Predicting Binding Affinities of MHC Class II Epitopes Across Alleles

Nico Pfeifer and Oliver Kohlbacher
Center for Bioinformatics
Eberhard Karls University Tübingen
Roadmap

• **Background**
  – Major Histocompatibility Complex Class II (MHCII)
  – Problem Definition

• **Methods**
  – Multiple Instance Learning
  – Normalized Set Kernel
  – Positionally-weighted RBF Kernel

• **Results**
MHCII

- Major Histocompatibility Complex class II (MHCII)
Background

MHCII

- Major Histocompatibility Complex class II (MHCII)
- Part of the adaptive immune system
MHCII

- Major Histocompatibility Complex class II (MHCII)
- Part of the adaptive immune system
- Can be found primarily on antigen-presenting cells (macrophages, dendritic cells and B-cells)
MHCII

- Major Histocompatibility Complex class II (MHCII)
- Part of the adaptive immune system
- Can be found primarily on antigen-presenting cells (macrophages, dendritic cells and B-cells)
- Presents peptides which are derived from exogenous proteins to trigger an immune response
Background

MHCII

adapted from: More than one reason to rethink the use of peptides in vaccine design
Anthony W. Purcell, James McCluskey & Jamie Rossjohn
Nature Reviews Drug Discovery 6, 404-414 (May 2007)
Why is MHCII interesting?

- Every human has up to eight different MHCII alleles
Why is MHCII interesting?

• Every human has up to eight different MHCII alleles

• Every corresponding MHCII molecule binds to a different set of peptides
Why is MHCII interesting?

- Every human has up to eight different MHCII alleles
- Every corresponding MHCII molecule binds to a different set of peptides
- Peptide vaccines can only activate immune responses if they bind to the MHC molecule
Why is MHCII interesting?

• Every human has up to eight different MHCII alleles

• Every corresponding MHCII molecule binds to a different set of peptides

• Peptide vaccines can only activate immune responses if they bind to the MHC molecule

⇒ Personalized vaccine design
Why is MHCII interesting?

- Every human has up to eight different MHCII alleles
- Every corresponding MHCII molecule binds to a different set of peptides
- Peptide vaccines can only activate immune responses if they bind to the MHC molecule

⇒ Personalized vaccine design
⇒ Good predictors for peptide binding needed
Definitions

Peptide-MHCII binding prediction

Given: Peptide sequence $s$ and sequence $m$ of an MHCII molecule

Problem: Decide whether $s$ binds to $m$ or not
Definitions

**Peptide-MHCII binding prediction**

Given: Peptide sequence $s$ and sequence $m$ of an MHCII molecule

Problem: Decide whether $s$ binds to $m$ or not

**Peptide-MHCII binding affinity prediction**

Given: Peptide sequence $s$ and sequence $m$ of an MHCII molecule

Problem: Decide how strong $s$ binds to $m$
Background

MHCII

- More than 700 different MHCII alleles are known [1]

MHCII

• More than 700 different MHCII alleles are known [1]

• For less than 3% there is sufficient experimental data from binding studies

MHCII

• More than 700 different MHCII alleles are known [1]

• For less than 3% there is sufficient experimental data from binding studies

• Problems of peptide-MHCII binding and affinity prediction are not solved yet

Background

What makes the problems difficult?

• Varying peptide length since binding clefts are open
What makes the problems difficult?

- Varying peptide length since binding clefts are open
- The binding core (9 residues) is, in most of the cases, unknown
Related work

- Positional Scoring Matrices
- Gibbs Samplers
- Hidden Markov Models
- Artificial Neural Networks
- Partial Least Squares
- Support Vector Machines
- Consensus approach

For more information on related work have a look at the paper
N. Pfeifer and O. Kohlbacher: Multiple Instance Learning Allows MHC Class II Epitope Predictions across Alleles
Lecture Notes in Bioinformatics: Proceedings of WABI 2008
Related work

- Positional Scoring Matrices
- Gibbs Samplers
- Hidden Markov Models
- Artificial Neural Networks
- Partial Least Squares
- Support Vector Machines
- Consensus approach

