Nucleic Acids Research Advance Access published April 24, 2008

Nucleic Acids Research, 2008, 1–4
doi:10.1093/nar/gkn229

EpiToolKit—a web server for computational immunomics

Magdalena Feldhahn[1,*], Philipp Thiel[1], Mathias M. Schuler[2], Nina Hillen[2], Stefan Stevanović[2], Hans-Georg Rammensee[2] and Oliver Kohlbacher[1]

[1]Simulation of Biological Systems, Center for Bioinformatics and [2]Department of Immunology, Institute for Cell Biology, Eberhard Karls University Tübingen, Germany

Received January 31, 2008; Revised April 4, 2008; Accepted April 11, 2008

ABSTRACT

Predicting the T-cell-mediated immune response is an important task in vaccine design and thus one of the key problems in computational immunomics. Various methods have been developed during the last decade and are available online. We present EpiToolKit, a web server that has been specifically designed to offer a problem-solving environment for computational immunomics. EpiToolKit offers a variety of different prediction methods for major histocompatibility complex class I and II ligands as well as minor histocompatibility antigens. These predictions are embedded in a user-friendly interface allowing refining, editing and constraining the searches conveniently. We illustrate the value of the approach with a set of novel tumor-associated peptides. EpiToolKit is available online at www.epitoolkit.org.

INTRODUCTION

Prediction of T-cell epitopes is a key problem in Immunoinformatics (1). Identifying peptides with high binding affinity to major histocompatibility complex (MHC) molecules is generally considered the best way to predict epitopes (2). Many different methods and tools for peptide–MHC binding have been developed in the recent past, most of them are accessible online (3–12). Most of the individual web servers, however, were designed to quickly offer online access to a newly developed method with usability not being the major concern. Nonexpert users may be overwhelmed by the variety of layouts, formats and options.

The main focus in the development of the EpiToolKit web server was usability. The purpose of EpiToolKit is to facilitate immunological research by providing a consistent and user-friendly interface for different methods from computational immunomics. The prediction pipeline is organized in four main steps: sequence input, sequence information, model selection and display of prediction results. Each page contains hints and short comments to guide the user through the pipeline. In addition, a detailed help and documentation is available through direct links on every page.

The service provides most of the commonly accepted prediction methods and allows their simultaneous application. Thereby, prediction results can be compared without the need to access the individual web-based services separately. The combination of different prediction methods also increases the number of available allelic models.

In addition to epitope prediction, EpiToolKit offers the functionality to examine the influence of sequence polymorphisms or mutations on potential T-cell epitopes. This feature is useful for the identification of minor histocompatibility antigens (mHags) and for the development of peptide-based vaccines against highly variable pathogens such as HCV and HIV.

EpiToolKit is based on a flexible and modular framework for predictions related to epitope prediction (13). The framework and EpiToolKit can easily be extended with new methods—e.g. new methods for MHC binding, or for the prediction of the epitope processing pathway.

WEB INTERFACE

The web interface is divided into two parts, the epitope prediction and the prediction on polymorphic proteins (SNEPv2). Both predictions use the same layout and are based on a common pipeline. To facilitate the use of EpiToolKit, the prediction pipeline has been broken down into four intuitive steps:

(1) Sequence input. Sequences can be retrieved from the most important resources for protein sequences, namely Swiss-Prot (14) and NCBI RefSeq (15). In case of polymorphic prediction, all reported

*To whom correspondence should be addressed. Tel: +49 7071 29 70460; Fax: +49 7071 29 5152; Email: feldhahn@informatik.uni-tuebingen.de

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

polymorphisms for the specified sequences are provided automatically. Alternatively, the user can paste sequences directly or upload a FASTA file. In case of user-defined polymorphic or mutated sequences, the changes can be specified in the FASTA header.

(2) Sequence information. The purpose of this step is to present the query results for the requested sequences. For all returned sequences, additional information such as sequence length, GeneID or RefSeq accession is displayed. Furthermore, SNEPv2 returns all annotated polymorphisms for each sequence. The sequences as well as the polymorphisms can be selected or deselected individually for further processing.

