Published online 31 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W513–W518***
doi:10.1093/nar/gkn254

Immune epitope database analysis resource (IEDB-AR)

Qing Zhang[1], Peng Wang[1], Yohan Kim[1], Pernille Haste-Andersen[2], John Beaver[3], Philip E. Bourne[3], Huynh-Hoa Bui[1], Soren Buus[2], Sune Frankild[2], Jason Greenbaum[1], Ole Lund[2], Claus Lundegaard[2], Morten Nielsen[2], Julia Ponomarenko[3], Alessandro Sette[1], Zhanyang Zhu[3] and Bjoern Peters[1,*]

[1]Immune Epitope Database and Analysis Resource (IEDB-AR), La Jolla Institute for Allergy and Immunology, La Jolla, CA, USA, [2]Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, DK-2800 Lyngby, Denmark and [3]San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA, USA

Received January 31, 2008; Revised April 14, 2008; Accepted April 20, 2008

ABSTRACT

We present a new release of the immune epitope database analysis resource (IEDB-AR, http://tools.immuneepitope.org), a repository of web-based tools for the prediction and analysis of immune epitopes. New functionalities have been added to most of the previously implemented tools, and a total of eight new tools were added, including two B-cell epitope prediction tools, four T-cell epitope prediction tools and two analysis tools.

INTRODUCTION

Understanding immune epitope recognition is important for the development of vaccines and diagnostics targeting infectious and autoimmune diseases, allergies and cancers. Epitopes can be defined as parts of molecules that are specifically recognized by molecules of the immune system. They are normally divided into T-cell epitopes that are presented by major histocompatibility complex (MHC) molecules and recognized by T-cell receptors (TCRs), and B-cell epitopes that are recognized by B-cell receptors (BCRs) or their soluble counterpart antibodies. The IEDB-AR covers a broad range of tools facilitating the prediction of new B- and T-cell epitopes in proteins of interest, and tools for the analysis of epitope sets collected from within the IEDB (1) or submitted by the user. This article presents an overview of the tool capabilities provided in the IEDB-AR with a focus on previously unpublished additions.

The overall goal of the IEDB-AR is to provide access to well-documented and tested epitope-related tools through a common style of web interface. This unified interface allows users to easily make direct comparisons between and among various prediction methods. Epitope and protein sequences can be submitted to each tool directly, or by selecting epitopes retrieved through a query of the IEDB. Help pages providing instructions and examples of input data accompany each tool.

THE WEB RESOURCE

Table 1 lists all the tools currently available from IEDB-AR.

T-cell epitope prediction

T-cell epitopes can be classified as either class I or class II epitopes. Class I epitopes are typically peptides presented by MHC class I (MHC-I) molecule and recognized by cytotoxic T lymphocytes (CTLs) that can kill infected cells. Class II epitopes are generally peptides that are presented by MHC class II (MHC-II) molecules and recognized by helper T lymphocytes (HTLs). Presentation of peptides-MHC-I complexes to T lymphocytes is a multistep process. In the IEDB-AR there are both tools that can predict the individual steps—such as cleavage of the polypeptide chain by the proteasome, transport of peptides into the endoplasmatic reticulum (ER) by the transporter associated with processing (TAP) and binding to MHC molecules—as well as methods that integrate these steps into one prediction.

All T-cell epitope prediction methods implemented in the IEDB-AR take protein sequences encoded by their single letter symbols as an input. In addition, the user can specify the MHC molecule for which they want to make predictions, as different MHC molecules have different binding specificities. Each input protein sequence is split

*To whom correspondence should be addressed. Tel: +1 858 483 1922; Fax: +1 858 752 6987; Email: bpeters@liai.org