Predictors only for a small fraction of all MHCII alleles

For more information on related work have a look at the paper
N. Pfeifer and O. Kohlbacher: Multiple Instance Learning Allows MHC Class II Epitope Predictions across Alleles
Lecture Notes in Bioinformatics: Proceedings of WABI 2008
Roadmap

• **Background**
  – Major Histocompatibility Complex Class II (MHCII)
  – Problem Definition

• **Methods**
  – Multiple Instance Learning
  – Normalized Set Kernel
  – Positionally-weighted RBF Kernel

• **Results**
Peptide-MHCII binding affinity prediction

- A core part of the peptide (9 amino acids long) binds to the MHCII molecule
Peptide-MHCII binding affinity prediction

• A core part of the peptide (9 amino acids long) binds to the MHCII molecule
• First idea: represent sequence $s$ by a bag $X$ containing all 9-mers of $s$
Peptide-MHCII binding affinity prediction

- A core part of the peptide (9 amino acids long) binds to the MHCII molecule
- First idea: represent sequence $s$ by a bag $X$ containing all 9-mers of $s$
  $\Rightarrow$ Lots of noise in the bags
Peptide-MHCII binding affinity prediction

- A core part of the peptide (9 amino acids long) binds to the MHCII molecule
- First idea: represent sequence \( s \) by a bag \( X \) containing all 9-mers of \( s \)
  \[ \Rightarrow \text{Lots of noise in the bags} \]
- Better idea: represent sequence \( s \) by a bag \( X \) containing all reasonable \([2,3]\) putative binding cores: aromatic (F,W,Y) or aliphatic (I,L,M,V) amino acid at position 1

Peptide-MHCII binding affinity prediction

ENPVVHFFKNIVTPR

VVHFFKNIV
VHFFKNIVT
FFKNIVTPR
Peptide-MHCII binding affinity prediction

Represent sequence ENPVVVHFFKNIIVTPR by bag

\[ X = \{VVHFFKNIV, VHFFKNIVT, FFKNIVTPR\} \]
Multiple Instance Learning

• Binary classification:

\[ S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) | x_i \in \mathcal{X} \land y_i \in \{-1, 1\} \ \forall i = 1, 2, \ldots, n\} \]
Multiple Instance Learning

• Binary classification:

\[ S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) | x_i \in X \land y_i \in \{-1, 1\} \ \forall i = 1, 2, \ldots, n\} \]

• Multiple instance classification [4]:

\[ S = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n) | X_i \subseteq X \land y_i \in \{-1, 1\} \ \forall i = 1, 2, \ldots, n\} \]

Multiple Instance Learning

- Binary classification:
  \[ S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) | x_i \in \mathcal{X} \land y_i \in \{-1, 1\} \ \forall i = 1, 2, \ldots, n\} \]

- Multiple instance classification [4]:
  \[ S = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n) | X_i \subseteq \mathcal{X} \land y_i \in \{-1, 1\} \ \forall i = 1, 2, \ldots, n\} \]

- Multiple instance regression [5,6]:
  \[ S = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n) | X_i \subseteq \mathcal{X} \land y_i \in \mathbb{R} \ \forall i = 1, 2, \ldots, n\} \]

Normalized Set Kernel [7]

\[ k(X, X') := \frac{\sum_{x \in X, x' \in X'} k_X(x, x')} {f_{\text{norm}}(X)f_{\text{norm}}(X')} \]

with \( k_X \) being a kernel on \( X \) and

\[ f_{\text{norm}}(X) = \sqrt{\sum_{x \in X, x' \in X} k_X(x, x')} \]
Single Predictor

- Create a bag $X$ for every sequence of the training set
Single Predictor

• Create a bag $X$ for every sequence of the training set

• Encode the $x \in X$ reflecting physicochemical properties [9]
Single Predictor

