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

Table of contents

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.

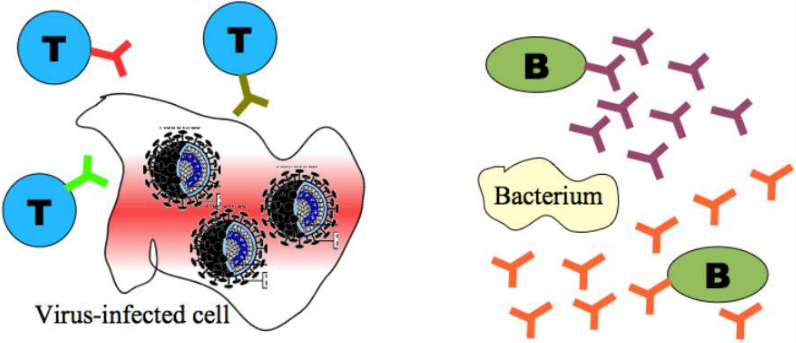
Two branches of immune system

- 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.

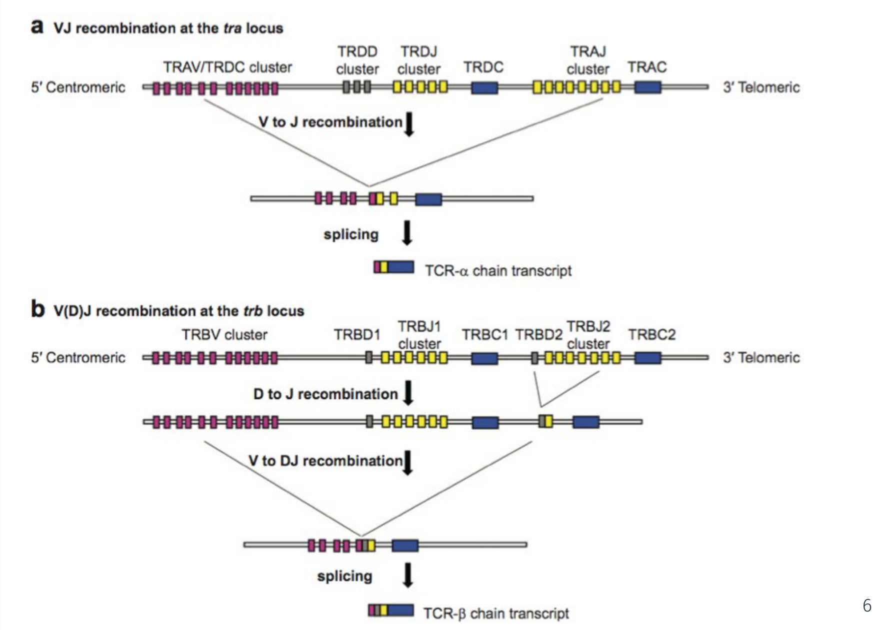
TCR chains

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

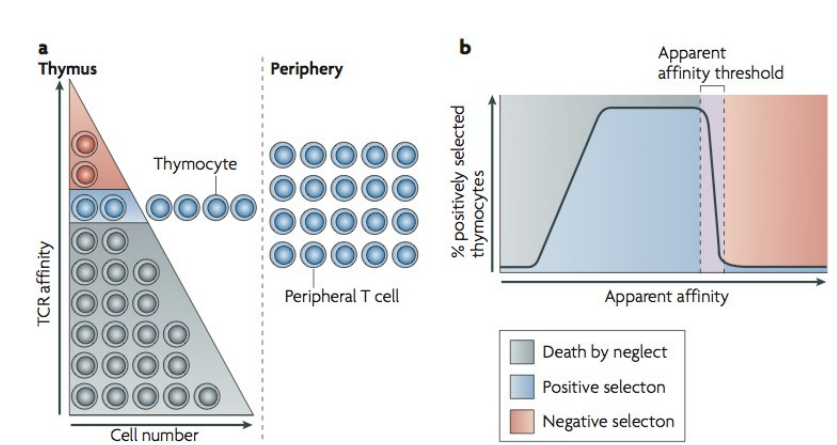
$\gamma\delta$ chain - terra incognita (phagocytosis, invariant cells)

Different generation processes!

