Structure Based Prediction of MHC II Binding Peptides

Josef Laimer, Markus Wiederstein, and Peter Lackner
University of Salzburg, Department of Biosciences
Hellbrunnerstraße 34, 5020 Salzburg/Austria

UNIVERSITY of SALZBURG

FWF Der Wissenschaftsfonds.

Introduction

The binding of an antigen peptide to MHC class II molecules is essential for initiating an immune response. Thus, fast and accurate identification of potential binding peptides is critical for basic research and clinical translation. Nowadays, various computational approaches exist for this task, which can roughly be divided into two classes: sequence-based methods employing machine learning (ML), and structure-based methods using physical concepts realized by docking, molecular dynamics, or threading. We present a novel structure-based approach which utilizes statistical scoring functions.

Methods

For our approach, statistical scoring functions (SSFs) are used to evaluate interactions between peptides and MHC II molecules. Thereby, predictions are performed on sets of allele-specific 3D models of MHC II molecules. SSFs have a wide range of applications in protein science, e.g. for the assessment of protein structures [1] or the scoring of protein-protein interactions [2]. Here, we utilize MAESTRO pSSFs [3]. Initially designed for the prediction of stability changes upon mutations, these SSFs have to be shown useful for other tasks. To put a focus on interactions between chains, we compiled the SSFs on a set of 1227 multimeric structures containing at least one polypeptide with a length between 5 and 20 residues. MHC models were excluded beforehand to prevent overfitting.

Based on a set of 161 experimental determined template structures (templates), at least 100 models for each HLA-DR allele were generated utilizing MODELLER [4]. All models include an alanine nonamer binding peptide as a placeholder. These placeholders represent the peptide binding core, their position was defined by a 3D alignment of the templates. Subsequently, the resulting models were scored with multiple scoring tools. The resulting models are provided at our M23D database [5].

Predictions are performed in three major steps: (i) First, a set of suitable models is selected based on their scores. Optionally, the peptide placeholders are replaced by alternative peptide conformations derived from the templates. We further investigated the effect of peptide conformations extended by two or four residues. These four variants of the peptide conformation are called 9-mer placeholder, 9-mer core, 11-mer core, and 13-mer core, for short. (ii) Then the potential binding peptide sequences are applied to each of the models and are subsequently scored. Multiple scores are computed if the target sequence is not of the same length as the peptide (placeholder) in the models. (iii) Finally, a consensus score is calculated by averaging the minimum scores achieved with each model. As the prediction is based on precalculated 3D models, the scoring of a certain binding peptide takes less than a second, which will allow predictions on a proteome scale.

Results

We tested our approach on a dataset provided by Jensen *et al.* [6]. The set contains 87363 log-transformed IC_{50} binding values for 36 HLA-DR molecules. For the performance tests below all peptides with an IC_{50} binding value $< 500nM$ were considered as binders. Table 1 shows performance results (AUC) per HLA allele in comparison with NetMHCIIpan-3.2. Figure 1 presents the accuracy on the whole dataset for the various peptide conformation setups.

MHC II allele	#peptides	#binders	NetMHCIIpan-3.2 5-fold[1,*]	NetMHCIIpan-3.2 LOMO[2,*]	MAESTRO pSSFs[3]
DRB1_0101	10412	6376	0.83	0.78	0.70
DRB1_0301	5352	1457	0.82	0.70	0.60
DRB1_0401	6317	3022	0.81	0.77	0.61
DRB1_0404	3657	1852	0.81	0.79	0.69
DRB1_0405	3962	1654	0.83	0.80	0.69
DRB1_0701	6325	3456	0.88	0.83	0.73
DRB1_0802	4465	2036	0.83	0.77	0.68
DRB1_0901	4318	2164	0.83	0.79	0.70
DRB1_1001	2066	1521	0.92	0.91	0.77
DRB1_1101	6045	2667	0.86	0.77	0.69
DRB1_1201	2384	759	0.87	0.80	0.73
DRB1_1301	1034	520	0.86	0.73	0.71
DRB1_1302	4477	2249	0.89	0.70	0.67
DRB1_1501	4850	2107	0.83	0.78	0.73
DRB1_1602	1699	989	0.88	0.87	0.72
DRB3_0101	4633	1415	0.89	0.80	0.60
DRB3_0202	3334	1055	0.87	0.76	0.69
DRB4_0101	3961	1540	0.82	0.73	0.67
DRB5_0101	5125	2430	0.85	0.77	0.65
Average			**0.85**	**0.78**	**0.69**

Table 1: Comparison of AUC performance between NetMHCIIpan-3.2 ([1]5-fold cross-validation and [2]Leave-one-molecule-out experiment) and [3]our approach based on MAESTRO pSSFs. Only alleles with more than 1000 sequences of potential binders are presented. [*]AUC values taken from Jensen *et al.* [6].

