BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 16 2010, pages 2066–2068
doi:10.1093/bioinformatics/btq324

*Systems biology*

Advance Access publication June 23, 2010

# EpiTOP—a proteochemometric tool for MHC class II binding prediction

Ivan Dimitrov[1], Panayot Garnev[1], Darren R. Flower[2] and Irini Doytchinova[1,*]

[1]Faculty of Pharmacy, Medical University of Sofia, 2 Dunav st, 1000 Sofia, Bulgaria and [2]Life and Health Sciences, Aston University, Aston Triangle, Birmingham, B4 7ET, UK

Associate Editor: Trey Ideker

**ABSTRACT**

**Motivation:** T-cell epitope identification is a critical immunoinformatic problem within vaccine design. To be an epitope, a peptide must bind an MHC protein.

**Results:** Here, we present EpiTOP, the first server predicting MHC class II binding based on proteochemometrics, a QSAR approach for ligands binding to several related proteins. EpiTOP uses a quantitative matrix to predict binding to 12 HLA-DRB1 alleles. It identifies 89% of known epitopes within the top 20% of predicted binders, reducing laboratory labour, materials and time by 80%. EpiTOP is easy to use, gives comprehensive quantitative predictions and will be expanded and updated with new quantitative matrices over time.

**Availability:** EpiTOP is freely accessible at http://www.pharmfac.net/EpiTOP

**Contact:** idoytchinova@pharmfac.net

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on October 23, 2009; revised on June 8, 2010; accepted on June 13, 2010

## 1 INTRODUCTION

T-cell epitope identification is a challenging immunoinformatic problem within vaccine design. To be an epitope, a peptide should bind a major histocompatibility complex (MHC) protein. For MHC class I, epitopes typically comprise 8–10 residues. The MHC class II binding site is open-ended, allowing much longer peptides to bind, although only nine amino acids occupy the site. Many computational methods have been developed for T-cell epitopes: see Flower (2008). Many work well and are widely used by immunologists and vaccinologists.

Most available epitope prediction methods separately address peptides binding particular MHC proteins, developing models for a single target allele. For MHC class II, only the generalized artificial neural network (ANN)-based server NetMHCIIpan uses both peptide and human leukocyte antigen (HLA) sequence information (Nielsen *et al.*, 2008). Recently, we developed a proteochemometrics-based approach to MHC class II prediction (Dimitrov *et al.*, 2010). Proteochemometrics, a quantitative structure-activity relationships (QSAR) approach originally developed by Wikberg (Lapinsh *et al.*, 2001), deals with ligands binding to several related proteins. In conventional QSAR, the X matrix of descriptors includes only information from ligands. In proteochemometrics, the X matrix contains information from proteins and ligands. A single proteochemometric model could potentially predict peptide binding to many MHC proteins.

We have developed and validated several models for binding to several HLA-DRB1 alleles, and now make the best model available in the server EpiTOP. It uses a quantitative matrix (QM) to predict peptide affinity to 12 HLA-DRB1 proteins: DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0404, DRB1*0405, DRB1*0701, DRB1*0802, DRB1*0901, DRB1*1101, DRB1*1201, DRB1*1301 and DRB1*1501.

## 2 ALGORITHM

The EpiTOP algorithm was described in detail elsewhere (Dimitrov *et al.*, 2010). Briefly, the QM was derived from 2666 known binders of different length, binding to 12 HLA-DRB1 alleles, and which were extracted from the Immune Epitope database (IEDB) (September 2008) (Peters *et al.*, 2005). Peptides are described using three $z$-scales per residue broadly corresponding to volume, hydrophobicity and polarizability (Hellberg *et al.*, 1987). Nonamers are encoded by a sequence of 27 $z$-descriptors (9 positions $\times$ 3 $z$-scales), forming the L block. HLA-DRB1 alleles are encoded by 54 descriptors (18 positions $\times$ 3 $z$-scales), forming the P block. We use only polymorphic residues within the binding site that interact with the peptide. The model also contains cross-terms for adjacent peptide positions (L12 block) and peptide–protein cross-terms (LP block). The LP block contains cross-terms for peptide–protein amino acid interactions in pockets 1, 4, 6, 7 and 9. The affinities of binders were assessed as p$IC_{50}$ values.

