Novel Machine Learning Methods for MHC Class I Binding Prediction

Christian Widmer[1,*], Nora C. Toussaint[2,*], Yasemin Altun[3], Oliver Kohlbacher[2], Gunnar Rätsch[1]

1 Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany
2 Center for Bioinformatics Tübingen, Eberhard-Karls-Universität, Tübingen, Germany
3 Max Planck Institute for Biological Cybernetics, Tübingen, Germany
*Authors contributed equally.

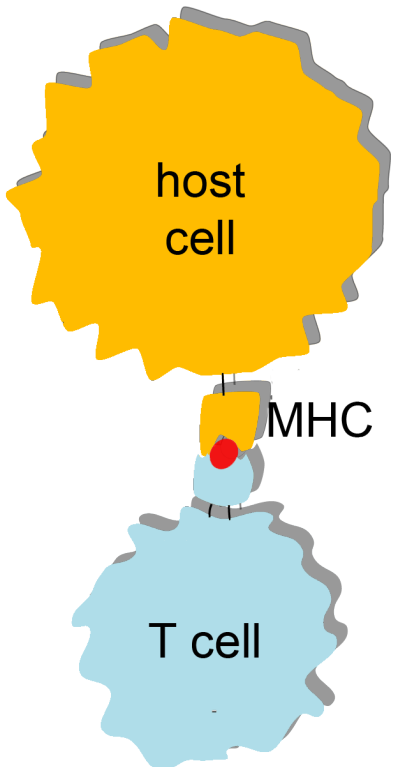
PRIB 2010

September 23rd, 2010

Talk Outline

- Introduction of MHC
- Ingredients of our approach
  - String kernel that respects physico-chemical properties
  - Multitask Learning
  - Fine-tuning the kernel with MKL variant
- Experiments (MHC benchmark dataset)

What is MHC and why does it matter?



MHC molecules

- Membrane-bound proteins
- Present small peptides on cell surface

T cells

- Immune system cells
- Recognize immunogenic peptides when bound to MHC

$\Rightarrow$ Induce immune response

- Immunogenic peptides are of great interest in vaccine design
- MHC-binding prediction: Does a peptide bind?

String kernels for MHC-binding prediction

MHC-I-binding peptides

- Bind in an extended conformation
- Side-chains interact with each other and with MHC

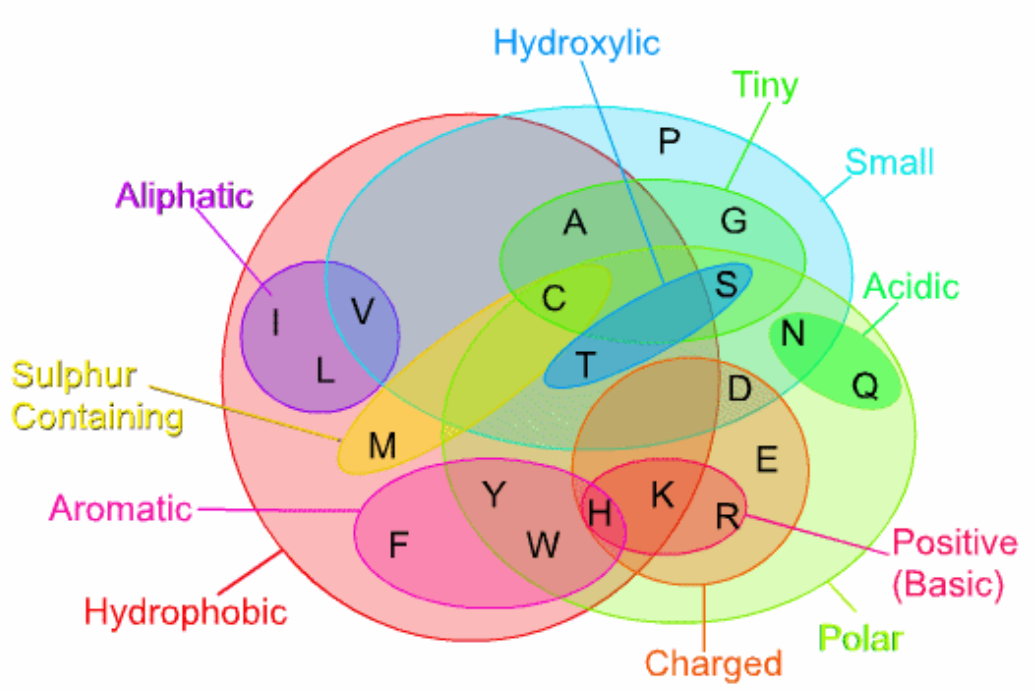
$\longrightarrow$ Sequential structure matters