- Create a bag $X$ for every sequence of the training set
- Encode the $x \in X$ reflecting physicochemical properties [9]
- Perform $\nu$-Support Vector Regression ($\nu$-SVR) with the normalized set kernel

$$
k(X, X') := \sum_{x \in X, x' \in X'} \frac{k_X(x, x')}{f_{\text{norm}}(X) f_{\text{norm}}(X')}
$$

with $k_X$ being the RBF kernel

MHCIIIMulti

- *Single Predictor* can only be used if there is sufficient experimental binding data for the target allele (~3% of all MHCII alleles)
MHCIIMulti

• *Single Predictor* can only be used if there is sufficient experimental binding data for the target allele (~3% of all MHCII alleles)

• Predictor for more alleles desirable
MHCIIMulti

- *Single Predictor* can only be used if there is sufficient experimental binding data for the target allele (~3% of all MHCII alleles)

- Predictor for more alleles desirable

- HLA-DRB1, HLA-DRB3, HLA-DRB4 and HLA-DRB5 alleles differ only at very few positions
Methods

Pocket 1

MHCII molecule

Peptide
Methods

MHCII molecule

Peptide
Methods

Pocket 4
Methods

- MHCII molecule
- Peptide
Methods

Pocket 6
Methods

- MHCII molecule
- Peptide
Methods

Pocket 7
Methods

- MHCII molecule
- Peptide
Pocket 9
It was shown that alleles which have the same amino acids in a pocket region show similar binding affinities for the corresponding amino acid of the peptide.
Covering more alleles

- Sturniolo *et al.* defined matrices for only ~7% of alleles
Covering more alleles

• Sturniolo et al. defined matrices for only ~7% of alleles

• More than 700 alleles known and more than 2/3 are HLA-DRB1, HLA-DRB3, HLA-DRB4 or HLA-DRB5 alleles
Covering more alleles

• Sturniolo et al. defined matrices for only ~7% of alleles

• More than 700 alleles known and more than 2/3 are HLA-DRB1, HLA-DRB3, HLA-DRB4 or HLA-DRB5 alleles

• Take alignments of the IMGT/HLA database [1] to get amino acids of the pocket regions for these alleles

Covering more alleles

• Sturniolo et al. defined matrices for only ~7% of alleles

• More than 700 alleles known and more than 2/3 are HLA-DRB1, HLA-DRB3, HLA-DRB4 or HLA-DRB5 alleles

• Take alignments of the IMGT/HLA database [1] to get amino acids of the pocket regions for these alleles

• Calculate a mean pocket vector for pockets 1, 4, 6, 7, and 9 using the same encoding as for the peptides

\[ p = \left[ p_1^T, p_4^T, p_6^T, p_7^T, p_9^T \right]^T \]

Multitask learning for MHC class I peptide binding [10]

• Kernel-based prediction of peptide MHC class I binding prediction:

\[ K((x, a), (x', a')) = K_{pep}(x, x') \times K_{all}(a, a') \]

– Simplifying assumption: all binding core positions are equally important

Multitask learning for MHC class I peptide binding [10]

• Kernel-based prediction of peptide MHC class I binding prediction:

\[ K((x, a), (x', a')) = K_{pep}(x, x') \times K_{all}(a, a') \]

  – Simplifying assumption: all binding core positions are equally important

• Our approach: weight positions according to similarities of the binding pockets

\[ p = [p_1^T, p_4^T, p_6^T, p_7^T, p_9^T]^T \]

Positionally-weighted RBF kernel

• Increase weights of important positions
Positionally-weighted RBF kernel

- Increase weights of important positions

- If sequences differ at these positions, there will be a greater effect using this kernel than with the standard RBF kernel:

\[ k_{pw-RBF}((p, x), (p', x')) = \exp\left(-\frac{\sum w_i \times \|x_i - x_i'\|^2}{2\sigma^2}\right) \]
Positionally-weighted RBF kernel

- Increase weights of important positions

- If sequences differ at these positions, there will be a greater effect using this kernel than with the standard RBF kernel:

$$k_{pw-RBF}((p, x), (p', x')) = \exp \left( -\frac{w_1 \times \|x_1 - x'_1\|^2 + \ldots + w_9 \times \|x_9 - x'_9\|^2}{2 \sigma^2} \right)$$

with

$$w_i := \begin{cases} 
\text{Pearson}(p_i, p'_i) + 1 & \text{if } i = 1, 4, 6, 7, 9 \\
0.5 & \text{otherwise}
\end{cases}$$
Predictor which uses data from multiple alleles ($MHCIIMulti$)

