**The other peptide-binding prediction tools**

**Scoring functions**

**CombLib**[1-4] is based on combinatorial libraries of single residue substitution peptides[5], which were synthesized and experimentally characterized through standard binding assays to evaluate their IC50 values. These values were standardized, normalized, and used to compute the average relative binding affinity for each residue and each position in 9-residue peptides. The method performance was assessed by 5 fold cross-validation and reached an AUC value at 0.74. This approach has been successfully used to determine the binding specificity in several applications, including T-cell receptor (TCR) recognition[6].

**TEPITOPE**[7-10] is another well-known scoring function-based predictor. In addition to the sequence properties of the peptides, it uses the sequence profile of the binding pocket of MHC Ⅱ alleles as well as gene expression data obtained from DNA microarray technology[10] to predict the HLA-peptide interactions[6] This method proceeds by constructing multiple alignments of HLA-DR sequences with known 3D structures, and generates pocket profiles according to the composition of the peptides and the HLA-peptide interactions as measured by an ELISA-based high-throughput competitive binding assay. In this way, TEPITOPE establishes a Pocket Profile database[10] for 11 HLA-DR alleles in the approximation that pockets sharing polymorphic residues exhibit a similar pocket specificity profile. Its prediction performance ranges from 0.714 to 0.736 for HLA-DR binding peptide prediction. TEPITOPEpan[11] is an extension of TEPITOPE which extrapolates from the characterized binding specificities of HLA-DR molecules to those uncharacterized.

**Machine-learning based methods**

**PERUN**[12] was the first approach to implement a neural network model  into the MHC Ⅱ-peptide binding prediction[13]. It integrates anchor positions and binding motif information into a peptide alignment matrix using an evolutionary algorithm. Then, this matrix is used as input of a feedforward neural network（FNN） with two hidden layers to classify the peptide as binder or non-binder. The AUC values of the PERUN model reach 0.88, 0.86 and 0.73 respectively for high-, moderate-, and low-affinity binders in the experimental test dataset. The dataset for training and testing the method was obtained from the MHC binding database MHCPEP[14] with 650 entries only. However, it inspired researchers to apply ML methods for MHC Ⅱ-peptide interaction prediction.

**SVRMHC**[15] predicts MHC-peptide binding affinity (IC50 value) using a support vector regression (SVR) model[16] and is trained based on the AntiJen database[17]. The training of SVRMHC is performed by implementing input peptide sequences in each of the MHC class as a set of nonamers and aligning them by employing anchor position information from SYFPEITHI[18]. After the alignment step, an iterative self-consistent (ISC) strategy is then used to refine the alignments and reconstruct the models until the highest prediction performance is reached. Six SVR models with different kernels and input encoding schemes were constructed for each MHC Ⅱ allele class. The average predicted performance of the method on the HLA-DR benchmark reaches an AUC at 0.688.

The **NetMHCII and NetMHCIIpan predictors**[1, 19-28]is a family of predictors that have been developed and upgraded over the past two decades.

**NetMHCII-1.0**[1], was the first version released, is a weight matrix based method that predicts the binding affinity between a peptide sequence and a specific HLA class Ⅱ allele. It is based on SMM-align which is able to determine the 9-mer core binding motif of a given peptide without the need of sequence alignment and instead simply extracts the 9-mer with the highest predicted binding affinity both during training and inference. SMM-align was trained on ~ 5.000 peptide-HLA (pHLA) quantitative binding affinity values across 17 different HLA alleles. The allele-specific weight matrices were optimized to reproduce the corresponding experimental IC50 values by a Monte Caro search.

One of the limitations of NetMHCII-1.0 was that users could only select among the HLA class Ⅱ alleles that were available in the training dataset. To overcome this limitation, it was replaced by a neural network-based pan HLA-DR predictor named **NetMHCIIpan-1.0**[20]. To enable the prediction of HLA-DR alleles that are not part of the training dataset, the authors developed the “HLA Pseudo-Sequence”, which is a sequence of 21 residues that represents the binding groove of MHCⅡ receptors and was determined after a careful analysis of 15 peptide-MHCⅡ 3D protein structures. The peptide and HLA pseudo-sequence residues are encoded using one-hot encoding and BLOSUM50 substitution matrix encoding, respectively, meaning each residue requires 20 input neurons. In total, the neural network has 658 input neurons. The model was trained on ~14.600 pHLA binding affinity data points across 14 human HLA-DR alleles.

**NetMHCII-2.0**[21], which was published soon after as an update of NetMHCII-1.0, is largely inspired by the neural network approach used in NetMHCIIpan-1.0 described above. As a result, SMM-align was replaced by NN-align, which essentially expands on the SMM-align method but replaces the weight matrices with a neural network. As in SMM-align, NN-align determines the 9–mer core binding motif of a peptide by simply identifying the one with the best binding affinity both during training of the neural network and inference.

**NetMHCIIpan-2.0**[22] was released shortly after. It essentially updated the previous version by replacing SMM-align with NN-align and also updated the training dataset to ~34.00 pHLA binding affinity data points covering 24 HLA-DR alleles. The neural network architecture and input features remain exactly the same as in NetMHCIIpan-1.0. On a number of different benchmarks, the method reached an average AUC across different HLA-DR alleles of between 0.78 and 0.85.