(3) Allele selection. In this step, the user can select allelic models for prediction. The model selection is organized in an expandable model/allele tree, sorted by allele name. The tree can be customized to contain only models for selected peptide lengths, prediction methods or alleles in the advanced options section.

(4) Prediction results. Results are displayed as tables. For epitope prediction, a single table is created for every peptide length. In case of polymorphic predictions using SNEPv2 for each polymorphism a separate table is created. Different methods for discrimination between predicted binders and nonbinders are available. Filter and display options can be changed in the advanced options section. Additionally, the prediction results can be exported in CSV (comma separated values) or XLS (Microsoft Excel) format.

DATASETS

EpiToolKit provides access to the Swiss-Prot database (14) and the NCBI RefSeq database (15). To improve the reliability of the service and to accelerate access time, both databases are kept as local copies. They are updated monthly. Current release information is displayed on the sequence input site.

Polymorphism information is provided from two different sources depending on the database used for sequence retrieval. For RefSeq sequences, polymorphisms are obtained from dbSNP (16). SNP entries contain links to protein sequences if the corresponding variation has been mapped onto coding regions. For requested RefSeq sequences, nonsynonymous coding SNPs are yielded by EpiToolKit. Furthermore, to limit the search to relevant polymorphisms the set of reported polymorphisms can be restricted to a user-defined heterozygosity range. Polymorphisms for Swiss-Prot sequences are extracted directly from the Swiss-Prot entries. Polymorphism data from dbSNP as well as from Swiss-Prot are also kept in local databases.

IMPLEMENTATION

This section briefly describes the prediction methods included in EpiToolKit and the implementation of the web server.

Prediction on polymorphic proteins

In addition to epitope prediction EpiToolKit provides the possibility to perform predictions on proteins containing polymorphisms. This tool is called SNEPv2. It is a new version of the SNP-derived Epitope Prediction program (SNEP) developed by Schuler *et al.* (11). One major improvement of SNEPv2 is the use of a second resource to retrieve sequence polymorphisms [dbSNP (16)]. Additionally, due to the embedding into EpiToolKit, SNEPv2 offers a comfortable user interface. The latter enables users to perform predictions for multiple polymorphic sequences from several sources, to apply a variety of prediction methods and to use different filtering methods to restrict the result set.

SNEPv2 creates a set of polymorphic peptides for each reported sequence polymorphism. These peptide sets are generated by extracting the peptides around the polymorphic position using a sliding window of specified length and subsequently mutating these peptides to all observed variants. The prediction results are displayed separately for each peptide set. On the result page, the user can additionally switch between the epitope prediction results for the source protein and the polymorphic predictions.

Epitope prediction methods

A major challenge in epitope prediction is that the manifold MHC alleles display a wide spectrum of binding specificities. Prediction methods must therefore provide models for different alleles. Not all prediction methods have a model for every allele or peptide length.

Five different methods for the prediction of peptides binding to MHC class I and two methods for MHC class II binding are currently available in EpiToolKit. These methods are described briefly in the following. For details on the prediction methods refer to the original publications. A table of all available allelic models and methods can be found in Supplementary Material 1. The description of the available methods in the documentation of EpiToolKit will be updated when new methods are included into EpiToolKit.

- SYFPEITHI (3) is based on position-specific scoring matrices. The matrices are manually generated based on expert knowledge and the occurrence of amino acids in naturally processed MHC ligands from the SYFPEITHI database.
- BIMAS/HLA_BIND (4) was developed at the BioInformatics and Molecular Analysis Section (BIMAS) at the NIH. The prediction method uses position specific scoring matrices that are derived from experimentally determined relative binding affinities. Dissociation rates of peptide:MHC:β_2-microglobulin complexes are used to measure binding affinities relative to a reference peptide. The original values in the matrices are log-transformed to obtain an additive scoring scheme.
- SVMHC (5) uses support vector machine classification to predict MHC-binding peptides. The method is trained on known MHC-binding peptides from

