Machine learning in immunology

Prediction of binding affinity of peptide-MHC

Vadim Nazarov

Genomics of Adaptive Immunity Lab, IBCH RAS
National Research University Higher School of Economics
1. Introduction to immunology
2. Introduction to deep learning
3. MHC:peptide binding affinity prediction
4. Conclusion
Introduction to immunology
Immune system

- Recognizes foreign / dangerous substances from the environment (mainly microbes).
- Is involved in elimination of old and damaged cells of the body.
- Attacks tumor and virus-infected cells.
Two branches of immune system

• Innate, nonspecific – very quickly recognizes most foreign substances and eliminates them. No memory or learning.

• Adaptive, specific – high degree of specificity in distinction between self and non-self. The reaction takes several days to be effectively triggered. It learns and memorizes the pathogen landscape.
Adaptive immune system

**T cells** destroy infected cells to eradicate intracellular pathogens. (Some bacteria, all viruses)

**B cells** secrete antibodies to attack extracellular pathogens (Most bacteria)

*The colors of the receptors indicate specificity: each can bind to one specific antigen. Adaptive immunity can only attack targets that it has prepared for.*
TCR chains

\(\alpha\beta\) chain - "classic" adaptive immunity (virus detection)

\(\gamma\delta\) chain - terra incognita (phagocytosis, invariant cells)

Different generation processes!
V(D)J recombination

(a) VJ recombination at the tra locus

5' Centromeric

TRAV/TRDC cluster

TRDD cluster

TRDJ cluster

TRDC

TRAJ cluster

TRAC

3' Telomeric

V to J recombination

splicing

TCR-\(\alpha\) chain transcript

(b) V(D)J recombination at the trab locus

5' Centromeric

TRBV cluster

TRBD1

TRBJ1 cluster

TRBC1

TRBD2

TRBJ2 cluster

TRBC2

3' Telomeric

D to J recombination

V to DJ recombination

splicing

TCR-\(\beta\) chain transcript
TCR selection

(a) Thymus

- Thymocyte

(b) Periphery

- Peripheral T cell

(b) Apparent affinity threshold

- % positively selected thymocytes

- Apparent affinity

- Death by neglect
- Positive selection
- Negative selection
TCR:peptide:MHC interaction
## TCR data example

<table>
<thead>
<tr>
<th>Count</th>
<th>Proportion</th>
<th>CDR3.nucleotide.sequence</th>
<th>CDR3.amino.acid.sequence</th>
<th>V.gene</th>
<th>J.gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>9959.760753</td>
<td>7.416466e-02</td>
<td>TGTGCCAGCAGCCAAGCTTAGCCGGGAGCAGATACCCG...</td>
<td>CASSQALAGADTQYF</td>
<td>TRBV4-2</td>
<td>TRBJ2-3</td>
</tr>
<tr>
<td>4425.389760</td>
<td>3.295335e-02</td>
<td>TGTGCCAGCAGCTTAGCCGGGAGCAGATACCCG...</td>
<td>CASSLPRNTGELFF</td>
<td>TRBV13</td>
<td>TRBJ2-2</td>
</tr>
<tr>
<td>3890.686845</td>
<td>2.897173e-02</td>
<td>TGTGCCAGCAGTTATGGCGGGGAGCAGATACCCG...</td>
<td>CASSYGAADTQYF</td>
<td>TRBV12-4, TRBV12-3</td>
<td>TRBJ2-3</td>
</tr>
<tr>
<td>221.330500</td>
<td>1.648122e-03</td>
<td>TGCAGTGTGAGGAGGATTAAACCTCCTACATGAGCA...</td>
<td>CSAGGIETYSEQFF</td>
<td>TRBV20-1</td>
<td>TRBJ2-1</td>
</tr>
<tr>
<td>1799.436602</td>
<td>1.339938e-02</td>
<td>TGTGCCAGCTACCCCACCTTACGGGAGCACTTCTC...</td>
<td>CASSPLGSEQFF</td>
<td>TRBV18</td>
<td>TRBJ2-1</td>
</tr>
<tr>
<td>1316.984630</td>
<td>9.806834e-03</td>
<td>TGTGCCAGCAAAAAAGACACGGGAATGCTACACCTTC...</td>
<td>CASKKDRDYGYTF</td>
<td>TRBV6-5</td>
<td>TRBJ1-2</td>
</tr>
<tr>
<td>2309.863250</td>
<td>1.720023e-02</td>
<td>TGTGCCAGCAGCCAACAGGGATGCAAACACCAATATA...</td>
<td>CASSQGSGNYTF</td>
<td>TRBV7-2</td>
<td>TRBJ1-3</td>
</tr>
<tr>
<td>3339.582627</td>
<td>2.486797e-02</td>
<td>TGTGCCAGCAGTTAGGTCTTCACTACGGACGACTCTTC...</td>
<td>CASSLGLHEQYF</td>
<td>TRBV28</td>
<td>TRBJ2-7</td>
</tr>
</tbody>
</table>
Introduction to deep learning
Deep network architecture ideas

- Fully connected / dense networks (DNN)
- Convolutional neural networks (CNN)
- Recurrent neural networks (RNN)
Fully connected networks 1
A 3-layers fully connected neural network (DNN)
Convolutions
Convolutional neural networks
Recurrent neural networks
MHC:peptide binding affinity prediction
Prediction of strong / weak binders (immunotherapy, etc.)
Data

