Chapter 18

Bioinformatics Tools for the Prediction of T-Cell Epitopes

Massimo Andreatta and Morten Nielsen

Abstract

T-cell responses are activated by specific peptides, called epitopes, presented on the cell surface by MHC molecules. Binding of peptides to the MHC is the most selective step in T-cell antigen presentation and therefore an essential factor in the selection of potential epitopes. Several in-vitro methods have been developed for the determination of peptide binding to MHC molecules, but these are all costly and time-consuming. In consequence, significant effort has been dedicated to the development of in-silico methods to model this event. Here, we describe two such tools, *NetMHCcons* and *NetMHCIIpan*, for the prediction of peptide binding to MHC class I and class II molecules, respectively, involved in the activation pathways of CD8+ and CD4+ T cells.

Key words T-cell epitopes, MHC binding, Prediction server, Artificial neural networks

1 Introduction

Major Histocompatibility Complex (MHC) molecules are transmembrane receptors that play an essential role in the cellular immune system of vertebrates. MHC molecules bind to short peptide fragments derived from pathogens and present them on the surface of antigen presenting cells, where they can be recognized by T cells [1, 2]. MHC class I molecules are primarily involved in the presentation of peptides derived from intracellular proteins to cytotoxic T cells, also called CD8+ T cells. In contrast, peptides presented by MHC class II molecules originate from proteins taken up from the extracellular environment, and can be recognized by helper T cells (CD4+ T cells). Because the structures of MHC class I and class II molecules are substantially different, the properties and size of the peptides that can bind to the two different classes are also distinct. The binding cleft of MHC class I molecules is closed at both ends and can accommodate only peptides of limited length, typically between 8 and 11 amino acids (Fig. 1a). Conversely, the binding groove of class II molecules is open at its extremities (Fig. 1b). This does not pose constraints on the length

Johan Rockberg and Johan Nilvebrant (eds.), *Epitope Mapping Protocols*, Methods in Molecular Biology, vol. 1785,
https://doi.org/10.1007/978-1-4939-7841-0_18, © Springer Science+Business Media, LLC, part of Springer Nature 2018

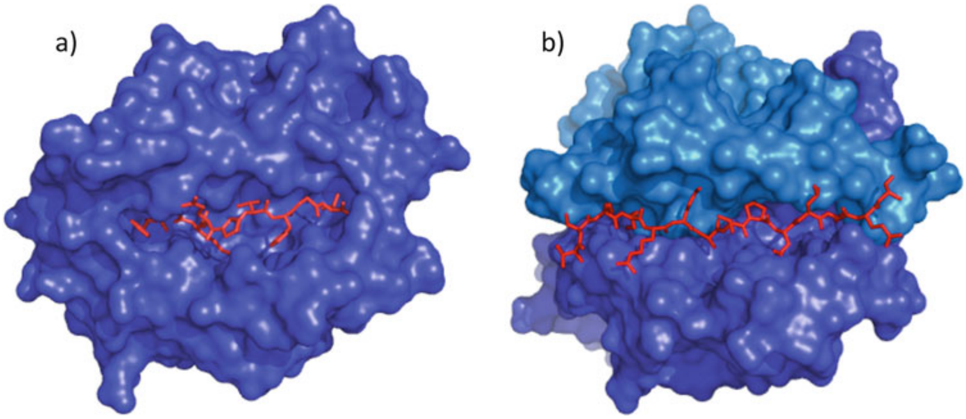


Fig. 1 The MHC class I and class II molecules with bound peptide ligands. (**a**) MHC class I molecule HLA-A2.1 (A2) with bound 9-mer peptide FLKEPVHGV in red sticks (PDB entry 1I1F). Note that the binding groove is closed at both ends and can accommodate only peptides of limited length. (**b**) MHC class II molecule HLA-DR1 with bound 14-mer peptide VSKMRMATPLLMQA (PDB entry 3QXA). The alpha chain is in light blue, the beta chain in dark blue. The HLA binding groove is open at both ends and the ligand can extend outside the extremities of the pocket

of the peptide ligands which can stick out freely at both ends and are typically between 11 and 20 amino acids long [3]. Computational methods for the prediction of binders to the two classes of MHC molecules, and ultimately for the prediction of CD8+ or CD4+ epitopes, have therefore been developed separately for the two problems. For a useful review of available methods for sequence-based T-cell epitope prediction *see* Lundegaard et al. [4]. Other factors than binding affinity to MHC determine if a peptide will induce a T-cell response. These factors include peptide processing [5–7], binding stability [8, 9], protein abundance [10, 11], and self-tolerance [12]. Several studies have investigated the relative importance of these other factors [5, 13–16] while most conclude that they do impact the predictability of T-cell epitopes, they all agree that MHC binding is the single most selective step in T-cell antigen presentation. In this chapter, we will therefore focus only on MHC binding, and describe two state-of-the-art methods: *NetMHCcons*, for the prediction of MHC class I binding; and *NetMHCIIpan*, for the prediction of MHC class II binding.

1.1 NetMHCcons

NetMHCcons [17] is a consensus method for the predictions of peptide binding to the MHC class I that combines the predictions of three state-of-the-art methods: *NetMHC* [18], *NetMHCpan* [19], and *PickPocket* [20]. *NetMHC* is a method based on artificial neural networks and it is allele-specific, i.e., it can only produce predictions for the molecules used to train the method. In contrast, as the name also suggests, *NetMHCpan* is pan-specific, i.e., it can be applied to any MHC class I molecule of known sequence, including alleles characterized by little or no experimental binding data. Finally, *PickPocket* is a matrix-based method that relies on receptor pocket similarities between MHC molecules and it is also pan-specific.