the SYFPEITHI database and randomly generated nonbinders.
- Epidemix (13) is based on position-specific scoring matrices. The matrices are statistically computed based on the positive training set of SVMHC. Sequence weighting and pseudo-count correction are applied to obtain the frequencies used to generate the matrices.
- UniTope (Feldhahn, Toussaint, Ziehm, and Kohlbacher, manuscript in preparation) is a support vector classification method recently developed in our group. UniTope combines structural and sequence information in a machine-learning framework. Based on a decomposition of the MHC-binding groove into distinct pockets, the correlation between the physicochemical properties of these pockets and peptide binding is learned. The allele encoding uses pocket profiles derived from crystal structures of peptide:MHC complexes. The peptides are also encoded using physicochemical properties. This enables binding prediction even for alleles where no experimental binding data are available.
- Hammer (17) is based on position-specific scoring matrices and predicts binding peptides for MHC class II. The virtual matrices were published by Sturniolo *et al.* (17) and are used by the TEPITOPE software.
- MHCIIMulti (18) is a new method based on multiple instance learning and support vector classification, which can be used to predict peptide binding for MHC class II. Therefore, a new kernel is introduced, which also takes similarities between alleles into account. The method can even be used to predict binding peptides for alleles without available binding data.

Filtering methods

Most prediction methods predict a score that represents the binding affinity of a peptide to an MHC allele. In order to discriminate binding peptides from nonbinding peptides, a threshold can be used to separate binders from nonbinders.

EpiToolKit uses thresholds to filter the results for potential epitopes. Only peptides classified as binders are displayed if a filter is activated. The following filter options are available:

(1) No filtering. All predicted scores are displayed. The peptides are not classified as binders or nonbinders.
(2) Filtering using *halfmax*-scores. For all matrix-based methods, the halfmax-score is defined as half of the maximal value obtainable from the matrix. The halfmax-scores are used as thresholds. For SVM-based predictions halfmax-scores are not defined. The 2%-thresholds are used instead (see 'Filtering by percentage' subsequently). The halfmax filter is the default filter method in EpiToolKit.
(3) Filtering by percentage. The thresholds are determined based on a background score distribution that was computed on a large set of peptides derived from natural proteins. The thresholds can be interpreted as follows: using as example a 2%-threshold 2% of the peptides used to compute the background distribution would be classified as binders. The advantage of this filtering method is that—in contrast to the halfmax-filtering—thresholds for different allelic models are comparable. The input format for the percentage thresholds is a float in the interval of [0, 1], e.g. 0.02 for a 2%-threshold.

Note that the current version of UniTope performs a binary classification. For consistency, all filter methods are also available for UniTope predictions but do not have an influence on the classification—peptides with score 1 are always classified as binders, whereas peptides with score 0 are always classified as nonbinders.

Web server

EpiToolKit utilizes the open source content management framework Plone (http://plone.org) based on the application server Zope (http://www.zope.org). Dynamic HTML pages provide forms for user input and prediction results. The pages use CSS and JavaScript and the service was tested for compatibility with the two most widely used web browsers, namely Mozilla Firefox (version 2.0) and MS Internet Explorer (version 7). Input data is pre-processed by Python scripts for data validation. Entered user data are temporarily stored on the server and automatically deleted after the corresponding session has expired. The program logic for data processing and for performing epitope prediction (13) is written in Python. Several tasks of the program logic use the Biopython package (http://biopython.org).

VALIDATION AND APPLICATION

To validate the correct function of EpiToolKit, the prediction results for 1000 randomly chosen proteins were compared to the results of the original methods/web servers. In all cases EpiToolKit produced the same results as the original methods.

In addition, a set of 27 novel HLA-A*0201 ligands characterized from tumor cells and in part derived from tumor antigens were used for validation. These peptides are listed in Supplementary Material 2. All available allelic models of appropriate length were used for prediction and default filtering (halfmax) was applied. Eighteen out of 20 nonameric peptides and six out of seven decameric peptides were correctly classified as binders.

To validate SNEPv2, we tried to retrospectively identify the mHags reported by Goulmy (19). The paper gives an introduction to the clinical relevance of mHags and contains a list of all then known human mHags. We refer to the single mHags with the names used in (19). The dataset is available in Supplementary Material 3.

Eight out of nine autosomally encoded mHags reported could directly be found using SNEPv2. The HLA-A*2902 restricted mHag UGT2B17 results from a gene-deletion and therefore no allelic counterpart is available, so epitope prediction was used. Since no decameric model was available for HLA-A*2902, the nonameric UniTope model was used to score the two nonameric substrings. Both substrings were predicted to bind to HLA-A*2902.