The QM was derived using the iterative self-consistent (ISC) algorithm (Doytchinova and Flower, 2003). Briefly, the initial training set included all nonamers with anchors (Tyr, Phe, Trp, Leu, Ile, Met and Val) at position 1 ($n = 10670$). This was used to extract the first model. The optimum number of principal components (PCs) was derived by cross-validation in seven groups. The first model was used to predict p$IC_{50}$s of the initial set and the best predicted nonamers from each parent peptide formed a second training set. This second set was used to produce the second model, which predicts p$IC_{50}$s of the initial training set. The best predicted nonamers from each parent peptide were selected and placed in a third training set. The selection procedure was repeated until the peptides in consecutive derived training sets were the same at the 99% level.

Protein sequences are submitted to EpiTOP in one letter format. A protein is divided into overlapping nonamers. Only nonamers

*To whom correspondence should be addressed.

 

bearing anchor residues at position 1 are assessed, the rest being omitted as non-binders. The binding affinities of the nonamers are predicted using the derived proteochemometric QM. In the results page, nonamers are arranged in descending order according to $pIC_{50}$. Results can be expressed using six different cutoffs: top 5%, 10%, 15%, 20%, 25% and all binders.

## 3 IMPLEMENTATION

EpiTOP 1.0 is a web-based application written in PHP and HTML, and integrating the MySQL database environment. It is freely accessible via http://www.pharmfac.net/EpiTOP. EpiTOP identifies peptides binding to HLA-DRB1 alleles within protein sequences, with options to vary HLA allele and cutoff.

## 4 PERFORMANCE

Three test sets were used to benchmark EpiTOP performance: AntiJen, IEDB and Lin's datasets. The evaluation based on AntiJen and IEDB datasets was performed under conditions similar to those an experimental immunologist might use: the complete protein sequences were submitted to a server and the results recorded. Five thresholds were used: top 5%, 10%, 15%, 20% and 25% of predicted binding nonamers. Identified binders are shown as a percentage of all binders (*sensitivity*). The predictive ability of EpiTOP was compared to eight other servers: SVMHC (Dönnes and Elofsson, 2002), ProPred (Singh and Raghava, 2001), RANKPEP (Reche *et al*., 2004), IEDB-ARB (Bui *et al*., 2005), IEDB-SMM_align (Nielsen *et al*., 2007), MHC2Pred (http://www.imtech.res.in/raghava/mhc2pred/), NetMHCII (http://www.cbs.dtu.dk/services/NetMHCII) and NetMHCIIpan (Nielsen *et al*., 2008). SVMHC, ProPred, RANKPEP, IEDB-ARB, IEDB-SMM_align and EpiTOP are QM-based methods; MHC2Pred uses SVM, NetMHCII and NetMHCIIpan are ANN based. Some of the servers do not predict binding to all DRB1 alleles used in the test sets. Only servers IEDB, NetMHCIIpan and EpiTOP make predictions for all 12 DRB1 alleles. Although many methods give quantitative predictions, in our evaluation they were used as classification methods. Each server was evaluated only on the alleles it predicts.

The evaluation using Lin's dataset was performed in terms of receiver operating characteristic (ROC) statistics (Bradley, 1997). Two variables *sensitivity* and *1-specificity* were calculated at different thresholds. The area under curve (AUC) is a quantitative measure of predictive ability and varies from 0.5 for random prediction to 1.0 for a perfect prediction. The performance of EpiTOP was compared to that of four other servers: SVMHC, ProPred, IEDB-SMM and NetMHCIIpan (Nielsen and Lund, 2009).