String kernels

- Exploit sequential structure
- Weighted degree kernel: [Rätsch and Sonnenburg, 2004]

$$K_\ell^{\mathrm{wd}}(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{L-d+1} \sum_{d=1}^{\ell} \beta_d \mathbf{I}\left(\mathbf{x}_{[i:i+d]} = \mathbf{z}_{[i:i+d]}\right)$$

Physico-chemical properties



[http://swift.cmbi.kun.nl/teach/ALIGN/IMAGE/AminoAcid_grouping.gif]

- String kernels ignore physico-chemical properties
- Valuable information especially when data is scarce

Incorporation of physico-chemical properties

- String kernels are based on ℓ-mer comparisons

$$\mathbf{I}(\bar{\mathbf{x}} = \bar{\mathbf{z}}) =: K_{\mathbf{I}}(\bar{\mathbf{x}}, \bar{\mathbf{z}}) = \prod_{i=1}^{\ell} \langle \phi(\bar{\mathbf{x}}_i), \phi(\bar{\mathbf{z}}_i) \rangle$$

- Replace ϕ with physico-chemical encoding ψ:

$$\phi(x_i) = \underbrace{\begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \text{alanine} \\ \text{aspargine} \\ \text{cysteine} \\ \vdots \\ \text{valine} \end{matrix}}_{\text{old encoding}} \quad \psi(x_i) = \underbrace{\begin{pmatrix} 0.3 \\ 0.2 \\ 0.1 \\ \vdots \\ 0.8 \end{pmatrix} \begin{matrix} \text{size} \\ \text{polarity} \\ \text{hydrophobicity} \\ \vdots \\ \text{aromaticity} \end{matrix}}_{\text{new encoding}}$$

Kernels on amino acid substrings

Linear

$$K^{\psi}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) = \prod_{i=1}^{\ell} \langle \psi(\overline{\mathbf{x}}_i), \psi(\overline{\mathbf{z}}_i) \rangle$$

Polynomial

$$K^{\psi}_{\ell,d}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) = \left(\sum_{i=1}^{\ell} \langle \psi(\overline{\mathbf{x}}_i), \psi(\overline{\mathbf{z}}_i) \rangle \right)^{d}$$

RBF

$$K^{\psi}_{\ell,\sigma}(\overline{\mathbf{x}}, \overline{\mathbf{z}}) = \exp\left(- \frac{\sum_{i=1}^{\ell} \| \psi(\overline{\mathbf{x}}_i) - \psi(\overline{\mathbf{z}}_i) \|^2}{\sigma^2} \right)$$

Improved WD kernel

Plug amino acid substring kernel into WD kernel.