- Create a bag $x$ for every sequence of the training set
Predictor which uses data from multiple alleles (*MHC\textit{II}Multi*)

- Create a bag $X$ for every sequence of the training set

- Encode $x \in X$ and $p$ reflecting physicochemical properties [9]
Methods

Predictor which uses data from multiple alleles (*MHCII*Multi)

- Create a bag \( X \) for every sequence of the training set

- Encode \( x \in X \) and \( p \) reflecting physicochemical properties [9]

- Perform \( \nu \)-SVR with the normalized set kernel

\[
k((p, X), (p', X')) = \frac{\sum_{x \in X, x' \in X'} k_X((p, x), (p', x'))}{f_{\text{norm}}((p, X))f_{\text{norm}}((p', X'))}
\]

with \( k_X \) being the pw-RBF kernel

Roadmap

• Background
  – Major Histocompatibility Complex Class II (MHCII)
  – Problem Definition

• Methods
  – Multiple Instance Learning
  – Normalized Set Kernel
  – Positionally-weighted RBF Kernel

• Results
Benchmark

- Recent study to compare peptide-MHCII binding predictors [11] on a new benchmark dataset

Benchmark

- Recent study to compare peptide-MHCII binding predictors [11] on a new benchmark dataset

- Performance comparison on all human MHCII alleles of benchmark to our methods (14 alleles, 9478 peptides)

Benchmark

• Recent study to compare peptide-MHCII binding predictors [11] on a new benchmark dataset

• Performance comparison on all human MHCII alleles of benchmark to our methods (14 alleles, 9478 peptides)

• ARB method performance measured by ten-fold CV
  ⇒ Single Predictor performance measured by ten-fold CV
  ⇒ MHCIIMulti performance measured by ten-fold CV

Recent study to compare peptide-MHCII binding predictors [11] on a new benchmark dataset

Performance comparison on all human MHCII alleles of benchmark to our methods (14 alleles, 9478 peptides)

ARB method performance measured by ten-fold CV
⇒ Single Predictor performance measured by ten-fold CV
⇒ MHCIIMulti performance measured by ten-fold CV