Based on a thorough benchmark, Karosiene et al. [17] defined a set of rules to combine in an optimal manner the predictions of the three methods. In particular, binding predictions for alleles included in the training set and sufficiently large training sets (at least 50 peptides, of which at least 10 binders) achieve highest performance using a linear combination of *NetMHCpan* and *NetMHC*. With fewer available data points, only *NetMHCpan* predictions are used. When the query allele is uncharacterized (that is, the allele was not used to train the method) but there is a MHC with similar sequence in the training set, *NetMHCpan* alone performs best. And finally, predictions for an allele with no close neighbors in the training set are defined in *NetMHCcons* as a linear combination of *NetMHCpan* and *PickPocket*.

1.2 NetMHCIIpan

NetMHCIIpan [21, 22] is a pan-specific method for the quantitative prediction of peptide binding to any MHC class II molecule of known sequence. *NetMHCIIpan* was trained on a large data set of over 50,000 data points covering 24 HLA-DR, 5 HLA-DP, 6 HLA-DQ, and 2 murine H-2 molecules, but can produce predictions for any other allele if it is provided with a complete MHC protein sequence (both alpha and beta chains). *NetMHCIIpan*, based on the artificial neural network algorithm *NNAlign* [23, 24], aims at solving two problems simultaneously: prediction of peptide-MHC binding affinity, and identification of the binding core. The peptide binding core is the region of usually nine amino acids directly in contact with the MHC-binding groove and the main determinant of binding. However, it has been shown that the peptide flanking regions (PFR) on either side of the binding core can affect peptide-MHC binding and, eventually, immunogenicity [25, 26]. The size and composition of PFRs, together with the length of the peptide itself, are taken into account and encoded in the NetMHCIIpan networks.

The method provides predictions both of peptide binding affinity and of the binding core register location within each peptide. We have recently shown [24] that the identification of the binding core by neural network ensembles can be greatly improved with the employment of a network alignment procedure called “offset correction,” which was incorporated into *NetMHCIIpan* to enhance MHC class II binding core recognition [22]. Besides accurately identifying the binding core, the method assigns reliability scores to each binding core prediction and allows the quantification of the likelihood of multiple binding cores within a single antigenic peptide.

2 Methods

2.1 NetMHCcons

The *NetMHCcons* server is hosted at *http://www.cbs.dtu.dk/services/NetMHCcons*. This guide refers to version 1.1 of the server—note that the available options may vary slightly in future updated versions.

1. The server accepts input in two formats: PEPTIDE and FASTA.
   - The PEPTIDE format is simply a list of amino acid sequences (of length 8–15) to be directly interrogated as potential MHC class I binders.
   - The FASTA format is intended for scans of protein sequences for potential epitopes. The protein sequence (or multiple sequences in FASTA format) is digested into overlapping peptides of the specified length(s), which are then submitted to the algorithm for prediction.
2. Specify the peptide length (for FASTA submissions). This parameter defines the length of the peptides to be generated from the FASTA sequences. Multiple lengths, between 8 and 15, can be selected for a single submission.
3. Select the method. As described in the introduction, *NetMHCcons* is a combination of the methods *NetMHC*, *NetMHCpan*, and *PickPocket*. Besides running predictions on the optimized consensus of the three methods (*NetMHCcons*), the user can also choose to use only one of the prediction methods.
4. Select species and allele. *NetMHCcons* has a large library of MHC protein sequences that include human (HLA-A, HLA-B, HLA-C, and HLA-E), chimpanzee (Patr), rhesus macaque (Mamu), pig (SLA), mouse (H-2), gorilla (Gogo), and bovine (BoLA) MHCs. Toggling the species displays the library of alleles with a characterized MHC sequence in the *NetMHCcons* library. Multiple alleles can be selected in a single submission.
5. If the query MHC allele is not present in the *NetMHCcons* library, or is a novel/mutated molecule, *NetMHCcons* can nevertheless produce a prediction. Simply upload the complete MHC protein sequence to the corresponding window in the server.
6. Conventionally, peptides with a IC_{50} binding affinity <50 nM are defined as strong binders to the MHC, and peptides with $IC_{50} > 500$ nM as weak binders [13, 27]. Studies have demonstrated that the repertoire of presented peptides varies dramatically between MHC molecules when defined in terms of IC_{50} binding affinity [28, 29]. In contrast the %Rank provides a robust filter for the identification of MHC-binding peptides

and, depending on the study and pathogen of interest, around 95% of validated CTL epitopes bind with a Rank score less than or equal to 2% [30, 31]. The IEDB currently recommends making selections based on a Rank score <1% to cover most of the immune responses (www.iedb.org). In *NetMHcons* a peptide will be identified as a strong binder if the %Rank is below 0.5% or the binding affinity (IC_{50}) is below 50 nM. Otherwise, the peptide will be identified as a weak binder if the %Rank is below 2% or the binding affinity (IC_{50}) is below 500 nM. These thresholds can be modified by the user.

7. Filtering options. In order to limit the size of the result files, they can be filtered by predicted affinity in terms of IC_{50} or % Rank by specifying filtering thresholds in the submission page. Additionally, toggling the corresponding option allows sorting the predictions by predicted affinity.
8. Save predictions to XLS file. For a more convenient visualization of the results, they can be saved in a spreadsheet format along with the default plain text output. The XLS format comprises global statistics on the epitope search, including the MHC allele coverage (NB column) and average predicted affinity (Ave column) for each peptide.
9. Submit your job. Clicking on the Submit button will initiate the run. You may wait for the job to terminate, or enter your email address and simply leave the window. You will be notified by email when it has terminated with a link to the results page.