- Original WD kernel

$$K_\ell^{\mathrm{wd}}(\mathbf{x},\mathbf{z}) = \sum_{i=1}^{L-d+1}\sum_{d=1}^{\ell}\beta_d \mathbf{I}\left(\mathbf{x}_{[i:i+d]} = \mathbf{z}_{[i:i+d]}\right)$$

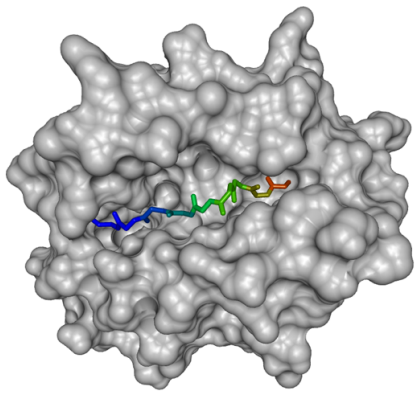
- WD-RBF kernel

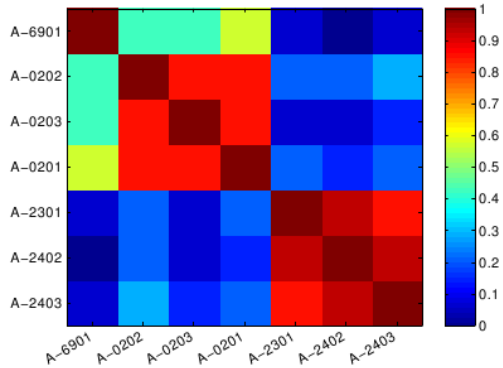
$$K_{\ell,\sigma}^{\mathrm{wd},\Psi}(\mathbf{x},\mathbf{z}) = \sum_{i=1}^{L-d+1}\sum_{d=1}^{\ell}\beta_d \exp\left(-\frac{\sum_{j=1}^{d}\|\Psi(\mathbf{x}_j)-\Psi(\mathbf{z}_j)\|^2}{\sigma^2}\right)$$

  - Sequential structure + physico-chemical properties

A new kernel for Multitask learning

- Many different alleles $\longrightarrow$ pool information





(a) MHC-I binding (b) Sim by pseudo-seq

- We employ a formulation proposed by Jacob and Vert [2008]:

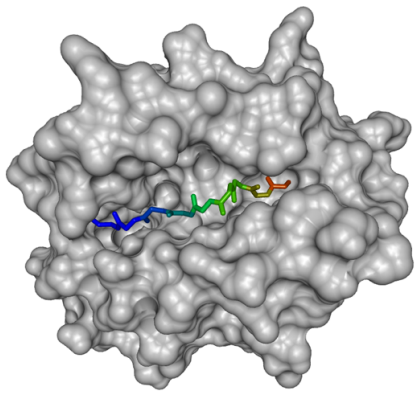
$$K^{\text{MT}}((x,s),(z,t)) = K^{\text{allele}}(s,t) \cdot K^{\text{peptide}}(x,z).$$

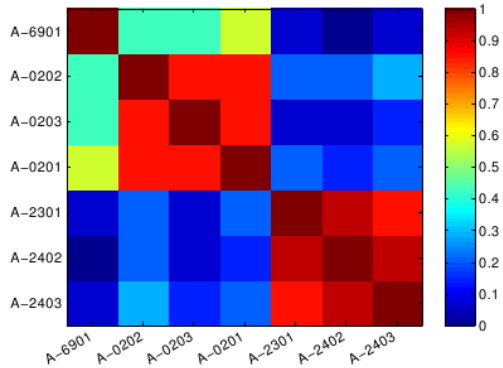
- Plugging in the new kernel leads to :

$$K^{\text{MT-WD-RBF}}((x,s),(z,t)) = K^{\text{WD}}(s,t) \cdot K^{\text{WD-RBF}}(x,z)$$

A new kernel for Multitask learning

- Many different alleles $\longrightarrow$ pool information





(a) MHC-I binding (b) Sim by pseudo-seq

- We employ a formulation proposed by Jacob and Vert [2008]:

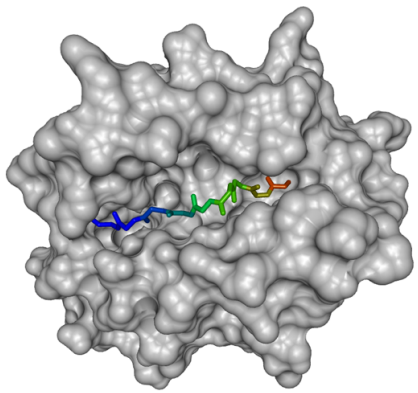
$$K^{\text{MT}}((x,s),(z,t)) = K^{\text{allele}}(s,t) \cdot K^{\text{peptide}}(x,z).$$