Performance measure: area under ROC curve

## Results

### Comparison

<table>
<thead>
<tr>
<th>Allele</th>
<th>ARB</th>
<th>ProPred</th>
<th>SMM-align</th>
<th>Consensus</th>
<th>Single Predictor</th>
<th>MHCIIMulti</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRB1*0101</td>
<td>0.76</td>
<td>0.74</td>
<td>0.77</td>
<td>0.79</td>
<td>0.81</td>
<td>0.75</td>
</tr>
<tr>
<td>DRB1*0301</td>
<td>0.66</td>
<td>0.65</td>
<td>0.69</td>
<td>0.72</td>
<td>0.73</td>
<td>0.72</td>
</tr>
<tr>
<td>DRB1*0401</td>
<td>0.67</td>
<td>0.69</td>
<td>0.68</td>
<td>0.69</td>
<td>0.67</td>
<td>0.78</td>
</tr>
<tr>
<td>DRB1*0404</td>
<td>0.72</td>
<td>0.79</td>
<td>0.75</td>
<td>0.80</td>
<td>0.79</td>
<td>0.80</td>
</tr>
<tr>
<td>DRB1*0405</td>
<td>0.67</td>
<td>0.75</td>
<td>0.69</td>
<td>0.72</td>
<td>0.83</td>
<td>0.79</td>
</tr>
<tr>
<td>DRB1*0701</td>
<td>0.69</td>
<td>0.78</td>
<td>0.78</td>
<td>0.83</td>
<td>0.82</td>
<td>0.90</td>
</tr>
<tr>
<td>DRB1*0802</td>
<td>0.74</td>
<td>0.77</td>
<td>0.75</td>
<td>0.82</td>
<td>0.76</td>
<td>0.79</td>
</tr>
<tr>
<td>DRB1*0901</td>
<td>0.62</td>
<td>-</td>
<td>0.66</td>
<td>0.68</td>
<td>0.64</td>
<td>0.66</td>
</tr>
<tr>
<td>DRB1*1101</td>
<td>0.73</td>
<td>0.80</td>
<td>0.81</td>
<td>0.80</td>
<td>0.85</td>
<td>0.87</td>
</tr>
<tr>
<td>DRB1*1302</td>
<td>0.79</td>
<td>0.58</td>
<td>0.69</td>
<td>0.73</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>DRB1*1501</td>
<td>0.70</td>
<td>0.72</td>
<td>0.74</td>
<td>0.72</td>
<td>0.72</td>
<td>0.75</td>
</tr>
<tr>
<td>DRB3*0101</td>
<td>0.59</td>
<td>-</td>
<td>0.68</td>
<td>-</td>
<td>0.72</td>
<td>0.57</td>
</tr>
<tr>
<td>DRB4*0101</td>
<td>0.74</td>
<td>-</td>
<td>0.71</td>
<td>0.74</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td>DRB5*0101</td>
<td>0.70</td>
<td>0.79</td>
<td>0.75</td>
<td>0.79</td>
<td>0.81</td>
<td>0.90</td>
</tr>
<tr>
<td><strong>mean</strong></td>
<td>0.70</td>
<td>0.73</td>
<td>0.73</td>
<td>0.76</td>
<td>0.76</td>
<td>0.77</td>
</tr>
</tbody>
</table>
## Comparison

<table>
<thead>
<tr>
<th>Allele</th>
<th>ARB</th>
<th>ProPred</th>
<th>SMM-align</th>
<th>Consensus</th>
<th>Single Predictor</th>
<th>MHCIIIMulti</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRB1*0101</td>
<td>0.76</td>
<td>0.74</td>
<td>0.77</td>
<td>0.79</td>
<td>0.81</td>
<td>0.75</td>
</tr>
<tr>
<td>DRB1*0301</td>
<td>0.66</td>
<td>0.65</td>
<td>0.69</td>
<td>0.72</td>
<td>0.73</td>
<td>0.72</td>
</tr>
<tr>
<td>DRB1*0401</td>
<td>0.67</td>
<td>0.69</td>
<td>0.68</td>
<td>0.69</td>
<td><strong>0.67</strong></td>
<td><strong>0.78</strong></td>
</tr>
<tr>
<td>DRB1*0404</td>
<td>0.72</td>
<td>0.79</td>
<td>0.75</td>
<td>0.80</td>
<td>0.79</td>
<td>0.80</td>
</tr>
<tr>
<td>DRB1*0405</td>
<td>0.67</td>
<td>0.75</td>
<td>0.69</td>
<td>0.72</td>
<td>0.83</td>
<td>0.79</td>
</tr>
<tr>
<td>DRB1*0701</td>
<td>0.69</td>
<td>0.78</td>
<td>0.78</td>
<td>0.83</td>
<td>0.82</td>
<td>0.90</td>
</tr>
<tr>
<td>DRB1*0802</td>
<td>0.74</td>
<td>0.77</td>
<td>0.75</td>
<td>0.82</td>
<td>0.76</td>
<td>0.79</td>
</tr>
<tr>
<td>DRB1*0901</td>
<td>0.62</td>
<td>-</td>
<td><strong>0.66</strong></td>
<td>0.68</td>
<td><strong>0.64</strong></td>
<td><strong>0.66</strong></td>
</tr>
<tr>
<td>DRB1*1101</td>
<td>0.73</td>
<td>0.80</td>
<td>0.81</td>
<td>0.80</td>
<td>0.85</td>
<td>0.87</td>
</tr>
<tr>
<td>DRB1*1302</td>
<td>0.79</td>
<td>0.58</td>
<td>0.69</td>
<td>0.73</td>
<td>0.74</td>
<td>0.73</td>
</tr>
<tr>
<td>DRB1*1501</td>
<td>0.70</td>
<td>0.72</td>
<td>0.74</td>
<td>0.72</td>
<td>0.72</td>
<td>0.75</td>
</tr>
<tr>
<td>DRB3*0101</td>
<td>0.59</td>
<td>-</td>
<td>0.68</td>
<td>-</td>
<td>0.72</td>
<td>0.57</td>
</tr>
<tr>
<td>DRB4*0101</td>
<td>0.74</td>
<td>-</td>
<td>0.71</td>
<td>0.74</td>
<td>0.79</td>
<td>0.78</td>
</tr>
<tr>
<td>DRB5*0101</td>
<td>0.70</td>
<td>0.79</td>
<td>0.75</td>
<td>0.79</td>
<td>0.81</td>
<td>0.90</td>
</tr>
<tr>
<td><strong>mean</strong></td>
<td>0.70</td>
<td>0.73</td>
<td>0.73</td>
<td>0.76</td>
<td>0.76</td>
<td>0.77</td>
</tr>
</tbody>
</table>