140,000 pairs of MHC-peptide for training

30,000 pairs of MHC-peptide for testing

<table>
<thead>
<tr>
<th>species</th>
<th>mhc</th>
<th>peptide_length</th>
<th>cv</th>
<th>sequence</th>
<th>inequality</th>
<th>meas</th>
</tr>
</thead>
<tbody>
<tr>
<td>cow</td>
<td>BoLA-HD6</td>
<td>9</td>
<td>TBD</td>
<td>ALFYKDGKLN</td>
<td>=</td>
<td>1.0</td>
</tr>
<tr>
<td>cow</td>
<td>BoLA-HD6</td>
<td>9</td>
<td>TBD</td>
<td>ALYEKKLAL</td>
<td>=</td>
<td>1.0</td>
</tr>
<tr>
<td>cow</td>
<td>BoLA-HD6</td>
<td>9</td>
<td>TBD</td>
<td>AMKDRFQPL</td>
<td>=</td>
<td>4.52170583277</td>
</tr>
<tr>
<td>cow</td>
<td>BoLA-HD6</td>
<td>9</td>
<td>TBD</td>
<td>AQRLEFFTL</td>
<td>=</td>
<td>1.0</td>
</tr>
<tr>
<td>cow</td>
<td>BoLA-HD6</td>
<td>9</td>
<td>TBD</td>
<td>FMKVKEAL</td>
<td>=</td>
<td>1.57674703262</td>
</tr>
<tr>
<td>cow</td>
<td>BoLA-HD6</td>
<td>9</td>
<td>TBD</td>
<td>FQHERLGQF</td>
<td>=</td>
<td>1.0</td>
</tr>
<tr>
<td>cow</td>
<td>BoLA-HD6</td>
<td>9</td>
<td>TBD</td>
<td>FQRAIMNAM</td>
<td>=</td>
<td>1.0</td>
</tr>
<tr>
<td>cow</td>
<td>BoLA-HD6</td>
<td>9</td>
<td>TBD</td>
<td>GQFLSFAVL</td>
<td>=</td>
<td>1.0</td>
</tr>
<tr>
<td>cow</td>
<td>BoLA-HD6</td>
<td>9</td>
<td>TBD</td>
<td>GQFNRYAAM</td>
<td>=</td>
<td>1.0</td>
</tr>
</tbody>
</table>
NetMHCpan

Paper: just google "netMHCpan paper"

Features:

• Onehot encoding
• Blosum encoding
• Lengths
• Indels

Pseudo-sequences – pan-allele approach

Model: DNN with 60 hidden neurons

F1 score - 0.8

\[
F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

\[
\text{precision} = \frac{TP}{TP + FP}
\]

\[
\text{recall} = \frac{TP}{TP + FN}
\]
word2vec
word2vec vectors
MICE: average multiple imputations generated using Gibbs sampling from the joint distribution of columns.
Paper: http://biorxiv.org/content/biorxiv/early/2016/05/22/054775.full.pdf

Features:

- Embeddings (per-pseudo-sequence!)

Model: DNN with 60 neurons

F1 score - 0.79
GRU

\[ z_t = \sigma (W_z \cdot [h_{t-1}, x_t]) \]
\[ r_t = \sigma (W_r \cdot [h_{t-1}, x_t]) \]
\[ \tilde{h}_t = \tanh (W \cdot [r_t \ast h_{t-1}, x_t]) \]
\[ h_t = (1 - z_t) \ast h_{t-1} + z_t \ast \tilde{h}_t \]
Paper:
http://www.biorxiv.org/content/biorxiv/early/2017/07/27/154757.full.pdf

Features:

- One-hot
- Model per-pseudo-sequence (64 units + sigmoid)
- Not even multi-layer or bidirectional!

Model: simple GRU

F1 score - 0.81
CNN for NLP

Sentence matrix $7 \times 5$

3 region sizes: (2,3,4)
2 filters for each region size
totally 6 filters

2 feature maps for each region size

6 univariate vectors concatenated together to form a single feature vector

1-max pooling
softmax function regularization in this layer

2 classes

Features:

- Embeddings on the overall data
- Model per-pseudo-sequence (64 units + sigmoid)
- Large convolutions

Model: CNN

F1 score - 0.75
ResNet - old networks’ problems

- Gradient vanishing
- Large number of parameters
- Shallowness
ResNet - proposed model
Revolution of Depth

ResNet - results

ImageNet Classification top-5 error (%)

ILSVRC'15 ResNet: 3.57
ILSVRC'14 GoogleNet: 6.7
ILSVRC'14 VGG: 7.3
ILSVRC'13: 11.7
ILSVRC'12 AlexNet: 16.4
ILSVRC'11: 25.8
ILSVRC'10: 28.2

152 layers

22 layers
19 layers
8 layers
8 layers
shallow
ResNet - old networks

**VGG, 19 layers (ILSVRC 2014)**

- 3x3 conv, 64
- 3x3 conv, 64, pool/2
- 3x3 conv, 128
- 3x3 conv, 128, pool/2
- 3x3 conv, 256
- 3x3 conv, 256
- 3x3 conv, 256
- 3x3 conv, 256, pool/2
- 3x3 conv, 512
- 3x3 conv, 512
- 3x3 conv, 512
- 3x3 conv, 512
- 3x3 conv, 512, pool/2
- fc, 4096
- fc, 4096
- fc, 1000
ResNet - current deep networks

VGG, 19 layers (ILSVRC 2014)

ResNet, 152 layers (ILSVRC 2015)
Our approach
Our approach - results

- F1 0.81
- Global models – prediction of binding affinities for unseen MHCs (mean F1 0.72)
- Better models for the per-pseudo-sequence approach.
Conclusion
Vadim I. Nazarov
Genomics of Adaptive Immunity Lab, IBCH RAS
National Research University Higher School of Economics

email: vdm.nazarov@gmail.com
telegram: @vadimnazarov