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

Table 1. Epitope related tools available at IEDB-AR

	Category	Tool description
T-cell epitope prediction	Peptide binding to MHC class I molecules[a]	Determines peptide's ability to bind to a specific MHC class I molecule
	Peptide binding to MHC class II molecules[a]	Provides the Sturniolo, ARB, SMM-align and a consensus approach to predict MHC class II binding peptides
	Proteasomal cleavage/TAP transport/MHC class I and combined predictor[a]	Combines predictors of proteasomal processing, TAP transport and MHC binding to produce an overall score for each peptide's potential of being an epitope. Two implementations are provided, one based on matrices and one on neural networks (NetChop and NetCTL)
B-cell epitope prediction	Prediction of B-cell epitopes from protein sequences[a]	Provides a common user interface to access a collection of previously published B cell prediction tools based on amino acid scales, including Bepipred that incorporates similar scales into a position-specific scoring matrix
	Prediction of B-cell epitopes from protein structures[a]	Predicts discontinuous epitopes based on amino acid statistics, spatial information and surface accessibility
Epitope analysis	Population coverage	Calculates the fraction of individuals predicted to respond to a given set of epitopes with known MHC restrictions
	Epitope conservancy analysis	Calculates degree of conservancy of an epitope within a given protein sequence set
	Epitope cluster analysis[a]	Groups epitopes into clusters based on sequence identity
	Homology mapping[a]	Maps linear epitopes to 3D structures of proteins

[a]Tools that have been implemented or updated since the previous release of the IEDB-AR.

into all possible peptides meeting the length preference of the selected MHC molecule. The output contains this list of peptides, each with a predicted score that reflects its likelihood of being a T-cell epitope. Depending on the chosen tool, this score reflects the disposition of a peptide to bind MHC, be processed from its parent protein or both.

MHC class I binding predictions

Three MHC class I peptide binding prediction methods are provided through the IEDB-AR: artificial neural network (ANN), stabilized matrix method (SMM) and average relative binding (ARB). Two of the methods, SMM and ARB, model binding specificity of an MHC molecule using position-specific scoring matrices (PSSM) (2–4). ANN, on the other hand, uses neural networks with diverse sequence encoding schemes to model-binding specificity (5).

The three prediction methods have been previously trained and evaluated (6) using 5-fold cross-validation. Since then, a new set of peptide-binding data for MHC class I molecules has become available, motivating the retraining and testing of all three MHC-I peptide prediction methods. For all MHC alleles with additional data available, new algorithms were trained that should exhibit improved prediction quality compared to the previous version, as the prediction accuracy is largely dependent on the amount of training data available. Also, the number of different MHC molecules for which predictions can be made was enhanced by eight new additions, including alleles from humans, mice, chimpanzees, macaques and gorillas. To accommodate historic comparisons, the IEDB-AR website now hosts two versions of the prediction methods, the previously published one and the updated implementations.

The availability of the new binding dataset gives us an opportunity to re-evaluate the 5-fold cross-validation used in ref. (6) to measure performance of these methods. Specifically, we made binding predictions for all peptides in the newly available data (called the 'Independent Data Set' from here on). The ability of each method to discriminate high affinity binders ($IC50 < 500\,nM$) from those with lower affinity was measured using the Area under Receiver-Operating-Characteristic curves (AUC) values (7). Figure 1 shows AUC values published in ref. (6) compared to those attained with the Independent Data Set. This clearly indicates that the two performance evaluation methods give similar results overall. Importantly, there is no evidence for the previous cross-validation overestimating the performance of the prediction methods: out of 127 cases total, the independent dataset had the higher AUC in 64 cases, while the cross-validated results were higher in the remaining 63 cases. Confirming our earlier analysis (6), SMM and ANN perform better than ARB using a different performance evaluation measure. While both ARB and SMM use PSSM to model-binding specificity, they arrive at these matrices in a different manner. For each element in a PSSM, ARB bases the matrix entry on the average affinity of peptides sharing the corresponding residue. In contrast, all elements of PSSM for SMM are calculated simultaneously using a least squares fitting approach. This apparently leads to a better representation of binding specificity.