AntiJen and IEDB datasets used for benchmark are given as Supplementary Data I. Lin's dataset is freely accessible at http://bio.dfci.harvard.edu/DFRMLI. The detailed results are given as Supplementary Material II.

*AntiJen benchmark dataset*: the AntiJen dataset consisted of 116 epitopes belonging to 29 proteins and binding to 6 HLA-DRB1 alleles (Supplementary Material I). It was extracted from the AntiJen database (Toseland *et al*., 2005). These epitopes bind to DRB1*0101 (22 binders), DRB1*0301 (7 binders), DRB1*0401 (62 binders), DRB1*0404 (1 binder), DRB1*1101 (2 binders) and DRB1*1501 (22 binders).

**Table 1.** *Sensitivity* at different cutoffs for AntiJen dataset

| Server | Top 5% (%) | Top 10% (%) | Top 15% (%) | Top 20% (%) | Top 25% (%) |
|---|---|---|---|---|---|
| SVMHC | 38 | 40 | 40 | 40 | 40 |
| ProPred | 58 | 69 | 69 | 69 | 69 |
| RANKPEP | 51 | 53 | 53 | 53 | 53 |
| IEDB-ARB | 44 | 58 | 64 | 65 | 66 |
| IEDB-SMM | 12 | 16 | 16 | 19 | 20 |
| MHC2Pred | 56 | 66 | 77 | 82 | 87 |
| NetMHCII | 55 | 75 | 87 | 91 | 97 |
| NetMHCIIpan | 65 | 80 | 90 | 96 | 97 |
| EpiTOP | 44 | 71 | 85 | 89 | 95 |

The total number of binders is 116. Time of evaluation: September 2009. Allele-specific performance is given in Supplementary Material II.

**Table 2.** *Sensitivity* at different cutoffs for IEDB dataset

| Server | Top 5% (%) | Top 10% (%) | Top 15% (%) | Top 20% (%) | Top 25% (%) |
|---|---|---|---|---|---|
| ProPred | 46 | 55 | 55 | 55 | 55 |
| RANKPEP | 44 | 67 | 80 | 88 | 88 |
| IEDB-ARB | 15 | 25 | 34 | 41 | 47 |
| IEDB-SMM | 22 | 35 | 45 | 53 | 59 |
| MHC2Pred | 19 | 29 | 38 | 46 | 52 |
| NetMHCII | 55 | 73 | 83 | 89 | 92 |
| NetMHCIIpan | 55 | 75 | 86 | 92 | 95 |
| EpiTOP | 45 | 66 | 80 | 89 | 93 |

The total number of binders is 4540. Time of evaluation: January 2010. Allele-specific performance is given in Supplementary Material II.

The results from the evaluation based on AntiJen test set are shown in Table 1. For the top 5% cutoff, EpiTOP is sixth in sensitivity, for the top 10%—fifth and for the top 15–25%—third after NetMHCIIpan and NetMHCII.

*IEDB benchmark dataset*: the dataset extracted from the Immune Epitope database (December 2009) consisted of 4540 epitopes, originating from 167 proteins (Supplementary Material I). The peptides from this set bind to 12 DRB1 alleles: DRB1*0101 (2051 binders), DRB1*0301 (190 binders), DRB1*0401 (392 binders), DRB1 *0404 (159 binders), DRB1*0405 (244 binders), DRB1*0701 (336 binders), DRB1*0802 (153 binders), DRB1*0901 (160 binders), DRB1*1101 (275 binders), DRB1*1201 (24 binders), DRB1*1302 (243 binders) and DRB1*1501 (313 binders).

The results from the evaluation based on IEDB are given in Table 2. At the time of the evaluation (January 2010), SVMHC was not accessible and it was excluded from the study. For the top 5% cutoff, EpiTOP is third in sensitivity; for the top 10% it is fourth; for the top 15%, third together with RANKPEP; and for the top 20% and 25%, it is second after NetMHCIIpan.