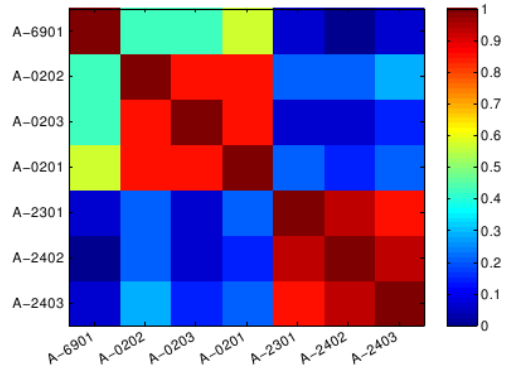
- Plugging in the new kernel leads to :

$$K^{\text{MT-WD-RBF}}((x,s),(z,t)) = K^{\text{WD}}(s,t) \cdot K^{\text{WD-RBF}}(x,z)$$

A new kernel for Multitask learning

- Many different alleles $\longrightarrow$ pool information





(a) MHC-I binding (b) Sim by pseudo-seq

- We employ a formulation proposed by Jacob and Vert [2008]:

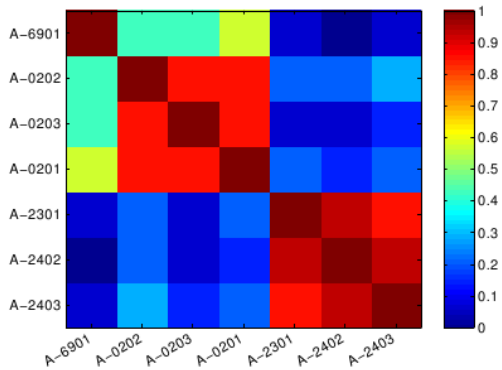
$$K^{\text{MT}}((x,s),(z,t)) = K^{\text{allele}}(s,t) \cdot K^{\text{peptide}}(x,z).$$

- Plugging in the new kernel leads to :

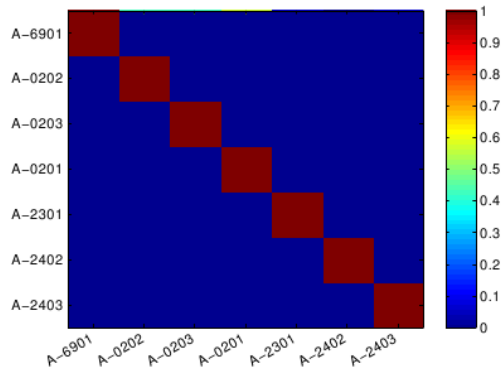
$$K^{\text{MT-WD-RBF}}((x,s),(z,t)) = K^{\text{WD}}(s,t) \cdot K^{\text{WD-RBF}}(x,z)$$

Fine tuning the kernel

- Not all alleles have same amount of training data
- Trade-off in-domain and out-of-domain