MHC class I processing predictions

Additional events apart from peptide–MHC binding determine if a peptide is likely to be a T-cell epitope. These include proteasomal cleavage of the source protein and transport of peptides into the ER from the cytosol by the TAP transporter. The IEDB-AR provides an implementation of such predictions described in ref. (8), which can be combined with the updated MHC class I binding predictions described above. In the new release, alternative approaches were implemented: NetChop (9) uses novel sequence encoding methods to predict proteasomal cleavage, and NetCTL (10,11) combines these with predictions of MHC binding and TAP transport. Both tools were implemented with a common user interface,

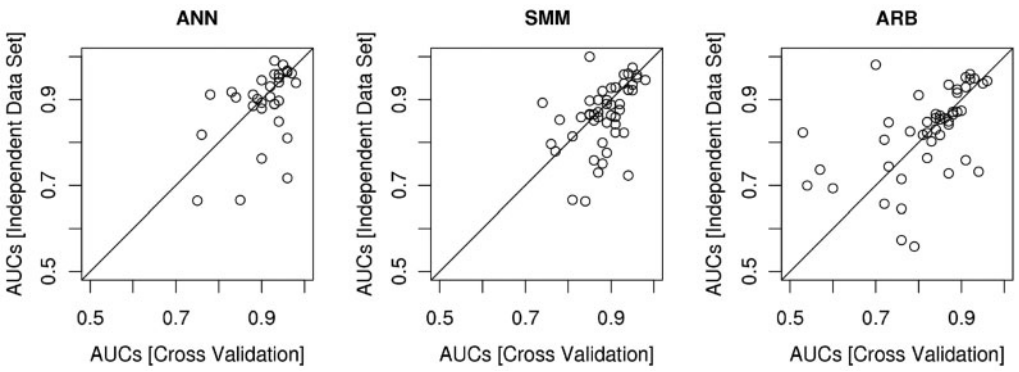


Figure 1. Performances of prediction methods measured using either cross-validation or independent data set. Each data point represents an area under a receiver-operating-characteristic curve (AUC) for an MHC molecule (and its preferred peptide length). AUCs generated using cross-validation were taken from ref. (6), while those using independent data set are presented here for the first time. For independent data set, prediction methods trained on the old data used in ref. (6) were tested on the new, and thus independent, peptide-binding data that recently became available. In the case of cross-validation, however, an AUC value is an average of AUCs generated by training a prediction method on 4/5 of a data set and testing on the remaining one. Each data set for an MHC molecule had at least 50 binding affinity measurements.

where a user can select to use either NetChop or NetCTL methods.

Compared to the highly selective peptide binding to MHC molecules, proteasomal cleavage and TAP transport have little influence on epitope generation. This makes predictions made by these tools less informative, but can be helpful to explain why some peptides that bind MHC molecules well are not naturally presented.

MHC class II binding predictions

Contrary to the closed binding groove of MHC class I molecules, the MHC class II groove that binds peptides is open at both ends and can thus accommodate peptides of variable length (12). This flexibility in binding makes MHC class II binding prediction substantially more difficult as compared to MHC class I binding prediction (6,13,14), leading to an overall lower prediction accuracy.