*Lin's benchmark dataset*: Lin's dataset (http://bio.dfci.harvard.edu/DFRMLI/) consists of 103 overlapping peptides derived from four protein antigens—bee venom phospholipase A2, dog lipocalin, tumor antigen LAGE-1 and viral antigen HIV NEF (Lin *et al*., 2008). The binding affinities to seven HLA-DR molecules (DRB1*0101, *0301, *0401, *0701, *1101, *1301 and *1501) were measured using a competition assay. We excluded allele DRB1*1301 as several servers did not predict binding to it.

Results from this evaluation are shown in Table 3. AUC values for SVMHC, ProPred, IEDB-SMM and NetMHCIIpan are taken from

**Table 3.** AUC values for Lin's dataset

| DRB1 allele | SVMHC | ProPred | IEDB-SMM | Net MHCII | EpiTOP |
|---|---|---|---|---|---|
| *0101 | 0.86 | 0.89 | 0.81 | 0.90 | 0.72 |
| *0301 | 0.69 | 0.70 | 0.71 | 0.78 | 0.89 |
| *0401 | 0.75 | 0.75 | 0.79 | 0.84 | 0.84 |
| *0701 | 0.74 | 0.74 | 0.67 | 0.75 | 0.73 |
| *1101 | 0.83 | 0.83 | 0.84 | 0.85 | 0.79 |
| *1501 | 0.66 | 0.66 | 0.67 | 0.79 | 0.74 |
| Average | 0.76 | 0.76 | 0.75 | 0.82 | 0.79 |

The total number of peptides is 103. Results for SVMHC, ProPred, IEDB-SMM and NetMHCII are taken from Nielsen and Lund (2009). Time of evaluation: April 2010. Protein-specific performance is given in Supplementary Material II.

**Table 4.** Identification of peptide binding core

| DRB1 allele | PDB code | Peptide | TEPITOPE | IEDB-SMM | NetMHCIIpan | NetMHCII | EpiTOP |
|---|---|---|---|---|---|---|---|
| *0101 | 2fse | AG**FKGEQGPKG**EPG | √ | √ | √ | √ | √ |
| *0101 | 2iam | GEL**IGILNAAKV**PAD | √ | √ | √ | √ | √ |
| *0101 | 1sje | PE**VIPMFSALS**EGATP | √ | √ | √ | √ | X |
| *0101 | 1dlh | PK**YVKQNTLKL**AT | √ | √ | √ | √ | √ |
| *0101 | 1aqd | VGSD**WRFLRGYHQ**YA | √ | √ | √ | √ | √ |
| *0101 | 1pyw | A**FVKQNAAAL**A | √ | X | √ | √ | √ |
| *0101 | 1t5w | AA**YSDQATPLL**LSPR | √ | √ | √ | √ | √ |
| *0301 | 1a6a | PVSK**MRMATPLLM**QA | √ | √ | √ | √ | X |
| *0401 | 2seb | AY**MRADAAAGG**A | √ | √ | X | X | X |
| *0401 | 1j8h | PK**YVKQNTLKL**AT | √ | √ | √ | √ | √ |
| *1501 | 1bx2 | ENPV**VHFFKNIVT**PR | √ | √ | √ | √ | √ |

Binding core is given in bold. Results for TEPITOPE, IEDB-SMM, NetMHCIIpan and NetMHCII are taken from Nielsen and Lund, 2009. Time of evaluation: April 2010. Detailed scores for EpiTOP are given in Supplementary Data II.

a previous study (Nielsen and Lund, 2009). EpiTOP has the second best result after NetMHCIIpan.

*Identification of the peptide binding core*: EpiTOP was tested to identify the peptide binding core on a set of X-ray data for peptide-DRB1 allele complexes (Nielsen *et al*., 2008). EpiTOP performance was compared to those of TEPITOPE (Sturniolo *et al*., 1999), IEDB-SMM, NetMHCIIpan and NetMHCII, published by Nielsen and Lund (2009; Table 4).

