SHORT COMMUNICATION

EpiDOCK: a molecular docking-based tool for MHC class II binding prediction

Mariyana Atanasova[1], Atanas Patronov[1], Ivan Dimitrov[1], Darren R.Flower[2] and Irini Doytchinova[1,3]

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

[3]To whom correspondence should be addressed.
E-mail: idoytchinova@pharmfac.net

Edited by Leo James

Cellular peptide vaccines contain T-cell epitopes. The main prerequisite for a peptide to act as a T-cell epitope is that it binds to a major histocompatibility complex (MHC) protein. Peptide MHC binder identification is an extremely costly experimental challenge since human MHCs, named human leukocyte antigen, are highly polymorphic and polygenic. Here we present EpiDOCK, the first structure-based server for MHC class II binding prediction. EpiDOCK predicts binding to the 23 most frequent human, MHC class II proteins. It identifies 90% of true binders and 76% of true non-binders, with an overall accuracy of 83%. EpiDOCK is freely accessible at http:// epidock.ddg-pharmfac.net.
Keywords: docking/MHC class II binding prediction/peptide vaccines/quantitative matrices

Introduction

Cellular peptide vaccines contain T-cell epitopes. The T-cell epitope is a continuous sequence of variable length. This sequence represents just a fragment of a larger foreign protein that has been digested inside the cell. In order to be recognised by the components of the immune system, the T-cell epitope needs to be presented on the surface of an antigen-presenting cell facilitated by a specialised family of proteins named major histocompatibility complex (MHC). The main prerequisite for a peptide to act as a T-cell epitope is that it binds to an MHC protein. The stable binding of the peptide to the MHC molecule is considered as the major bottleneck in the complicated pathway of antigen presentation. Peptide MHC binder identification is an extremely costly experimental challenge as human MHCs, named human leukocyte antigen (HLA), are highly polymorphic and polygenic. More than 8700 molecules are listed in IMGT/HLA database (Robinson *et al.*, 2011). The only tractable alternative approach is MHC binding prediction using computational methods. Dependent on the available input data, the techniques range from plain

sequence-based approaches, to more detailed, structure-based models (Patronov and Doytchinova, 2013).

There are two classes of MHC molecules: class I and class II. MHC class I molecules typically present peptides from proteins synthesised within the cell (endogenous processing pathway). MHC class II proteins primarily present peptides derived from endocytosed extracellular proteins (exogenous processing pathway). MHC class I proteins are encoded by three loci: HLA-A, HLA-B and HLA-C. MHC class II proteins also are encoded by three loci: HLA-DR, HLA-DQ and HLA-DP. The peptide binding site of class I proteins has a closed cleft, formed by a single protein chain (α-chain) (Janeway *et al.*, 1999). Usually, only short peptides of 8–11 amino acids bind in extended conformation. In contrast, the cleft of class II proteins is open-ended, allowing much longer peptides to bind, although only nine amino acids actually occupy the site. The class II cleft is formed by two separate protein chains: α and β (Janeway *et al.*, 1999). Both clefts have binding pockets, corresponding to primary and secondary anchor positions on the binding peptide. The combination of two or more anchors is called a motif.

Here we present EpiDOCK, the first structure-based server for MHC class II binding prediction. EpiDOCK predicts binding to the 23 most frequent human MHC class II proteins: 12 HLA-DR, 6 HLA-DQ and 5 HLA-DP proteins. These alleles cover more than 95% of the human population. EpiDOCK is freely accessible at: http://epidock.ddg-pharmfac.net/.

Input data description

EpiDOCK accepts as input the target protein sequence in fasta format (Fig. 1). Multi-fasta protein format is also supported. The next step is to select the HLA class II protein of interest. There are two options for selection: single protein and all proteins. The last step is 'Get the result'.

Output data description

EpiDOCK converts the input sequence into a collection of overlapping nonamers, because the peptide binding core consists of nine contiguous residues. Every nonamer is evaluated by a docking score-based quantitative matrix (DS-QM) derived for the selected HLA class II protein and assigned a score. A threshold for binding to each HLA protein is given (Fig. 2). Peptides with scores higher or equal to the given threshold are predicted to be binders, otherwise they are non-binders. The data can be exported either in xls or csv formats.