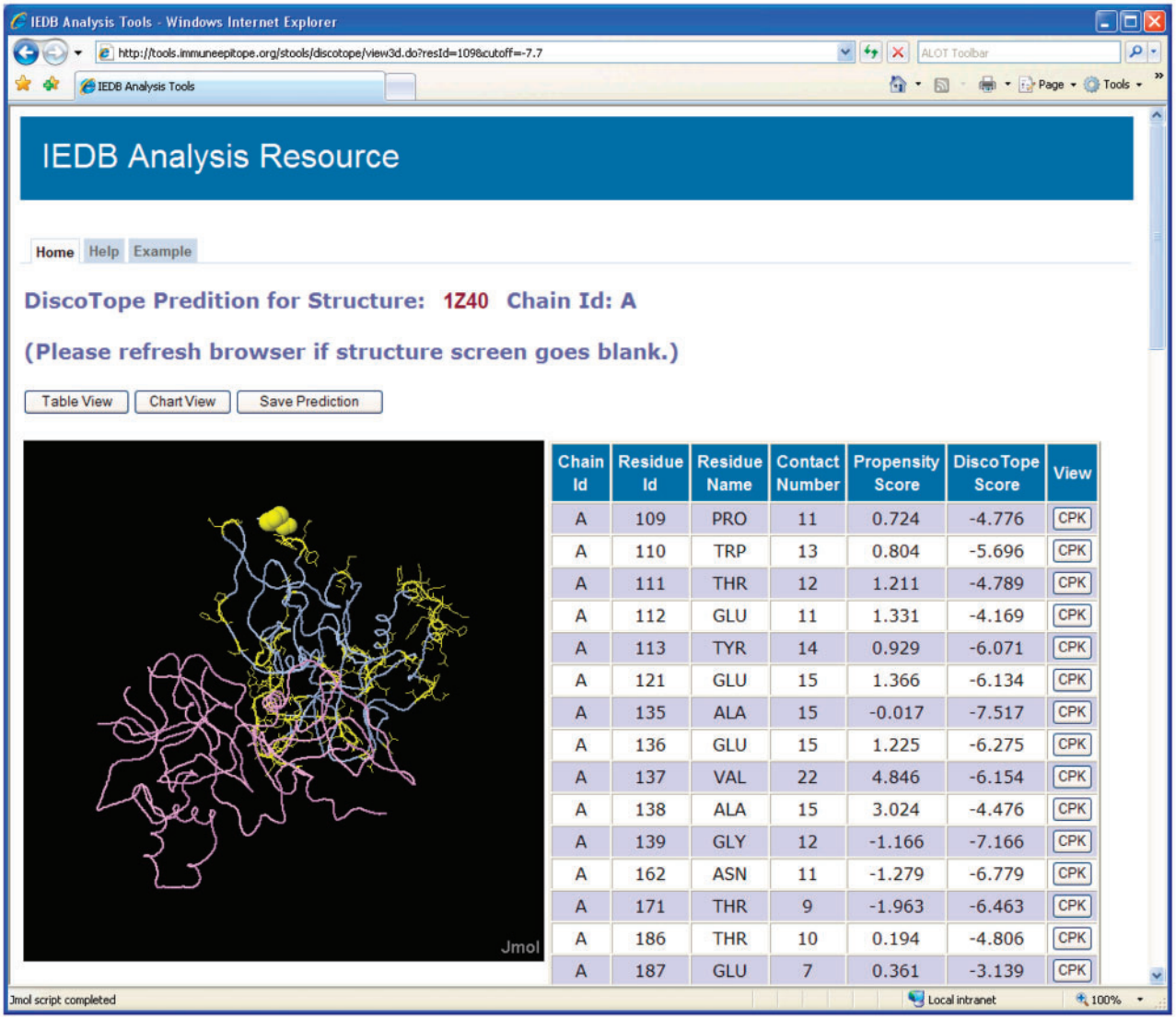
The MHC class II binding prediction service hosted at the IEDB-AR provides four prediction methods: ARB, SMM_align, Sturniolo's method (which is also the basis of TEPITOPE), and a consensus approach. ARB was the only method implemented in the last release, and is based on ARB matrices (2). SMM-align (15) is a matrix-based method with extensions incorporating flanking residues outside of binding grooves. Both methods can predict the IC50 values of peptides. The Sturniolo method implemented at IEDB-AR is a direct implementation of the matrices published in Sturniolo's original article (16), which provides scores on a nominal scale rather than IC50 values.

The consensus method is developed to combine the results of those three methods. We first scanned a random set of Swiss-Prot proteins and generated scores for 2 000 000 random peptides. The set of scores were then used as a reference to rank new predictions. The consensus approach then uses the median rank of the three methods as the final prediction score.

In order to independently evaluate the performance of the four methods, we generated an independent dataset by combining MHC class II binding data from several recent publications (17–20). The average AUCs for the four methods tested on the independent dataset are as follows: ARB-0.75, SMM-align-0.74, Sturniolo-0.72 and consensus-0.77. In addition to this new dataset, we have previously evaluated the performance of those methods with ~20 000 peptide–MHC binding affinities generated by the Sette group (manuscript submitted for publication). The four methods implemented at the IEDB-AR are the top performing ones with average AUC values ranging from 0.71 to 0.76, which are consistent with the new independent evaluation.