EpiTOP identified 8 out of 11 binding cores correctly (73%). Two of the misaligned cores (MRMATPLLM and MRADAAAGG) are second best binders with predicted p$IC_{50}$ values very close to the best binders (Supplementary Material II).

## 5 DISCUSSION

We undertook a rigorous evaluation of the performance of EpiTOP across three datasets, comparing it to that of either four or eight other servers, using either recall statistics or ROC analysis. Overall, EpiTOP compares very favourably with other more specialized models. For the AntiJen and IEDB datasets, EpiTOP performs sub-optimally only at the highest specificity; since it is strongly inclusive and much broader in its predictive potential, other much more highly focused, allele-specific models outperform it at this highly stringent level. At more permissive thresholds, and within statistical error, EpiTOP performs identically to best-in-class servers.

For Lin's dataset, the five tested binders were much close in performance: within the limits of error, EpiTOP performs sub-optimally only for DRB1*0101. It is interesting to note that EpiTOP is very much the best model for DRB1*0301.

For the identification of binding cores, EpiTOP again performs well, but interestingly core identification correlates inversely with overall statistical perform, perhaps suggesting that much remains to be understood regarding class MHC–peptide interaction.

Taken together, results from these various benchmarking exercises both validate EpiTOP and indicate that development of a synergistic meta-server, which integrates results from several servers should prove a useful exercise, hopefully yielding significant overall enhancements.

## 6 CONCLUSION

EpiTOP is the first proteochemometrics-based server for T-cell epitope prediction. It is a tool for performing preliminary computational analyses of large datasets for accelerated epitope-based vaccine design. It is easy to use, gives comprehensive quantitative predictions and will be expanded and updated with new QMs.

*Funding*: National Science Fund of Ministry of Education and Science, Bulgaria (02-115/2008); Wellcome Trust (WT079287MA).

*Conflict of Interest*: none declared.

## REFERENCES

Bradley,A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**, 1145–1159.

Bui,H.H. *et al*. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics*, **57**, 304–314.

Dimitrov,I. *et al*. (2010) Peptide binding to the HLA-DRB1 supertype: a proteochemometrics analysis. *Eur. J. Med. Chem.*, **45**, 236–243.

Dönnes,P. and Elofsson,A. (2002) Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics*, **3**, 25.

Doytchinova,I.A. and Flower,D.R. (2003) Towards the in silico identification of class II restricted T-cell epitopes: a partial least squares iterative selfconsistent algorithm for affinity prediction. *Bioinformatics*, **19**, 2263–2270.

Flower,D.R. (2008) Vaccines: data driven prediction of binders, epitopes and immunog enicity. In Flower,D.R. (ed), *Bioinformatics for Vaccinology*, Wiley-Blackwell, Chichester, UK, pp.167–216.

Hellberg,S. *et al*. (1987) Peptide quantitative structure-activity relationships, a multivariate approach. *J. Med. Chem*., **30**, 1126–1135.

Lapinsh,M. *et al*. (2001) Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta*, **1525**, 180–190.

Lin,H.H. *et al*. (2008) Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. *BMC Bioinformatics*, **9**, S22.

Nielsen,M. *et al*. (2007) Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics*, **8**, 238.

Nielsen,M. *et al*. (2008) Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. *PLoS Comput. Biol*., **4**, e1000107.

Nielsen,M. and Lund,O. (2009) NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics*, **10**, 296.

Peters,B. *et al*. (2005) The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol*., **3**, e91.

Reche,P.A. *et al*. (2004) Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics*, **56**, 405–419.

Singh,H. and Raghava,G.P.S. (2001) ProPred: prediction of HLA-DR binding sites. *Bioinformatics*, **17**, 1236–1237.

Sturniolo,T. *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.

Toseland,C.P. *et al*. (2005) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical and cellular data. *Immunome Res.*, **1**, 4.