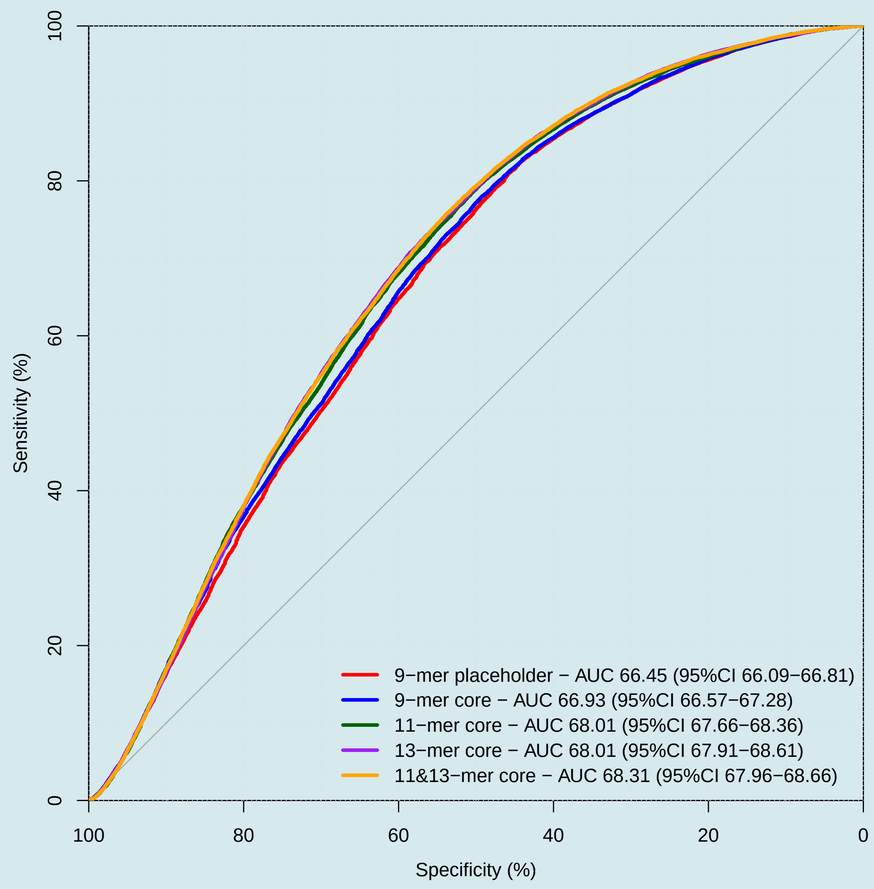

Figure 1: Prediction performance on a set of 87363 binding values. 9-mer placeholder: predictions are performed on 20 allele-specific models and their peptide placeholders. 9/11/13-mer core: predictions are performed on a single model, but with alternative peptide conformations derived from the available template structures. 11&13-mer core: combines the results of 11-mer and 13-mer core.

While the results clearly show that NetMHCIIpan-3.2 reaches better AUC values, they also show the potential of our approach, especially since our method was not specifically optimized or trained on these data. By extending the core binding peptide by the flanking residues (11/13-mer core), the results can be slightly increased. Besides the classification, SSFs also provide a promising base for the prediction of IC_{50} binding values itself. Our approach reaches a Pearson's ρ of 0.36 and a Spearman's ρ of 0.38, respectively.

Conclusion and Outlook

Our method is not limited to specific alleles or the availability of an experimentally determined structure. The utilized SSFs are not specifically trained on MHC binding data. Thus the risk of overfitting on certain training data is reduced to a minimum.

In ongoing work, the enrichment of our approach with various machine learning techniques is investigated. First experiments provide promising results. Further, the prediction capabilities will be extended to all loci.

An easy to use web interface and a RESTful web service is under development. The final version will be provided for free to the scientific community.

References

[1] **Recognition of errors in three-dimensional structures of proteins.** Sippl, 1993. https://www.came.sbg.ac.at/

[2] **DrugScorePPI knowledge-based potentials used as scoring and objective function in protein-protein docking.** Krüger, 2014

[3] **MAESTRO - multi agent stability prediction upon point mutations.** Laimer *et al.*, 2015. https://biwww.che.sbg.ac.at/maestro

[4] **Comparative protein modelling by satisfaction of spatial restraints.** Sali and Blundell, 1993. https://salilab.org/modeller/

[5] **M23D - MHC-II Model Database.** Laimer *et al.*, 2018. https://biwww.che.sbg.ac.at/m23d/

[6] **Improved methods for predicting peptide binding affinity to MHC class II molecules.** Jensen *et al.*, 2018. http://www.cbs.dtu.dk/services/NetMHCIIpan/

josef.laimer@sbg.ac.at

This work was funded by the Austrian Science Fund (FWF, grant P30042).