B-cell epitope prediction

B-cell epitopes (or, antibody epitopes) can be defined as regions on molecules that are specifically bound by B-cell receptors (antibodies). They can include native protein structures, linear peptides, carbohydrates and other biological macromolecules. The majority of short linear peptides have the potential to be B-cell epitopes (21,22). This makes their prediction much harder than for T-cell epitopes, which require binding to the highly selective MHC molecules. The IEDB-AR includes seven methods for B-cell epitope prediction; five of them are previously published approaches for which no web implementation had been available. Two recently developed tools, BepiPred and DiscoTope, were implemented in the new release of IEDB-AR. BepiPred combines a PSSM with a propensity scale method to predict linear epitopes (23). DiscoTope is a method for discontinuous epitope prediction that uses protein 3D structural data as input (24). It is based on amino acid statistics, spatial information and surface accessibility for a set of discontinuous epitopes determined by X-ray crystallography of antibody/antigen protein complexes. It exhibits better performance for predicting residues of discontinuous epitopes than methods based solely on sequence information (24). Both BepiPred and DiscoTope were implemented on the IEDB-AR website with new features that include structural visualization of predicted discontinuous epitopes on protein 3D structures (Figure 2), based on the Jmol



Chain Id	Residue Id	Residue Name	Contact Number	Propensity Score	DiscoTope Score	View
A	109	PRO	11	0.724	-4.776	CPK
A	110	TRP	13	0.804	-5.696	CPK
A	111	THR	12	1.211	-4.789	CPK
A	112	GLU	11	1.331	-4.169	CPK
A	113	TYR	14	0.929	-6.071	CPK
A	121	GLU	15	1.366	-6.134	CPK
A	135	ALA	15	-0.017	-7.517	CPK
A	136	GLU	15	1.225	-6.275	CPK
A	137	VAL	22	4.846	-6.154	CPK
A	138	ALA	15	3.024	-4.476	CPK
A	139	GLY	12	-1.166	-7.166	CPK
A	162	ASN	11	-1.279	-6.779	CPK
A	171	THR	9	-1.963	-6.463	CPK
A	186	THR	10	0.194	-4.806	CPK
A	187	GLU	7	0.361	-3.139	CPK

Figure 2. Screenshot of the '3D View' page for the DiscoTope tool on IEDB-AR. The input for this prediction is the PDB structure file of AMA1 from *Plasmodium falciparum* (28) (PDB ID 1Z40). On the left is the 3D display of this structure, with all predicted epitope residues for chain A highlighted in yellow. The table on the right lists details of these predicted residues. Clicking on the 'CPK' button in the table highlights that specific residue in the 3D display, as seen on the top.

package (Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/).

Epitope analysis

The IEDB-AR provides five analytical tools for: (i) calculating the epitope population coverage; (ii) assessing the degree of conservancy of an epitope; (iii) visualization and analysis of 3D structures of molecules containing epitopes; (iv) clustering of epitope sequences and (v) mapping of epitopes onto 3D protein structures. The last two tools have not been published before.

Clustering of epitope sequences

In evaluating groups of peptide epitopes for vaccine applications, it is important to remove redundant sequences or sequences with high similarity. The clustering tool was designed for this purpose. It groups epitopes into clusters based on sequence identity. A cluster is defined as a group of sequences, which have a sequence similarity greater than the minimum sequence identity threshold specified by a user.

Analysis of conserved epitopes

In an epitope-based vaccine setting, the use of conserved epitopes would be expected to provide broader protection across multiple strains, or even species, than epitopes derived from highly variable genome regions (25). The conservancy tool was designed to analyze the variability or conservation of epitopes. This tool calculates the degree of conservancy of a set of peptide/epitope sequences within a given set of protein sequences. The degree of