In Fig. 2 is shown an example of *NetMHCcons* output. In this example the 30 amino acids region between positions 180 and 209 of the Gag polyprotein from HIV virus was submitted to the program in FASTA format:

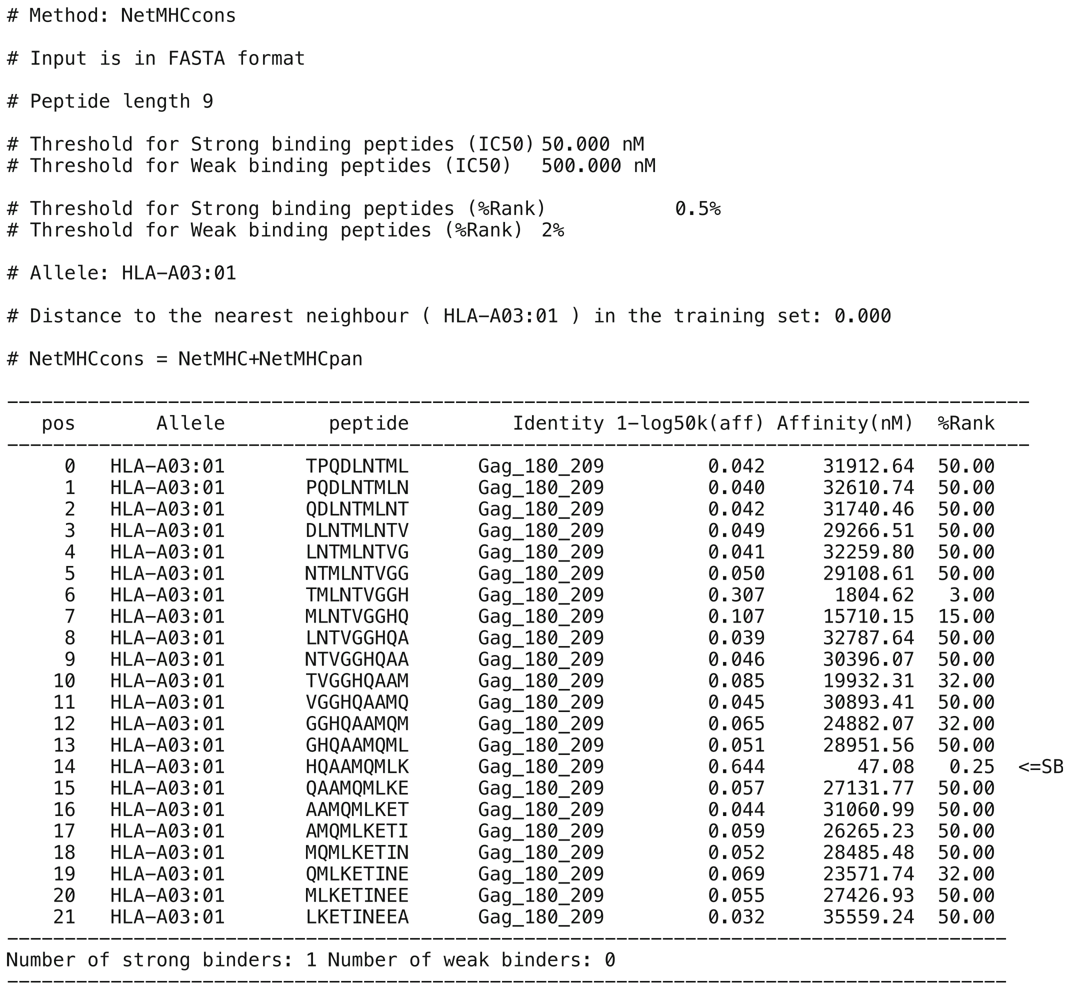
```
>Gag_180_209
TPQDLNTMLNTVGGHQAAMQMLKETINEEA
```

The peptide length was set to 9, which resulted in the digestion of the protein region into 22 overlapping peptides. The method predicts a strong binder to the allele HLA-A*03:01 corresponding to the peptide HQAAMQMLK with a predicted binding affinity IC_{50} of 47 nM and %Rank of 0.25%.

2.2 NetMHCIIpan

The *NetMHCIIpan* server is hosted at *http://www.cbs.dtu.dk/services/NetMHCIIpan*. This guide refers to version 3.1 of the server—note that the available options may vary slightly in future updated versions.

1. The server accepts input in two formats: PEPTIDE and FASTA.



```
# Method: NetMHCcons

# Input is in FASTA format

# Peptide length 9

# Threshold for Strong binding peptides (IC50) 50.000 nM
# Threshold for Weak binding peptides (IC50)   500.000 nM

# Threshold for Strong binding peptides (%Rank)          0.5%
# Threshold for Weak binding peptides (%Rank)  2%

# Allele: HLA-A03:01

# Distance to the nearest neighbour ( HLA-A03:01 ) in the training set: 0.000

# NetMHCcons = NetMHC+NetMHCpan

-----------------------------------------------------------------------------------------
   pos       Allele         peptide        Identity 1-log50k(aff) Affinity(nM)  %Rank
-----------------------------------------------------------------------------------------
     0   HLA-A03:01       TPQDLNTML      Gag_180_209         0.042     31912.64  50.00
     1   HLA-A03:01       PQDLNTMLN      Gag_180_209         0.040     32610.74  50.00
     2   HLA-A03:01       QDLNTMLNT      Gag_180_209         0.042     31740.46  50.00
     3   HLA-A03:01       DLNTMLNTV      Gag_180_209         0.049     29266.51  50.00
     4   HLA-A03:01       LNTMLNTVG      Gag_180_209         0.041     32259.80  50.00
     5   HLA-A03:01       NTMLNTVGG      Gag_180_209         0.050     29108.61  50.00
     6   HLA-A03:01       TMLNTVGGH      Gag_180_209         0.307      1804.62   3.00
     7   HLA-A03:01       MLNTVGGHQ      Gag_180_209         0.107     15710.15  15.00
     8   HLA-A03:01       LNTVGGHQA      Gag_180_209         0.039     32787.64  50.00
     9   HLA-A03:01       NTVGGHQAA      Gag_180_209         0.046     30396.07  50.00
    10   HLA-A03:01       TVGGHQAAM      Gag_180_209         0.085     19932.31  32.00
    11   HLA-A03:01       VGGHQAAMQ      Gag_180_209         0.045     30893.41  50.00
    12   HLA-A03:01       GGHQAAMQM      Gag_180_209         0.065     24882.07  32.00
    13   HLA-A03:01       GHQAAMQML      Gag_180_209         0.051     28951.56  50.00
    14   HLA-A03:01       HQAAMQMLK      Gag_180_209         0.644        47.08   0.25  <=SB
    15   HLA-A03:01       QAAMQMLKE      Gag_180_209         0.057     27131.77  50.00
    16   HLA-A03:01       AAMQMLKET      Gag_180_209         0.044     31060.99  50.00
    17   HLA-A03:01       AMQMLKETI      Gag_180_209         0.059     26265.23  50.00
    18   HLA-A03:01       MQMLKETIN      Gag_180_209         0.052     28485.48  50.00
    19   HLA-A03:01       QMLKETINE      Gag_180_209         0.069     23571.74  32.00
    20   HLA-A03:01       MLKETINEE      Gag_180_209         0.055     27426.93  50.00
    21   HLA-A03:01       LKETINEEA      Gag_180_209         0.032     35559.24  50.00
---------------------------------------------------------------------------------------
Number of strong binders: 1 Number of weak binders: 0
---------------------------------------------------------------------------------------
```