### Results

Our methods
Leave-One-Allele-Out (LOAO) Prediction

- Train Leave-One-Allele-Out (LOAO) version of \textit{MHCIIMulti} on binding data of all alleles but the target allele (10 times randomly drawn subset)
Leave-One-Allele-OUT (LOAO) Prediction

- Train Leave-One-Allele-OUT (LOAO) version of MHCIIIMulti on binding data of all alleles but the target allele (10 times randomly drawn subset)

- Build aggregating predictor MHCIIIMulti* from all 10 random draws (mean prediction value)
Leave-One-Allele-Out (LOAO) Prediction

• Train Leave-One-Allele-Out (LOAO) version of \textit{MHCIIMulti} on binding data of all alleles but the target allele (10 times randomly drawn subset)

• Build aggregating predictor \textit{MHCIIMulti*} from all 10 random draws (mean prediction value)

• Compare to \textit{Single Predictor} of most similar allele

⇒ Measure performance on the target allele
### Comparison

<table>
<thead>
<tr>
<th>Allele</th>
<th>Single Predictor Trained on Nearest Neighbor</th>
<th>LOAO MHCII Multi</th>
<th>LOAO MHCII Multi*</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRB1*0101</td>
<td>0.74</td>
<td>0.64</td>
<td>0.69</td>
</tr>
<tr>
<td>DRB1*0301</td>
<td>0.61</td>
<td>0.69</td>
<td>0.70</td>
</tr>
<tr>
<td>DRB1*0401</td>
<td>0.64</td>
<td>0.76</td>
<td>0.78</td>
</tr>
<tr>
<td>DRB1*0404</td>
<td>0.70</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td>DRB1*0405</td>
<td>0.78</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>DRB1*0701</td>
<td>0.72</td>
<td>0.89</td>
<td>0.91</td>
</tr>
<tr>
<td>DRB1*0802</td>
<td>0.71</td>
<td>0.77</td>
<td>0.79</td>
</tr>
<tr>
<td>DRB1*0901</td>
<td>0.57</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td>DRB1*1101</td>
<td>0.79</td>
<td>0.87</td>
<td>0.9</td>
</tr>
<tr>
<td>DRB1*1302</td>
<td>0.62</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>DRB1*1501</td>
<td>0.66</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td>DRB3*0101</td>
<td>0.51</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>DRB4*0101</td>
<td>0.77</td>
<td>0.69</td>
<td>0.72</td>
</tr>
<tr>
<td>DRB5*0101</td>
<td>0.73</td>
<td>0.89</td>
<td>0.92</td>
</tr>
<tr>
<td>mean</td>
<td>0.68</td>
<td>0.74</td>
<td>0.76</td>
</tr>
</tbody>
</table>
Summary

• Methods for peptide-MHCII binding affinity prediction perform quite well on alleles with sufficient data
Summary

• Methods for peptide-MHCII binding affinity prediction perform quite well on alleles with sufficient data

• Presentation of method for building predictor for about 2/3 of all MHCII alleles (instead of ~7%)
Summary

- Methods for peptide-MHCII binding affinity prediction perform quite well on alleles with sufficient data.