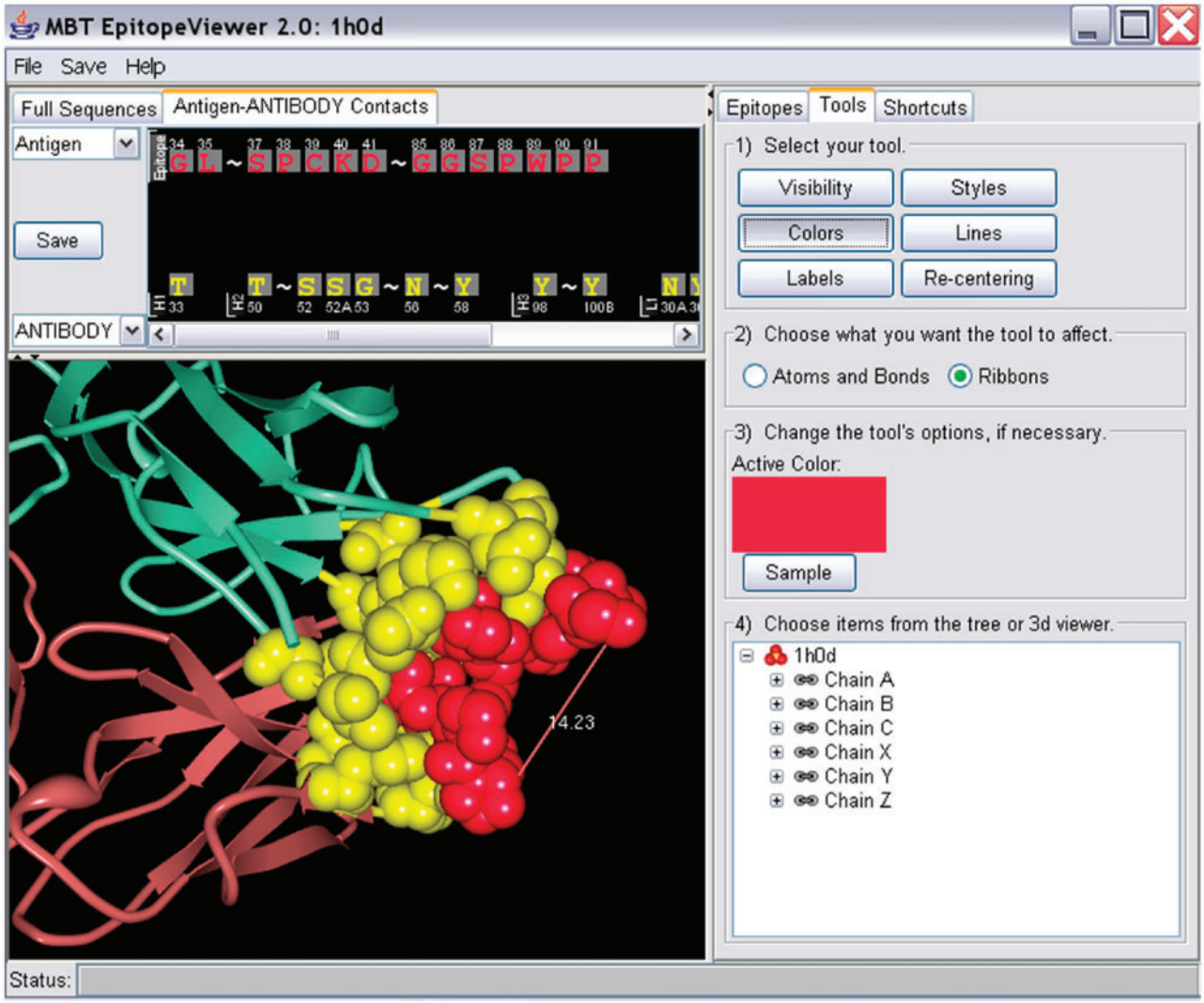


Figure 3. Screenshot of the EpitopeViewer 2.0 showing the 26-2F epitope of human angiogenin [PDB ID: 1H0D, chain C] in red (the rest of the antigen is invisible) and paratope in yellow in CPK representation. The distance between the atom N of residue 34 Gly and atom Cα of residue 91 Pro of the epitope is 14.23 Å (red line).

conservation is defined as the fraction of protein sequences containing the peptide sequence at a given identity level.