Fig. 2 Example of *NetMHCcons* output for the scan of potential MHC class I binders to HLA-A*03:01 in the region (180..209) of the Gag polyprotein from HIV virus. *NetMHCcons* predicts a strong nonamer binder (HQAAMQMLK) with predicted binding affinity $IC_{50} < 50$ nM

- The PEPTIDE format is simply a list of sequences of at least nine amino acids to be directly interrogated as potential MHC class II binders.
- The FASTA format is intended for scans of protein sequences for potential epitopes. The protein sequence (or multiple sequences in FASTA format) is digested into overlapping peptides of the specified length, which are then submitted to the algorithm for prediction.

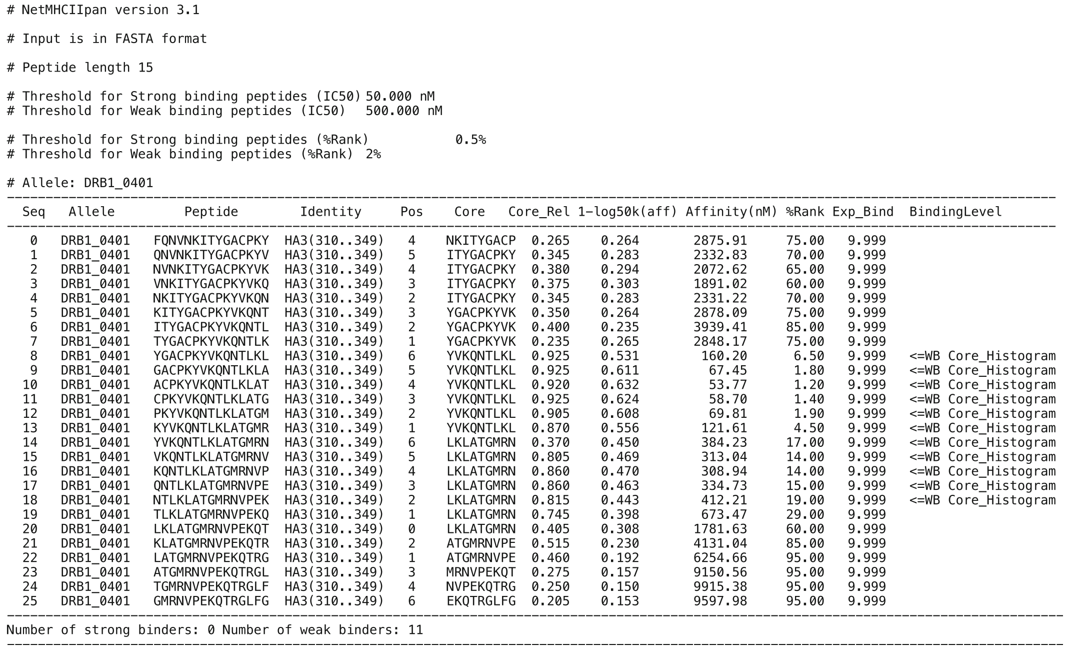
2. Specify the peptide length (for FASTA submissions). This parameter defines the length of the peptides to be generated from the FASTA sequences. By default the server uses 15mer peptides.

3. Select the species/loci and alleles. Predictions can be obtained for human HLA-DR, HLA-DP, and HLA-DQ molecules, and for H-2 mouse molecules. Selecting the species/locus displays the list of available alleles for the locus in question. As only the beta chain of HLA-DR is polymorphic, only the HLA-DRB allele should be specified. In contrast, both the alpha and beta chains of HLA-DP and HLA-DQ must be selected from the drop-down list.
4. If the MHC molecule is not in the list, or is an uncharacterized allelic variant, the user can upload the full MHC protein sequence in FASTA format. As above, only the beta chain is needed for HLA-DR molecules. For all other loci, both the alpha and beta chain should be uploaded using the dedicated boxes.
5. Optionally specify thresholds for strong and weak binders. Two different types of thresholds can be set: based on the binding affinity (in nanomolar IC_{50} values) or expressed in terms of % Rank of the prediction value relative to the background distribution of predictions on 200,000 random natural peptides. The peptide will be identified as a strong binder if the %Rank or IC_{50} affinity is below the specified threshold. The peptide will be identified as a weak binder if the %Rank or IC_{50} affinity is above the strong binder threshold but below the specified threshold for weak binders. As for MHC class I, the repertoire of presented peptides can vary dramatically between MHC molecules when defined in terms of binding affinity (IC_{50}), and it is recommended to use %Rank scores to categorize antigenic peptides. The IEDB currently recommends making selections based on a 10% rank score (www.iedb.org).
6. Filtering options. In order to limit the size of the result files, they can be filtered by predicted affinity in terms of IC_{50} or % Rank by specifying filtering thresholds in the submission page.
7. Optionally run the program in Fast mode (recommended for very large submissions), which uses a reduced ensemble of ten neural networks. It gives a faster but generally less accurate response.
8. The results of FASTA submissions can be filtered further by only displaying the strongest binding core in overlapping consecutive peptides with the same predicted core. Additionally, toggling the corresponding option allows sorting the predictions by predicted affinity.
9. Offset correction is a procedure that improves the identification of MHC class II binding cores by optimizing the combined information content of multiple networks in an ensemble [24, 32]. Excluding this step by toggling the corresponding