- Presentation of method for building predictor for about 2/3 of all MHCII alleles (instead of ~7%).

- LOAO performance results are quite promising.
  \[\Rightarrow\] Can be assumed that the methods perform well for alleles for which no experimental data exist.
Summary

• Methods for peptide-MHCII binding affinity prediction perform quite well on alleles with sufficient data

• Presentation of method for building predictor for about 2/3 of all MHCII alleles (instead of ~7%)

• LOAO performance results are quite promising
  ⇒ Can be assumed that the methods perform well for alleles for which no experimental data exist

• Methods are integrated into the EpiToolKit [12] (http://www.epitoolkit.org/mhciimulti)

Thanks to

... Oliver Kohlbacher

... Lena Feldhahn and Philipp Thiel

... SFB 685, project B1

... you for listening
## Comparison to NetMHCIIPan

<table>
<thead>
<tr>
<th>Allele</th>
<th>Number of training samples (MHCIIMulti*)</th>
<th>MHCIIMulti*</th>
<th>Number of training samples (NetMHCIIPan)</th>
<th>NetMHCIIPan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1_0101</td>
<td>4430</td>
<td>0.670</td>
<td>9441</td>
<td>0.778</td>
</tr>
<tr>
<td>1_0401</td>
<td>4609</td>
<td>0.765</td>
<td>13583</td>
<td>0.775</td>
</tr>
<tr>
<td>1_0405</td>
<td>4595</td>
<td>0.777</td>
<td>13977</td>
<td>0.808</td>
</tr>
<tr>
<td>1_1101</td>
<td>4604</td>
<td>0.863</td>
<td>13657</td>
<td>0.799</td>
</tr>
<tr>
<td>1_1302</td>
<td>4224</td>
<td>0.651</td>
<td>14109</td>
<td>0.658</td>
</tr>
<tr>
<td>1_1501</td>
<td>2900</td>
<td>0.744</td>
<td>13673</td>
<td>0.738</td>
</tr>
<tr>
<td>1_0301</td>
<td>3240</td>
<td>0.738</td>
<td>13587</td>
<td>0.746</td>
</tr>
<tr>
<td>1_0404</td>
<td>3426</td>
<td>0.796</td>
<td>13944</td>
<td>0.852</td>
</tr>
<tr>
<td>1_0701</td>
<td>4076</td>
<td>0.833</td>
<td>13754</td>
<td>0.825</td>
</tr>
<tr>
<td>1_0802</td>
<td>4538</td>
<td>0.789</td>
<td>14187</td>
<td>0.841</td>
</tr>
<tr>
<td>1_0901</td>
<td>3779</td>
<td>0.684</td>
<td>14077</td>
<td>0.653</td>
</tr>
<tr>
<td>3_0101</td>
<td>4302</td>
<td>0.630</td>
<td>14058</td>
<td>0.716</td>
</tr>
<tr>
<td>4_0101</td>
<td>3206</td>
<td>0.746</td>
<td>14161</td>
<td>0.724</td>
</tr>
<tr>
<td>5_0101</td>
<td>4511</td>
<td>0.848</td>
<td>13683</td>
<td>0.831</td>
</tr>
<tr>
<td>mean:</td>
<td>4031</td>
<td>0.752</td>
<td>13564</td>
<td>0.768</td>
</tr>
</tbody>
</table>
Positive Definiteness

THEOREM (SCHOENBERG THEOREM). Let $\mathcal{X}$ be a space in which a distance function $d(x, y)$ is defined subject to the following conditions:

1. $d(x, y) = d(y, x) \geq 0$

2. $d(x, x) = 0$

for all $x, y \in \mathcal{X}$. The function $\exp(-d^p(x, y))$ is positive definite if $0 < p \leq 2$ and not positive definite if $p > 2$.


Schoenberg, I.J.: Metric spaces and positive definite functions. Trans Amer Math Soc 44(3) (1938) 522-536