(a) out-of-domain

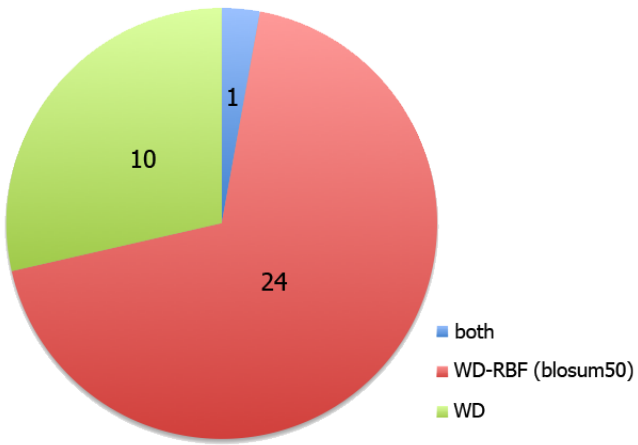


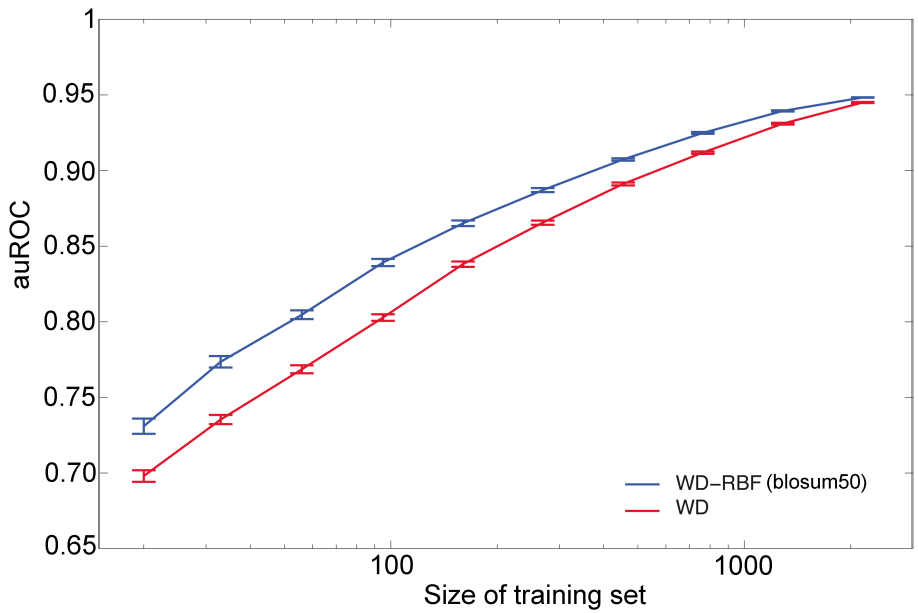
(b) in-domain

$$K^{\text{MT-WD-RBF}}((x,s),(z,t)) = \beta_{s,1} \cdot K^{\text{WD}}(s,t) \cdot K^{\text{WD-RBF}}(x,z) + \beta_{s,2} \cdot \delta_{s,t} \cdot K^{\text{WD-RBF}}(x,z)$$

Idea: Use multiple kernel learning (MKL) for optimal combination

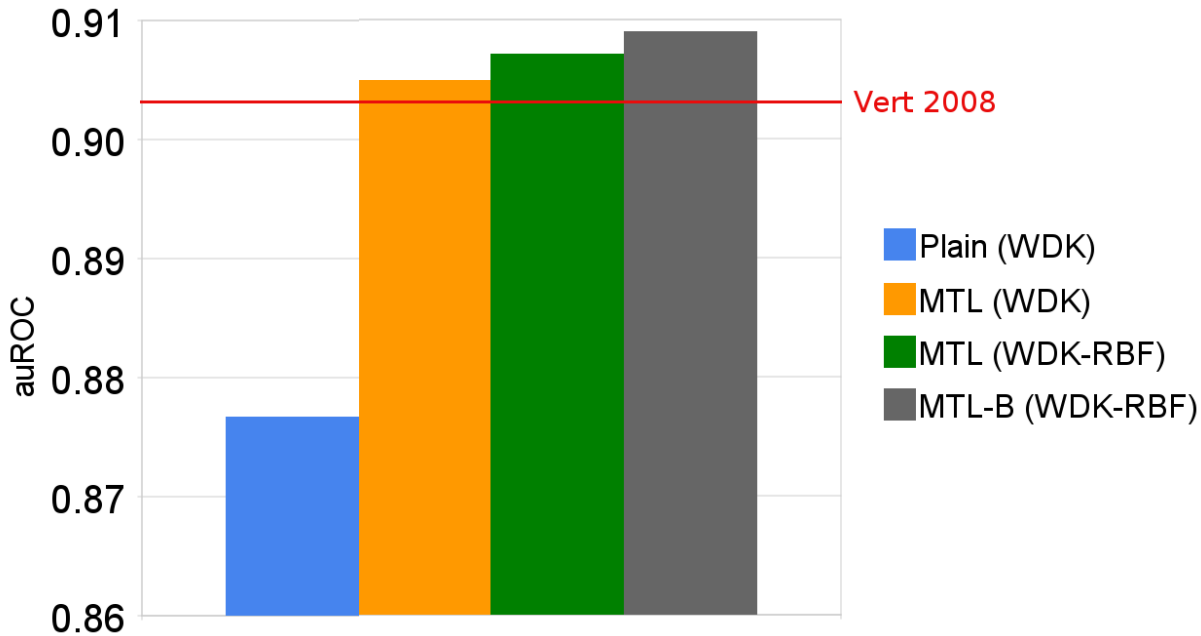
Improved WD kernel vs. standard WD kernel