Processing method

The method implemented in EpiDOCK was described elsewhere (Atanasova *et al.*, 2011; Patronov *et al.*, 2011, 2012,

M.Atanasova et al.

Fig. I. EpiDOCK front page.

Downloaded from https://academic.oup.com/peds/article/26/10/631/1508922 by guest on 23 April 2024

2013). Briefly, three X-ray structures of the peptide–HLA class II protein complex were used as starting templates to model by homology the structure of the studied HLA proteins. These were: 2g9h for DR proteins, 1s9v and 1jk8 for DQ proteins, and 3lqz for DP proteins. The resulting models were used to generate virtual combinatorial peptide libraries using the single amino acid substitution (SAAS) principle for the nine amino acids of the peptide binding core. Peptides with flexible SAAS residues were docked into the rigid HLA binding site using either AutoDock (Morris *et al.*, 2009) or GOLD (Jones *et al.*, 1997). The resulting scores were normalised and transformed into DS-QMs with dimensions of 9 positions × 20 amino acids. One DS-QM was generated for each HLA protein.

Validation of predictions

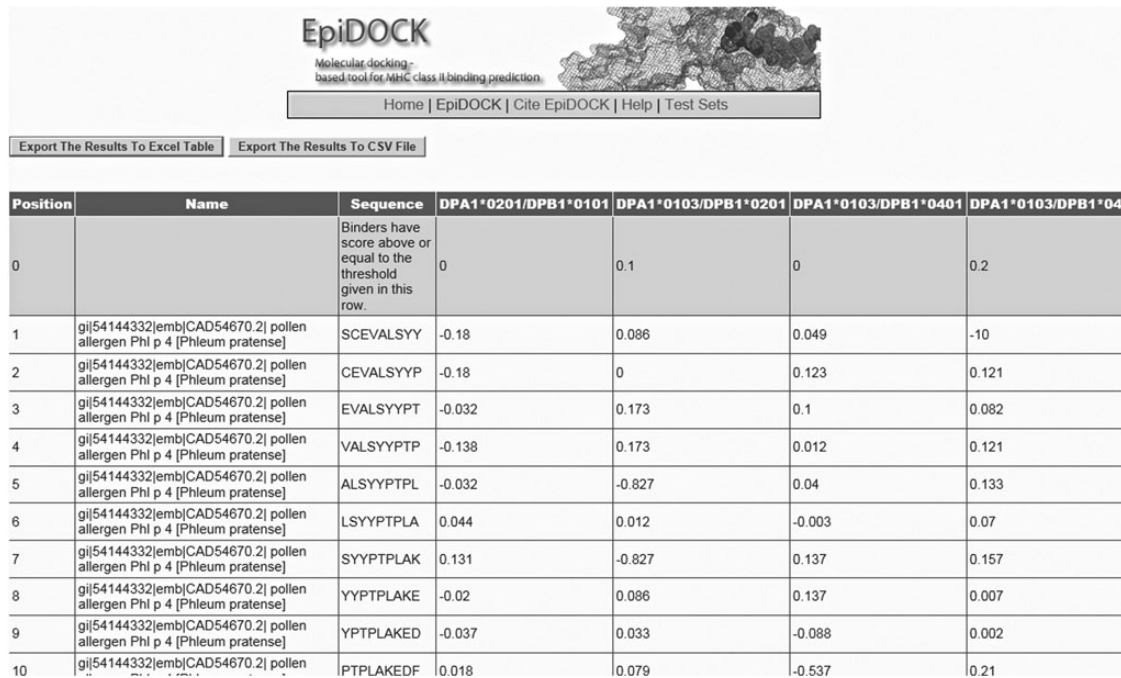A test set of 7050 known binders to HLA-DR, HLA-DQ and HLA-DP proteins, originating from 1195 proteins, was

632

Fig. 2. EpiDOCK results page.