Homology mapping tool

The knowledge of a T-cell epitope location in the protein 3D structure may elucidate the mechanisms of the epitope processing, MHC binding and recognition by TCRs. Likewise, the mapping of B-cell epitopes to protein structure is useful for assessing epitope structural features recognized by antibodies. The IEDB-AR homology mapping tool provides the mapping of a linear epitope from a source protein to the proteins with known 3D structures by sequence similarity search of the epitope source sequence against protein sequences in the Protein Data Bank (PDB) (26). The tool's output displays the alignments between the epitope source sequence and homologous sequences from the PDB and, using the EpitopeViewer application (27), allows visualization and analysis of the 3D structure of the homologous protein with the epitope mapped to it.

VISUALIZATION AND ANALYSIS OF 3D STRUCTURES OF MOLECULES CONTAINING EPITOPES

The 2.0 version of the EpitopeViewer (24) has been enhanced by adding several new visualization features, permitting a more detailed 3D structural analysis of the epitope/antigen and immune receptor molecules, and the interactions between them. Specifically, the following new features have been added: hiding/showing selected residues, chains or atoms; centering the structure on the selected atom, residue or chain; displaying the distance between any two selected atoms/residues; representing atoms, bonds and ribbons in different styles; coloring residues by types, secondary structures

and hydrophobicity; and coloring atoms by elements, B-factors and colors of corresponding residues. In addition to the image export file formats available in the 1.0 version are options to save images of the 3D structure and interaction plots. The 2.0 release expands on this functionality by allowing to save both the curated and on-the-fly calculated atomic intermolecular contacts between the epitope/antigen and the receptor (Figure 3).

SUMMARY

We presented in this article a new release of IEDB-AR. It provides multiple tools to predict MHC class I and MHC class II restricted T-cell epitopes, linear and discontinuous B-cell epitopes and tools to analyze epitope sequences and structures. The updates in this new release take advantage of the additional data available in the IEDB to create improved prediction methods, and implement feature requests from the user community.