Three years later, **NetMHCIIpan-3.0** was released[23], which is the first truly pan specific method as, in addition to HLA-DR alleles, it is capable of predicting pHLA binding affinity for any HLA-DP and HLA-DQ alleles. This was achieved by generalizing the concept of the HLA pseudo-sequence to all three HLA allele isotypes, which resulted in a new definition of the HLA pseudo-sequence that includes 34 binding residues instead of 21. The training dataset was also updated to ~ 52.000 pHLA binding affinity data points.

In 2018, **NetMHCII-2.3**[25] and **NetMHCIIpan-3.2**[25] were released following the update of the training datasets of both methods with a new training dataset containing ~134.000 pHLA binding affinity data points. This update slightly increased the performance of the methods, with an average AUC-ROC across all alleles of 0.76 - 0.86 for NetMHCII-2.3 and 0.78 - 0.86 for NetMHCIIpan-3.2.

A major update of NN-align was published in 2019[26] which enabled the method to be able to deal both with pHLA binding affinity data as well as LC-MS data. LC-MS data is inherently poly-specific, which means it returns a set of multiple HLA alleles and peptides that bind to them, but the exact pairing of each pHLA is unknown. To deconvolute this data and obtain pHLA pairs, they developed NN-align\_MA (Multi Allele). Briefly, NN-align was adapted by first training the neural network for a number of iterations on Single Allele (SA) data (i.e binding affinity pHLA data and some SA LC-MS experimental data) and the resulting model was used to annotate and thus deconvolute the MA LC-MS data. All the data is then combined and used to train the neural network. The latest pan allelic method, **NetMHCIIpan-4.1**[28], used the newly developed NN-align\_MA algorithm and a training dataset of > 500.00 binding affinity and LC-MS pHLA data, reaching a median ROC-AUC of 0.98 in cross-validation.

1. Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method, BMC bioinformatics 2007;8:1-12.

2. Sidney J, Assarsson E, Moore C et al. Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries, Immunome research 2008;4:1-14.

3. Wang P, Sidney J, Kim Y et al. Peptide binding predictions for HLA DR, DP and DQ molecules, BMC bioinformatics 2010;11:1-12.

4. Xu Y, Luo C, Qian M et al. MHC2MIL: a novel multiple instance learning based method for MHC-II peptide binding prediction by considering peptide flanking region and residue positions, BMC genomics 2014;15:1-9.

5. Uebel S, Kraas W, Kienle S et al. Recognition principle of the TAP transporter disclosed by combinatorial peptide libraries, Proceedings of the National Academy of Sciences 1997;94:8976-8981.

6. Wucherpfennig KW, Allen PM, Celada F et al. Polyspecificity of T cell and B cell receptor recognition. In: Seminars in immunology. 2007, p. 216-224. Elsevier.

7. Bian H, Hammer J. Discovery of promiscuous HLA-II-restricted T cell epitopes with TEPITOPE, Methods 2004;34:468-475.

8. Hammer J, Sturniolo T, Sinigagua F. HLA class II peptide binding specificity and autoimmunity, Advances in Immunology 1997;66:67-100.

9. Manici S, Sturniolo T, Imro MA et al. Melanoma cells present a MAGE-3 epitope to CD4+ cytotoxic T cells in association with histocompatibility leukocyte antigen DR11, The Journal of experimental medicine 1999;189:871-876.

10. Sturniolo T, Bono E, Ding J et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices, Nature biotechnology 1999;17:555-561.

11. Zhang L, Chen Y, Wong H-S et al. TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules, PLoS One 2012;7:e30483.

12. Brusic V, Rudy G, Honeyman G et al. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network, Bioinformatics (Oxford, England) 1998;14:121-130.

13. Zurada J. Introduction to artificial neural systems. West Publishing Co., 1992.

14. Brusic V, Rudy G, Kyne AP et al. MHCPEP—a database of MHC-binding peptides: update 1995, Nucleic acids research 1996;24:242-244.

15. Wan J, Liu W, Xu Q et al. SVRMHC prediction server for MHC-binding peptides, BMC bioinformatics 2006;7:1-5.

16. Liu W, Meng X, Xu Q et al. Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models, BMC bioinformatics 2006;7:1-13.

17. Toseland CP, Clayton DJ, McSparron H et al. AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data, Immunome research 2005;1:1-12.

18. Rammensee H-G, Bachmann J, Emmerich NPN et al. SYFPEITHI: database for MHC ligands and peptide motifs, Immunogenetics 1999;50:213-219.

19. Nielsen M, Lundegaard C, Worning P et al. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach, Bioinformatics 2004;20:1388-1397.

20. Nielsen M, Lundegaard C, Blicher T et al. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan, PLoS computational biology 2008;4:e1000107.

21. Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction, BMC bioinformatics 2009;10:1-10.

22. Nielsen M, Justesen S, Lund O et al. NetMHCIIpan-2.0-Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure, Immunome research 2010;6:1-10.

23. Karosiene E, Rasmussen M, Blicher T et al. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ, Immunogenetics 2013;65:711-724.

24. Andreatta M, Karosiene E, Rasmussen M et al. Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification, Immunogenetics 2015;67:641-650.

25. Jensen KK, Andreatta M, Marcatili P et al. Improved methods for predicting peptide binding affinity to MHC class II molecules, Immunology 2018;154:394-406.

26. Alvarez B, Reynisson B, Barra C et al. NNAlign\_MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved T-cell epitope predictions, Molecular & Cellular Proteomics 2019;18:2459-2477.

27. Reynisson B, Barra C, Kaabinejadian S et al. Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data, Journal of proteome research 2020;19:2304-2315.

28. Reynisson B, Alvarez B, Paul S et al. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data, Nucleic acids research 2020;48:W449-W454.