Table I. Validation of EpiDOCK

HLA protein	Binding peptides/ parent proteins	Threshold	Sensitivity	Specificity	Accuracy	AUC	NetMHCII AUC
DRB1*01:01	1883/122	0.3	0.887	0.770	0.828	0.894	0.655
DRB1*03:01	166/48	0.4	0.896	0.782	0.838	0.896	0.681
DRB1*04:01	346/105	0.3	0.845	0.827	0.836	0.906	0.688
DRB1*04:04	149/32	0.3	0.905	0.755	0.830	0.891	0.675
DRB1*04:05	225/61	0.3	0.949	0.733	0.841	0.902	0.682
DRB1*07:01	308/77	0.3	0.948	0.733	0.841	0.902	0.689
DRB1*08:02	123/53	0.3	0.912	0.742	0.828	0.893	0.678
DRB1*09:01	140/48	0.4	0.926	0.769	0.847	0.914	0.687
DRB1*11:01	247/73	0.5	0.851	0.864	0.858	0.922	0.670
DRB1*12:01	22/16	0.7	0.894	0.936	0.916	0.960	–[a]
DRB1*13:02	221/56	0.3	0.928	0.731	0.829	0.896	0.708
DRB1*15:01	294/71	0.2	0.961	0.730	0.845	0.900	0.693
DQA1*05:01/DQB1*02:01	333/35	0.1	0.928	0.664	0.796	0.865	0.687
DQA1*05:01/DQB1*03:01	558/22	0.3	0.820	0.817	0.818	0.892	0.624
DQA1*03:01/DQB1*03:02	340/31	0.1	0.822	0.730	0.776	0.860	0.670
DQA1*04:01/DQB1*04:02	296/18	0.1	0.932	0.711	0.822	0.889	0.696
DQA1*01:01/DQB1*05:01	174/19	0.1	0.939	0.726	0.833	0.897	0.679
DQA1*01:02/DQB1*06:02	418/24	0.1	0.921	0.615	0.768	0.854	0.664
DPA*02:01/DPB1*01:01	102/60	0	1.000	0.674	0.837	0.864	0.630
DPA*01:03/DPB1*02:01	323/21	0.1	0.864	0.791	0.828	0.898	0.614
DPA*01:03/DPB1*04:01	152/71	0	1.000	0.876	0.938	0.943	0.661
DPA*01:03/DPB1*04:02	122/66	0.2	0.684	0.776	0.730	0.807	0.643
DPA*02:01/DPB1*05:01	108/66	0.1	0.952	0.710	0.831	0.881	0.621
Average			0.903	0.759	0.831	0.892	0.667

[a]NetMHCII does not predict peptide binding to DRB1*12:01.

collected from the Immune Epitope Database (http://www .immuneepitope.org) in July 2012 (Vita *et al.*, 2010). Most peptides bind to more than one HLA. Each protein was presented as a set of overlapping nonamers and for each nonamer a score was calculated using the corresponding DS-QM. The thresholds were defined at the maximum accuracy for each DS-QM. Peptides with scores above or equal this threshold were predicted as binders. If the predicted nonamer binder was part of the known binder sequence, the predicted peptide was

considered as a true predicted binder, otherwise it was considered as a false binder. The predictions were assessed by *sensitivity* (*true binders/all binders*), *specificity* (*true non-binders/ all non-binders*), *accuracy* (*true binders and non-binders/all peptides*) and *area under curve* (*AUC*) using coordinates *sensitivity/1-specificity* (Bradley, 1997). The average values for *sensitivity*, *specificity*, *accuracy* and *AUC* were 0.903, 0.759, 0.831 and 0.892, respectively (Table I). The test set used for validation is available via the EpiDOCK website.

Using the same test set, EpiDOCK was benchmarked against the *de facto* 'state-of-the-art' server for MHC class II binding prediction: NetMHCII (Andreatta and Nielsen, 2012). NetMHCII is sequence-based and uses artificial neural networks; it has been shown previously to be the-best-in-class (Lin *et al*., 2008). The *AUC* values for NetMHCII are given in Table I (last column). EpiDOCK shows significantly higher average *AUC* value than NetMHCII: 0.892 vs. 0.667.

Discussion

In contrast to sequence-based methods, structure-based approaches do not require extensive pre-existing experimental data. The only information needed is an X-ray structure of the peptide—MHC protein complex. Analysis of the peptide—MHC protein binding interface by structure-based methods such as molecular docking and molecular dynamics reveal clear and unambiguous amino acid preferences at each peptide binding core position (Khan and Ranganathan, 2010; Atanasova *et al*., 2011; Doytchinova *et al*., 2011; Patronov *et al*., 2011). Here, we summarise, collate and extend results from our recent docking studies, making them freely accessible *via* EpiDOCK: the first structure-based server for MHC binding prediction.