option reproduces the behavior of the older version (3.0) of the server for the task of binding core identification.

10. The server can produce a graphical representation of the peptide-binding core registers. For each possible register, the plot depicts the fraction of networks in the ensemble that placed the optimal core at that starting position.
11. Save predictions to XLS file. For a more convenient visualization of the results, they can be saved in a spreadsheet format along with the default plain text output. The XLS format comprises global statistics on the epitope search, including the MHC molecule coverage (NB column) and average predicted affinity (Ave column) for each peptide.
12. Submit your job. Clicking on the Submit button will initiate the run. You may wait for the job to terminate, or enter your email address and simply leave the window. You will be notified by email when it has terminated with a link to the results page.

In Fig. 3 is shown an example of *NetMHCIIpan* output. In this example, a 40 amino acids region between positions 310 and 349 of the Hemagglutinin protein serotype H3 from Influenza virus was submitted to the program in FASTA format:



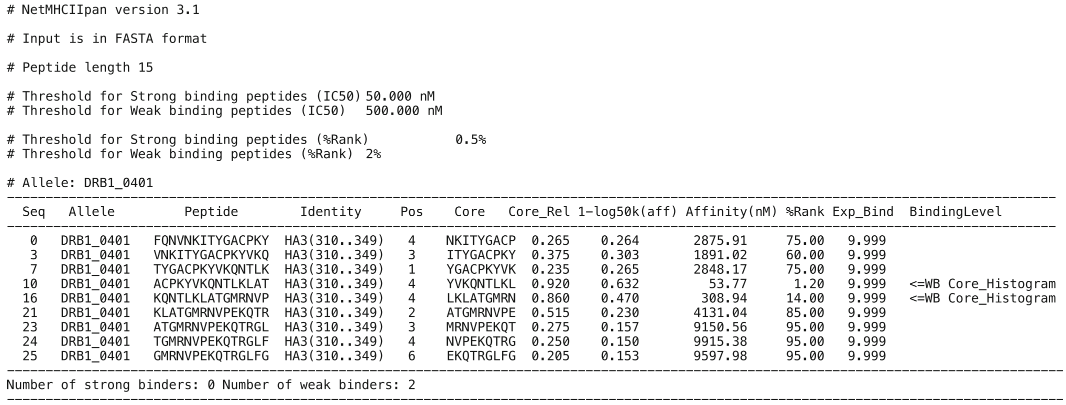
```
# NetMHCIIpan version 3.1

# Input is in FASTA format

# Peptide length 15

# Threshold for Strong binding peptides (IC50) 50.000 nM
# Threshold for Weak binding peptides (IC50)   500.000 nM

# Threshold for Strong binding peptides (%Rank)          0.5%
# Threshold for Weak binding peptides (%Rank)  2%

# Allele: DRB1_0401
-------------------------------------------------------------------------------------------------------------------------
  Seq   Allele          Peptide        Identity     Pos    Core   Core_Rel 1-log50k(aff) Affinity(nM) %Rank Exp_Bind  BindingLevel
-------------------------------------------------------------------------------------------------------------------------
   0   DRB1_0401   FQNVNKITYGACPKY  HA3(310..349)   4    NKITYGACP  0.265    0.264       2875.91     75.00   9.999
   1   DRB1_0401   QNVNKITYGACPKYV  HA3(310..349)   5    ITYGACPKY  0.345    0.283       2332.83     70.00   9.999
   2   DRB1_0401   NVNKITYGACPKYVK  HA3(310..349)   4    ITYGACPKY  0.380    0.294       2072.62     65.00   9.999
   3   DRB1_0401   VNKITYGACPKYVKQ  HA3(310..349)   3    ITYGACPKY  0.375    0.303       1891.02     60.00   9.999
   4   DRB1_0401   NKITYGACPKYVKQN  HA3(310..349)   2    ITYGACPKY  0.345    0.283       2331.22     70.00   9.999
   5   DRB1_0401   KITYGACPKYVKQNT  HA3(310..349)   3    YGACPKYVK  0.350    0.264       2878.09     75.00   9.999
   6   DRB1_0401   ITYGACPKYVKQNTL  HA3(310..349)   2    YGACPKYVK  0.400    0.235       3939.41     85.00   9.999
   7   DRB1_0401   TYGACPKYVKQNTLK  HA3(310..349)   1    YGACPKYVK  0.235    0.265       2848.17     75.00   9.999
   8   DRB1_0401   YGACPKYVKQNTLKL  HA3(310..349)   6    YVKQNTLKL  0.925    0.531        160.20      6.50   9.999   <=WB Core_Histogram
   9   DRB1_0401   GACPKYVKQNTLKLA  HA3(310..349)   5    YVKQNTLKL  0.925    0.611         67.45      1.80   9.999   <=WB Core_Histogram
  10   DRB1_0401   ACPKYVKQNTLKLAT  HA3(310..349)   4    YVKQNTLKL  0.920    0.632         53.77      1.20   9.999   <=WB Core_Histogram
  11   DRB1_0401   CPKYVKQNTLKLATG  HA3(310..349)   3    YVKQNTLKL  0.925    0.624         58.70      1.40   9.999   <=WB Core_Histogram
  12   DRB1_0401   PKYVKQNTLKLATGM  HA3(310..349)   2    YVKQNTLKL  0.905    0.608         69.81      1.90   9.999   <=WB Core_Histogram
  13   DRB1_0401   KYVKQNTLKLATGMR  HA3(310..349)   1    YVKQNTLKL  0.870    0.556        121.61      4.50   9.999   <=WB Core_Histogram
  14   DRB1_0401   YVKQNTLKLATGMRN  HA3(310..349)   6    LKLATGMRN  0.370    0.450        384.23     17.00   9.999   <=WB Core_Histogram
  15   DRB1_0401   VKQNTLKLATGMRNV  HA3(310..349)   5    LKLATGMRN  0.805    0.469        313.04     14.00   9.999   <=WB Core_Histogram
  16   DRB1_0401   KQNTLKLATGMRNVP  HA3(310..349)   4    LKLATGMRN  0.860    0.470        308.94     14.00   9.999   <=WB Core_Histogram
  17   DRB1_0401   QNTLKLATGMRNVPE  HA3(310..349)   3    LKLATGMRN  0.860    0.463        334.73     15.00   9.999   <=WB Core_Histogram
  18   DRB1_0401   NTLKLATGMRNVPEK  HA3(310..349)   2    LKLATGMRN  0.815    0.443        412.21     19.00   9.999   <=WB Core_Histogram
  19   DRB1_0401   TLKLATGMRNVPEKQ  HA3(310..349)   1    LKLATGMRN  0.745    0.398        673.47     29.00   9.999
  20   DRB1_0401   LKLATGMRNVPEKQT  HA3(310..349)   0    LKLATGMRN  0.405    0.308       1781.63     60.00   9.999
  21   DRB1_0401   KLATGMRNVPEKQTR  HA3(310..349)   2    ATGMRNVPE  0.515    0.230       4131.04     85.00   9.999
  22   DRB1_0401   LATGMRNVPEKQTRG  HA3(310..349)   1    ATGMRNVPE  0.460    0.192       6254.66     95.00   9.999
  23   DRB1_0401   ATGMRNVPEKQTRGL  HA3(310..349)   3    MRNVPEKQT  0.275    0.157       9150.56     95.00   9.999
  24   DRB1_0401   TGMRNVPEKQTRGLF  HA3(310..349)   4    NVPEKQTRG  0.250    0.150       9915.38     95.00   9.999
  25   DRB1_0401   GMRNVPEKQTRGLFG  HA3(310..349)   6    EKQTRGLFG  0.205    0.153       9597.98     95.00   9.999
-------------------------------------------------------------------------------------------------------------------------
Number of strong binders: 0 Number of weak binders: 11
-------------------------------------------------------------------------------------------------------------------------
```