ACKNOWLEDGEMENTS

The IEDB is funded by NIH contract HHSN2662004 0006C. Funding to pay the Open Access publication charges for this article was provided by NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Peters,B., Sidney,J., Bourne,P., Bui,H.H., Buus,S., Doh,G., Fleri,W., Kronenberg,M., Kubo,R., Lund,O. *et al.* (2005) The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol.*, **3**, e91.
2. Bui,H.-H., Sidney,J., Peters,B., Sathiamurthy,M., Sinichi,A., Purton,K.-A., Mothé,B.R., Chisari,F.V., Watkins,D.I. and Sette,A. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*, **57**, 304–314.
3. Peters,B. and Sette,A. (2005) Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinform.*, **6**, 132.
4. Peters,B., Tong,W., Sidney,J., Sette,A. and Weng,Z. (2003) Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules. *Bioinformatics*, **19**, 1765–1772.
5. Nielsen,M., Lundegaard,C., Worning,P., Lauemoller,S.L., Lamberth,K., Buus,S., Brunak,S. and Lund,O. (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.*, **12**, 1007–1017.
6. Peters,B., Bui,H.-H., Frankild,S., Nielsen,M., Lundegaard,C., Kostem,E., Basch,D., Lamberth,K., Harndahl,M., Fleri,W. *et al.* (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput. Biol.*, **2**, e65.
7. Swets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.
8. Tenzer,S., Peters,B., Bulik,S., Schoor,O., Lemmel,C., Schatz,M.M., Kloetzel,P.M., Rammensee,H.G., Schild,H. and Holzhütter,H.G. (2005) Modeling the MHC class I pathway by combining predictions of proteasomal cleavage,TAP transport and MHC class I binding. *Cell. Mol. Life Sci.*, **62**, 1025–1037.
9. Nielsen,M., Lundegaard,C., Lund,O. and Keşmir,C. (2005) The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*, **57**, 33–41.
10. Larsen,M., Lundegaard,C., Lamberth,K., Buus,S., Lund,O. and Nielsen,M. (2007) Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinform.*, **8**, 424.
11. Larsen, Mette,V., Lundegaard,C., Lamberth,K., Buus,S., Brunak,S., Lund,O. and Nielsen,M. (2005) An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur. J. Immunol.*, **35**, 2295–2303.
12. Jones,E.Y., Fugger,L., Strominger,J.L. and Siebold,C. (2006) MHC class II proteins and disease: a structural perspective. *Nat. Rev. Immunol.*, **6**, 271–282.
13. Gowthaman,U. and Agrewala,J.N. (2008) In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion. *J. Proteome Res.*, **7**, 154–163.
14. Lin,H.H., Ray,S., Tongchusak,S., Reinherz,E.L. and Brusic,V. (2008) Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC Immunol.*, **9**, 8.
15. Nielsen,M., Lundegaard,C. and Lund,O. (2007) Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinform.*, **8**, 238.
16. Sturniolo,T., Bono,E., Ding,J., Raddrizzani,L., Tuereci,O., Sahin,U., Braxenthaler,M., Gallazzi,F., Protti,M.P., Sinigaglia,F. *et al.* (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.*, **17**, 555–561.
17. Depil,S., Moralès,O., Castelli,F.A., Delhem,N., François,V., Georges,B., Dufossé,F., Morschhauser,F., Hammer,J., Maillère,B. *et al.* (2007) Determination of a HLA II promiscuous peptide cocktail as potential vaccine against EBV latency II malignancies. *J. Immunother.*, **30**, 215–226.
18. Immonen,A., Kinnunen,T., Sirven,P., Taivainen,A., Houitte,D., Perasaari,J., Narvanen,A., Saarelainen,S., Rytkonen-Nissinen,M., Maillere,B. *et al.* (2007) The major horse allergen Equ c 1 contains one immunodominant region of T cell epitopes. *Clin. Exp. Allergy*, **37**, 939–947.
19. Malmassari,S.L., Deng,Q., Fontaine,H., Houitte,D., Rimlinger,F., Thiers,V., Maillere,B., Pol,S. and Michel,M.-L. (2007) Impact of hepatitis B virus basic core promoter mutations on T cell response to an immunodominant HBx-derived epitope. *Hepatology*, **45**, 1199–1209.
20. Stone,S.P., Cooper,B.S., Kibbler,C.C., Cookson,B.D., Roberts,J.A., Medley,G.F., Duckworth,G., Lai,R., Ebrahim,S., Brown,E.M. *et al.* (2007) The ORION statement: guidelines for transparent reporting of outbreak reports and intervention studies of nosocomial infection. *J. Antimicrob. Chemother.*, **59**, 833–840.
21. Blythe,M.J. and Flower,D.R. (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci.*, **14**, 246–248.
22. Greenbaum,J.A., Andersen,P.H., Blythe,M., Bui,H.-H., Cachau,R.E., Crowe,J., Davies,M., Kolaskar,A.S., Lund,O., Morrison,S. *et al.* (2007) Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J. Mol. Recognit.*, **20**, 75–82.
23. Larsen,J., Lund,O. and Nielsen,M. (2006) Improved method for predicting linear B-cell epitopes. *Immunome Res.*, **2**, 2.
24. Andersen,P.H., Nielsen,M. and Lund,O. (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.*, **15**, 2558–2567.
25. Bui,H.-H., Sidney,J., Li,W., Fusseder,N. and Sette,A. (2007) Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. *BMC Bioinform.*, **8**, 361.
26. 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.
27. Beaver,J., Bourne,P. and Ponomarenko,J. (2007) EpitopeViewer: a Java application for the visualization and analysis of immune epitopes in the immune epitope database and analysis resource (IEDB). *Immunome Res.*, **3**, 3.
28. Bai,T., Becker,M., Gupta,A., Strike,P., Murphy,V.J., Anders,R.F. and Batchelor,A.H. (2005) Structure of AMA1 from Plasmodium falciparum reveals a clustering of polymorphisms that surround a conserved hydrophobic pocket. *Proc. Natl Acad. Sci. USA*, **102**, 12736–12741.