The external validation of EpiDOCK indicates the server's high predictive ability. EpiDOCK identifies 90% of true binders and 76% of true non-binders; with an overall accuracy of 83%. In terms of lab work, the usage of good *in silico* predictors means significantly less expenditure on materials, labour and time. Compared with the 'state-of-the-art' server for MHC class II binding prediction NetMHCII, EpiDOCK gave better predictions in terms of *AUC* values: 0.892 compared with 0.667. On an individual HLA-by-HLA basis, EpiDOCK was significantly higher in all cases: between 0.16 and 0.28 greater than NetMHCII.

Given the open-minded nature of scientific endeavour, there is no reason to think that technology such as EpiDOCK will not find a place among the tried-and-tested techniques available within immunology. There are many problems in need of solutions: the personalisation of cancer treatment and vaccines, the immunotherapy of asthma and allergy, as well as autoimmune disorders, are all among those crying out for innovation and success. The ability to identify epitopes of class II MHC has proved a significant technical challenge hitherto, but EpiDOCK represents a significant step-forward in this direction. In time, the clear predictive power of this approach will be complemented by the accurate prediction of proteolytic degradation of endocytosed peptides by cathepsins.

The enormous and confounding degeneracy inherent within the T-cell response is manifest by the vast intersection of the T-cell receptor repertoire, MHC haplotype and peptide epitome generated from pathogen proteomes. Structure-based methods, such as EpiDOCK, offer the most flexible, adaptable and practical approach to address the daunting combinatorial explosion implicit within the cellular response. EpiDOCK is accurate and reliable; we anticipate that it will prove valuable as a tool, able to predict class II mediated T-cell epitopes that form the basis of reagents, diagnostics and vaccines.

Funding

References

Andreatta,M. and Nielsen,M. (2012) *Immunology*, **136**, 306–311.
Atanasova,M., Dimitrov,I., Flower,D.R. and Doytchinova,I. (2011) *Mol. Inform.*, **30**, 368–375.
Bradley,A.P. (1997) *Pattern Recognit.*, **30**, 1145–1159.
Doytchinova,I., Petkov,P., Dimitrov,I., Atanasova,M. and Flower,D.R. (2011) *Protein Sci.*, **20**, 1918–1928.
Janeway,C.A., Travers,P., Walport,M. and Capra,J.D. (1999) In Janeway,C.A., Travers,P., Walport,M. and Capra,J.D. (eds), *Immunobiology: The Immune System in Health and Disease*. 4th ed. Elsevier, London, pp. 78–193.
Jones,G., Willett,P., Glen,R.C., Leach,A.R. and Taylor,R. (1997) *J. Mol. Biol.*, **267**, 727–748.
Khan,J.M. and Ranganathan,S. (2010) *Immunome Res.*, **6**(Suppl 1), S2.
Lin,H.H., Zhang,G.L., Tongchusak,S., Reinherz,E.L. and Brusic,V. (2008) *BMC Bioinformatics*, **9**, S22.
Morris,G.M., Huey,R., Lindstrom,W., Sanner,M.F., Belew,R.K., Goodsell,D.S. and Olson,A.J. (2009) *J. Comput. Chem.*, **30**, 2785–2791.
Patronov,A. and Doytchinova,I. (2013) *Open Biol.*, **3**, 120139.
Patronov,A., Dimitrov,I., Flower,D.R. and Doytchinova,I. (2011) *BMC Struct. Biol.*, **11**, 32.
Patronov,A., Dimitrov,I., Flower,D.R. and Doytchinova,I. (2012) *BMC Struct. Biol.*, **12**, 20.
Patronov,A., Salamanova,E., Dimitrov,I., Flower,D.R. and Doytchinova,I. (2013) *Curr. Comput. Aided Drug Des.*, in press.
Robinson,J., Mistry,K., McWilliam,H., Lopez,R., Parham,P. and Marsh,S.G.E. (2011) *Nucleic Acids Res.*, **39**(Suppl 1), D1171–D1176.
Vita,R., Zarebski,L., Greenbaum,J.A., Emami,H., Hoof,I., Salimi,N., Damle,R., Sette,A. and Peters,B. (2010). *Nucleic Acids Res.*, **38**(Database issue), D854–D862.