Fig. 3 Example of *NetMHCIIpan* output for the scan of potential MHC class II binders to HLA-DRB1*04:01 in the region (310.349) of the Hemagglutinin H3 protein from influenza virus. A number of candidate epitopes with predicted binding affinity close to 50 nM are centered around the 9mer binding core YVKQNTLKL, with the 15mer ACPKYVKQNTLKLAT obtaining the highest predicted affinity

>HA3(310.0.349)
FQNVNKITYGACPKYVKQNTLKLATGMRNVPEKQ
TRGLFG

The peptide length was set to 15, which resulted in the digestion of the protein sequence into 26 overlapping peptides. A region spanned by eleven 15mer peptides was predicted to contain potential MHC class II binders, especially centered on the 9mer core YVKQNTLKL. The 15mer ACPKYVKQNTLKLAT obtained the highest predicted affinity of 54 nM and %Rank of 1.20%. The column Core_Rel lists the reliability scores of the core prediction, i.e., it expresses the fraction of networks in the ensemble that agreed on the identification of the optimal 9mer binding core. The clickable links Core_Histogram in the last column display plots of the reliability scores for all possible registers within the corresponding peptide.

For a more compact output, the same search can be performed with the *Print only the strongest binding core option* turned on. Using this option, the results include only the peptide with highest predicted affinity among overlapping peptides with the same predicted binding core. Figure 4 shows the results of the epitope search in the Hemagglutinin fragment described above using the strongest core option. For instance, of the six alternative peptides with predicted 9mer core YVKQNTLKL, only the 15mer having the highest predicted binding affinity to HLA-DRB1*04:01 (ACPKYVKQNTLKLAT) is included in the results.



```
# NetMHCIIpan version 3.1

# Input is in FASTA format

# Peptide length 15

# Threshold for Strong binding peptides (IC50) 50.000 nM
# Threshold for Weak binding peptides (IC50)   500.000 nM

# Threshold for Strong binding peptides (%Rank)           0.5%
# Threshold for Weak binding peptides (%Rank)  2%

# Allele: DRB1_0401
----------------------------------------------------------------------------------------------------------------------------
  Seq   Allele          Peptide        Identity     Pos    Core    Core_Rel 1-log50k(aff) Affinity(nM) %Rank Exp_Bind  BindingLevel
----------------------------------------------------------------------------------------------------------------------------
   0   DRB1_0401   FQNVNKITYGACPKY  HA3(310..349)   4    NKITYGACP  0.265    0.264        2875.91     75.00   9.999
   3   DRB1_0401   VNKITYGACPKYVKQ  HA3(310..349)   3    ITYGACPKY  0.375    0.303        1891.02     60.00   9.999
   7   DRB1_0401   TYGACPKYVKQNTLK  HA3(310..349)   1    YGACPKYVK  0.235    0.265        2848.17     75.00   9.999
  10   DRB1_0401   ACPKYVKQNTLKLAT  HA3(310..349)   4    YVKQNTLKL  0.920    0.632          53.77      1.20   9.999   <=WB Core_Histogram
  16   DRB1_0401   KQNTLKLATGMRNVP  HA3(310..349)   4    LKLATGMRN  0.860    0.470         308.94     14.00   9.999   <=WB Core_Histogram
  21   DRB1_0401   KLATGMRNVPEKQTR  HA3(310..349)   2    ATGMRNVPE  0.515    0.230        4131.04     85.00   9.999
  23   DRB1_0401   ATGMRNVPEKQTRGL  HA3(310..349)   3    MRNVPEKQT  0.275    0.157        9150.56     95.00   9.999
  24   DRB1_0401   TGMRNVPEKQTRGLF  HA3(310..349)   4    NVPEKQTRG  0.250    0.150        9915.38     95.00   9.999
  25   DRB1_0401   GMRNVPEKQTRGLFG  HA3(310..349)   6    EKQTRGLFG  0.205    0.153        9597.98     95.00   9.999
----------------------------------------------------------------------------------------------------------------------------
Number of strong binders: 0 Number of weak binders: 2
----------------------------------------------------------------------------------------------------------------------------
```

