.
13. El-Manzalawy,Y., Dobbs,D. and Honavar,V. (2008) Predicting flexible length linear B-cell epitopes. *Comput. Syst. Bioinformatics Conf.*, **7**, 121–132.
14. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
15. Klausen,M.S., Anderson,M.V., Jespersen,M.C., Nielsen,M. and Marcatili,P. (2015) LYRA, a webserver for lymphocyte receptor structural modeling. *Nucleic Acids Res.*, **43**, W349–W355.
16. Bigelow,C.C. (1967) On the average hydrophobicity of proteins and the relation between it and protein structure. *J.Theor. Biol.*, **16**, 187–211.
17. Sweet,R.M. and Eisenberg,D. (1983) Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.*, **171**, 479–488.
18. Radzicka,A. and Wolfenden,R. (1988) Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*, **27**, 1664–1670.
19. Petersen,B., Petersen,T., Andersen,P., Nielsen,M. and Lundegaard,C. (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.*, **9**, 51.
20. Swets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.
21. Vita,R., Overton,J.A., Greenbaum,J.A., Ponomarenko,J., Clark,J.D., Cantrell,J.R., Wheeler,D.K., Gabbard,J.L., Hix,D., Sette,A. *et al.* (2015) The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.*, **43**, D405–D412.
22. Menze,B.H., Kelm,B.M., Masuch,R., Himmelreich,U., Bachert,P., Petrich,W. and Hamprecht,F.A. (2009) A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, **10**, 213.
23. Kringelum,J.V., Lundegaard,C., Lund,O. and Nielsen,M. (2012) Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput. Biol.*, **8**, e1002829.
24. Kunik,V. and Ofran,Y. (2013) The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. *Protein Eng. Des. Sel.*, **26**, 599–609.
25. Krawczyk,K., Liu,X., Baker,T., Shi,J. and Deane,C.M. (2014) Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics*, **30**, 2288–2294.
26. El-Manzalawy,Y., Dobbs,D. and Honavar,V. (2008) Predicting linear B-cell epitopes using string kernels. *J. Mol. Recognit.*, **21**, 243–255.
27. Ansari,H.R. and Raghava,G.P. (2010) Identification of conformational B-cell Epitopes in an antigen from its primary sequence. *Immunome Res.*, **6**, 6.