- Data: IEDB benchmark (35 human alleles) [Peters et al., 2006]
- Performance Measure: auROC (5-fold cross validation)

Improved Multitask Learning Kernel



- Multitask learning greatly improves performance
- WDK-RBF improves performance
- Fine tuning kernel improves results
- Slightly improves over previous results (0.903 auROC)
  [Jacob and Vert, 2008]

Conclusion

- Combination of several orthogonal ideas
  - String kernel that exploits physico-chemical properties
  - Multitask Learning algorithms
  - Fine tuning the kernel by MKL/Boosting
- Sum of several small improvements: Competitive results
- Kernel implementations available in SHOGUN [Sonnenburg et al., 2010]
  `http://www.shogun-toolbox.org`

Acknowledgments

- My co-authors
- Jose Leiva [1,2]
- Yasemin Altun [2]
- Sören Sonnenburg [3]
- Klaus Robert Müller [3]

1 Friedrich Miescher Laboratory of the Max Planck Society
2 Max Planck Institute for Biological Cybernetics
3 Berlin Institute of Technology

Supported by Deutsche Forschungsgemeinschaft and Max Planck Society

Thank you for your attention.

Acknowledgments

- My co-authors
- Jose Leiva [1,2]
- Yasemin Altun [2]
- Sören Sonnenburg [3]
- Klaus Robert Müller [3]

1 Friedrich Miescher Laboratory of the Max Planck Society
2 Max Planck Institute for Biological Cybernetics
3 Berlin Institute of Technology

Supported by Deutsche Forschungsgemeinschaft and Max Planck Society

Thank you for your attention.

References I

P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Proc. ICCV*, volume 1, page 6, 2009.

Laurent Jacob and Jean-Philippe Vert. Efficient peptide-MHC-I binding prediction for alleles with few known binders. *Bioinformatics (Oxford, England)*, 24(3):358–66, February 2008.

B Peters, H-H Bui, S Frankild, M Nielsen, C Lundegaard, E Kostem, D Basch, K Lamberth, M Harndahl, W Fleri, S S Wilson, J Sidney, O Lund, S Buus, and A Sette. A Community Resource Benchmarking Predictions of Peptide Binding to MHC-I Molecules. *PLoS Comput Biol*, 2(6):e65, 06 2006. doi: 10.1371/journal.pcbi.0020065.

G Rätsch and S Sonnenburg. *Accurate Splice Site Detection for Caenorhabditis elegans*. MIT Press, 2004.

B Schölkopf, A J Smola, R C Williamson, and P L Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.

S. Sonnenburg, G. Raetsch, S. Henschel, C. Widmer, A. Zien, F. de Bona, C. Gehl, A. Binder, and V. Franc. The SHOGUN machine learning toolbox. *Journal of Machine Learning Research*, 2010.