Fig. 4 Example of *NetMHCIIpan* output for the scan of potential MHC class II binders to HLA-DRB1*04:01 in the region (310..349) of the Hemagglutinin H3 protein from influenza virus, using the option of printing only the strongest binding core in overlapping consecutive peptides. Compared to the complete protein scan shown in Fig. 3, only unique binding cores are displayed in this more compact output

3 Guidelines and Remarks

1. All input sequences should be expressed in the conventional uppercase 20-letter amino acid code plus the letter X to represent unknown amino acids: A C D E F G H I K L M N P Q R S T V W Y X. The server converts all other characters to Xs.

2. Large submissions, for example in the case of several protein sequences in FASTA format interrogated on multiple MHC alleles, can generate output of considerable size. Because only a small fraction of peptides can usually bind to the MHC, the majority of these results will relate to predicted non-binders. In order to limit the size of the result files, they can be filtered by predicted affinity in terms of IC_{50} or %Rank by specifying filtering thresholds in the submission page.

3. The core reliability plots in *NetMHCIIpan* can be made only for a maximum of 20 peptides. Using the graphics together with the sorting option is generally a good idea in order to display the plots for the strongest predicted binders.

4. The predicted binding affinity distribution in IC_{50} can vary greatly between different alleles. In other words, at the same threshold of IC_{50} affinity certain MHC molecules will have a large number of binders whereas other molecules will have few or none. If we assume that fraction of binding peptides is approximately the same for most molecules, then the %Rank is a more reliable quantity to identify predicted binders, as it is independent of the distribution of affinities. The Immune Epitope Database (IEDB) [33] recommends selecting candidate epitopes based on a Rank score <1% for MHC class I and Rank score <10% for MHC class II to cover most of the immune responses.

5. *NetMHCcons* is trained only on 9mer peptide data. Predictions for peptides of length different from nine are extrapolated using an approximation that conforms longer and shorter peptides to a series of 9mers [34]. Predictions for peptides of length different from nine, especially very long peptides (12mers and longer) should be therefore taken with caution.

6. Stand-alone software packages for both *NetMHCcons* and *NetMHCIIpan* are available for download for academic users on the servers web pages.

7. While binding affinity to MHC molecules is the single most selective event in the T-cell antigen presentation pathways, other factors have been demonstrated to impact the likelihood of a peptide becoming a T-cell epitope. Several prediction tools have been developed to incorporate these factors into the

antigen selection pipeline. Some of these are listed below (all available at www.cbs.dtu.dk/services):

(a) *NetChop* [7]: Prediction of proteasomal cleavage;

(b) *NetCTLpan* [15]/*NetCTL* [5]: Integration of peptide-MHC class I binding, proteasomal C terminal cleavage, and TAP transport efficiency for the prediction of CTL epitopes;

(c) *NetTepi* [35]: Integration of peptide-MHC-binding affinity, peptide-MHC stability, and T-cell propensity for the prediction of CTL epitopes;

(d) *NetMHCstab* [36]: Prediction of stability of peptide-MHC class I complexes.

Acknowledgments

This study was supported in part with Federal funds from the National Institutes of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272201200010C. This work was partially funded by the Agencia Nacional de Promoción Científica y Tecnológica, Argentina (PICT-2012-0115). MN is a researcher at the Argentinean national research council (CONICET).

References

1. Germain RN (1994) MHC-dependent antigen processing and peptide presentation: providing ligands for T lymphocyte activation. Cell 76:287–299
2. Rudolph MG, Stanfield RL, Wilson IA (2006) How TCRs bind MHCs, peptides, and coreceptors. Annu Rev Immunol 24:419–466. https://doi.org/10.1146/annurev.immunol.23.021704.115658
3. Rudensky AY, Preston-Hurlburt P, Hong SC et al (1991) Sequence analysis of peptides bound to MHC class II molecules. Nature 353:622–627. https://doi.org/10.1038/353622a0
4. Lundegaard C, Hoof I, Lund O, Nielsen M (2010) State of the art and challenges in sequence based T-cell epitope prediction. Immunome Res 6:S3. https://doi.org/10.1186/1745-7580-6-S2-S3
5. Larsen M, Lundegaard C, Lamberth K (2005) An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. Eur J Immunol 35:2295–2303
6. Assarsson E, Sidney J, Oseroff C et al (2007) A quantitative analysis of the variables affecting the repertoire of T cell specificities recognized after vaccinia virus infection. J Immunol 178:7890–7901
7. Nielsen M, Lundegaard C, Lund O, Keşmir C (2005) The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. Immunogenetics 57:33–41. https://doi.org/10.1007/s00251-005-0781-7
8. Harndahl M, Rasmussen M, Roder G, Buus S (2011) Real-time, high-throughput measurements of peptide-MHC-I dissociation using a scintillation proximity assay. J Immunol Methods 374:5–12. https://doi.org/10.1016/j.jim.2010.10.012
9. Harndahl M, Rasmussen M, Roder G et al (2012) Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. Eur J Immunol 42:1405–1416. https://doi.org/10.1002/eji.201141774
10. Juncker AS, Larsen MV, Weinhold N et al (2009) Systematic characterisation of cellular localisation and expression profiles of proteins

containing MHC ligands. PLoS One 4:e7448. https://doi.org/10.1371/journal.pone.0007448