The Y-chromosomal encoded mHags do not have allelic counterparts, except for the HLA-A*01 restricted mHag DFFRY. Both versions of DFFRY were predicted to bind to HLA-A*0101. For all other Y-chromosomal mHags, epitope prediction was used. RPS4Y/HLA-DRB3*0301 was correctly classified as binder. No allelic model was available to predict the mHag DBY/HLA-DQ. For two gonosomal mHags (UTY/HLA-B60 and SMCY/HLA-B7), no model of the appropriate length was available, so all related models of shorter length were used. For both mHags, at least one nonameric substring was predicted to bind by the respective model. The remaining three mHags were correctly predicted to bind to the restricting allele.

These examples demonstrate the usefulness and applicability of EpiToolKit for the identification of potential epitopes and mHags.

CONCLUSION

As a convenient and user-friendly program, EpiToolKit enables the direct comparison of different epitope prediction software packages and thus allows for precise selection of HLA-presented peptides from any protein of choice. In addition, its unique feature of screening polymorphic proteins for HLA ligands with SNEPv2 will promote and facilitate the identification of mHags, including tissue-specific peptides, as well as epitopes from quickly mutating pathogens. We expect that EpiToolKit with its epitope prediction and in particular with the SNEPv2 screening will provide valuable support in the future identification of T cell epitopes of major clinical relevance.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

This work was supported by Deutsche Forschungsge meinschaft (SFB 685). Funding to pay the Open Access publication charges for this article was provided by Deutsche Forschungsgemeinschaft (SFB685).

Conflict of interest statement. None declared.

REFERENCES

1. DeLuca,D.S. and Blasczyk,R. (2007) The immunoinformatics of cancer immunotherapy. *Tissue Antigens*, **70**, 265–271.
2. Rötzschke,O., Falk,K., Stevanović,S., Jung,G., Walden,P. and Rammensee,H.G. (1991) Exact prediction of a natural T cell epitope. *Eur. J. Immunol.*, **21**, 2891–2894.
3. Rammensee,H., Bachmann,J., Emmerich,N.P., Bachor,O.A. and Stevanović,S. (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, **50**, 213–219.
4. Parker,K.C., Bednarek,M.A. and Coligan,J.E. (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.*, **152**, 163–175.
5. Dönnes,P. and Kohlbacher,O. (2006) SVMHC: a server for prediction of MHC-binding peptides. *Nucleic Acids Res.*, **34**, W194–W197.
6. Nielsen,M., Lundegaard,C., Worning,P., Lauemøller,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.
7. 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.
8. Honeyman,M.C., Brusic,V., Stone,N.L. and Harrison,L.C. (1998) Neural network-based prediction of candidate T-cell epitopes. *Nat. Biotechnol.*, **16**, 966–969.
9. Bui,H., Sidney,J., Peters,B., Sathiamurthy,M., Sinichi,A., Purton,K., 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.
10. Peters,B., Bui,H., Frankild,S., Nielson,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.
11. Schuler,M.M., Dönnes,P., Nastke,M.D., Kohlbacher,O., Rammensee,H.G. and Stevanović,S. (2005) SNEP: SNP-derived epitope prediction program for minor H antigens. *Immunogenetics*, **57**, 816–820.
12. Halling-Brown,M., Quartey-Papafio,R., Travers,P.J. and Moss,D.S. (2006) SiPep: a system for the prediction of tissue-specific minor histocompatibility antigens. *Int. J. Immunogenet.*, **33**, 289–295.
13. Feldhahn,M. (2006) FRED: a framework for T-cell epitope detection. Diploma thesis, University of Tuebingen, Tuebingen, Germany.
14. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
15. The Uniport Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D65.
16. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
17. 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.
18. Pfeifer,N. and Kohlbacher,O. (2008) Multiple Instance Learning allows MHC class II epitope predictions for alleles without experimental data. Manuscript submitted.
19. Goulmy,E. (2004) Minor histocompatibility antigens: allo target molecules for tumor-specific immunotherapy. *Cancer J.*, **10**, 1–7.