11. Hoof I, van Baarle D, Hildebrand WH, Keşmir C (2012) Proteome sampling by the HLA class I antigen processing pathway. PLoS Comput Biol 8:e1002517. https://doi.org/10.1371/journal.pcbi.1002517
12. Frankild S, de Boer RJ, Lund O et al (2008) Amino acid similarity accounts for T cell cross-reactivity and for "holes" in the T cell repertoire. PLoS One 3:e1831. https://doi.org/10.1371/journal.pone.0001831
13. Yewdell JW, Bennink JR (1999) Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. Annu Rev Immunol 17:51–88. https://doi.org/10.1146/annurev.immunol.17.1.51
14. Tenzer S, Peters B, Bulik S et al (2005) Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. Cell Mol Life Sci CMLS 62:1025–1037. https://doi.org/10.1007/s00018-005-4528-2
15. Stranzl T, Larsen MV, Lundegaard C, Nielsen M (2010) NetCTLpan: pan-specific MHC class I pathway epitope predictions. Immunogenetics 62:357–368. https://doi.org/10.1007/s00251-010-0441-4
16. Doytchinova IA, Guan P, Flower DR (2006) EpiJen: a server for multistep T cell epitope prediction. BMC Bioinformatics 7:131. https://doi.org/10.1186/1471-2105-7-131
17. Karosiene E, Lundegaard C, Lund O, Nielsen M (2012) NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. Immunogenetics 64:177–186. https://doi.org/10.1007/s00251-011-0579-8
18. Lundegaard C, Lamberth K, Harndahl M et al (2008) NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res 36:W509–W512. https://doi.org/10.1093/nar/gkn202
19. Hoof I, Peters B, Sidney J et al (2009) NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics 61:1–13. https://doi.org/10.1007/s00251-008-0341-z
20. Zhang H, Lund O, Nielsen M (2009) The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. Bioinforma Oxf Engl 25:1293–1299. https://doi.org/10.1093/bioinformatics/btp137
21. Karosiene E, Rasmussen M, Blicher T et al (2013) 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 65:711–724. https://doi.org/10.1007/s00251-013-0720-y
22. Andreatta M, Karosiene E, Rasmussen M et al (2015) Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. Immunogenetics 67(11-12):641–650
23. Nielsen M, Lund O (2009) NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics 10:296
24. Andreatta M, Schafer-Nielsen C, Lund O et al (2011) NNAlign: a web-based prediction method allowing non-expert end-user discovery of sequence motifs in quantitative peptide data. PLoS One 6:e26781. https://doi.org/10.1371/journal.pone.0026781
25. Godkin AJ, Smith KJ, Willis A et al (2001) Naturally processed HLA class II peptides reveal highly conserved immunogenic flanking region sequence preferences that reflect antigen processing rather than peptide-MHC interactions. J Immunol 166:6720–6727
26. Carson RT, Vignali KM, Woodland DL, Vignali DA (1997) T cell receptor recognition of MHC class II-bound peptide flanking residues enhances immunogenicity and results in altered TCR V region usage. Immunity 7:387–399
27. Moutaftsi M, Peters B, Pasquetto V et al (2006) A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. Nat Biotechnol 24:817–819. https://doi.org/10.1038/nbt1215
28. Rao X, Costa AI, van Baarle D, Kesmir C (2009) A comparative study of HLA binding affinity and ligand diversity: implications for generating immunodominant CD8+ T cell responses. J Immunol 182:1526–1532
29. Paul S, Weiskopf D, Angelo MA et al (2013) HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. J Immunol 191:5831–5839. https://doi.org/10.4049/jimmunol.1302101
30. Erup Larsen M, Kloverpris H, Stryhn A et al (2011) HLArestrictor--a tool for patient-specific predictions of HLA restriction elements and optimal epitopes within peptides. Immunogenetics 63:43–55. https://doi.org/10.1007/s00251-010-0493-5

31. Braendstrup P, Mortensen BK, Justesen S et al (2014) Identification and HLA-tetramer-validation of human CD4+ and CD8+ T cell responses against HCMV proteins IE1 and IE2. PLoS One 9:e94892. https://doi.org/10.1371/journal.pone.0094892

32. Andreatta M, Nielsen M (2012) Characterizing the binding motifs of 11 common human HLA-DP and HLA-DQ molecules using NNAlign. Immunology 136:306–311. https://doi.org/10.1111/j.1365-2567.2012.03579.x

33. Vita R, Overton JA, Greenbaum JA et al (2015) The immune epitope database (IEDB) 3.0. Nucleic Acids Res 43:D405–D412. https://doi.org/10.1093/nar/gku938

34. Lundegaard C, Lund O, Nielsen M (2008) Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. Bioinformatics 24:1397–1398. https://doi.org/10.1093/bioinformatics/btn128

35. Trolle T, Nielsen M (2014) NetTepi: an integrated method for the prediction of T cell epitopes. Immunogenetics 66:449–456. https://doi.org/10.1007/s00251-014-0779-0

36. Jørgensen KW, Rasmussen M, Buus S, Nielsen M (2014) NetMHCstab - predicting stability of peptide-MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. Immunology 141:18–26. https://doi.org/10.1111/imm.12160