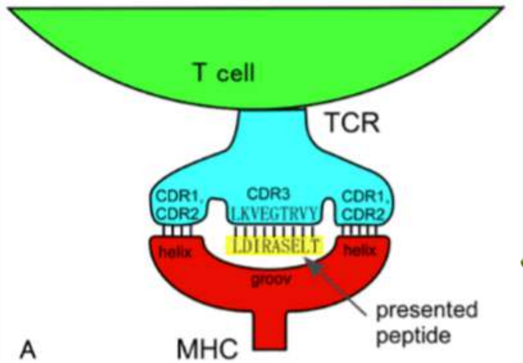
V(D)J recombination

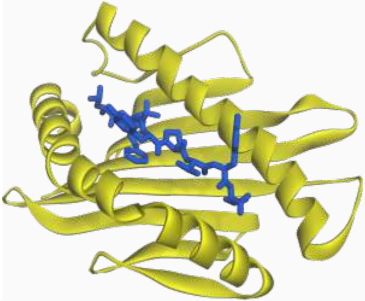


TCR selection

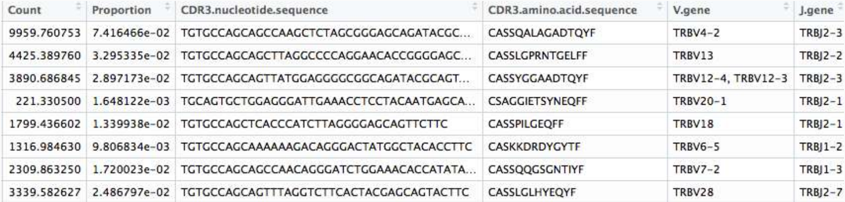


TCR:peptide:MHC interaction





TCR data example



Count	Proportion	CDR3.nucleotide.sequence	CDR3.amino.acid.sequence	V.gene	J.gene
9959.760753	7.416466e-02	TGTGCCAGCAGCCAAGCTCTAGCGGGAGCAGATACGC...	CASSQALAGADTQYF	TRBV4-2	TRBJ2-3
4425.389760	3.295335e-02	TGTGCCAGCAGCTTAGGCCCCAGGAACACCGGGGAGC...	CASSLGPRNTGELFF	TRBV13	TRBJ2-2
3890.686845	2.897173e-02	TGTGCCAGCAGTTATGGAGGGGCGGCAGATACGCAGT...	CASSYGGAADTQYF	TRBV12-4, TRBV12-3	TRBJ2-3
221.330500	1.648122e-03	TGCAGTGCTGGAGGGATTGAAACCTCCTACAATGAGCA...	CSAGGIETSYNEQFF	TRBV20-1	TRBJ2-1
1799.436602	1.339938e-02	TGTGCCAGCTCACCCATCTTAGGGGAGCAGTTCTTC	CASSPILGEQFF	TRBV18	TRBJ2-1
1316.984630	9.806834e-03	TGTGCCAGCAAAAAAGACAGGGACTATGGCTACACCTTC	CASKKDRDYGYTF	TRBV6-5	TRBJ1-2
2309.863250	1.720023e-02	TGTGCCAGCAGCCAACAGGGATCTGGAAACACCATATA...	CASSQQGSGNTIYF	TRBV7-2	TRBJ1-3
3339.582627	2.486797e-02	TGTGCCAGCAGTTTAGGTCTTCACTACGAGCAGTACTTC	CASSLGLHYEQYF	TRBV28	TRBJ2-7

Introduction to deep learning

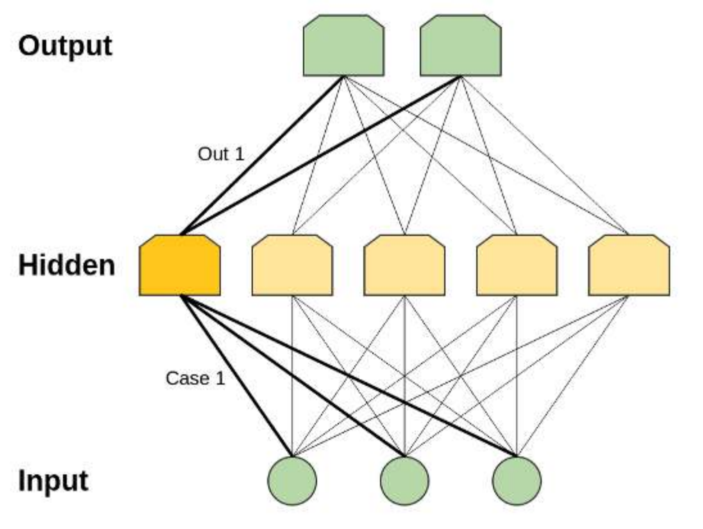
Deep network architecture ideas

Fully connected / dense networks (DNN)

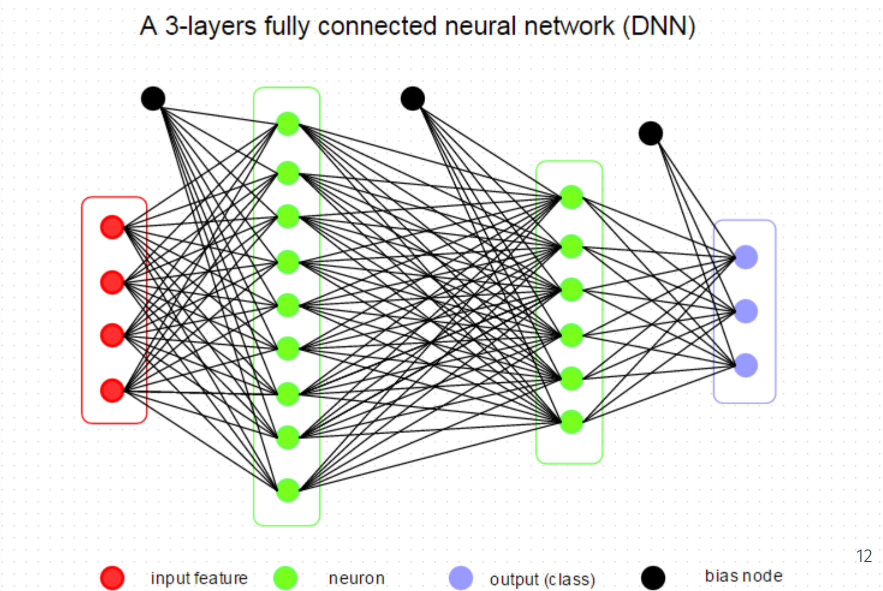
Convolutional neural networks (CNN)

Recurrent neural networks (RNN)

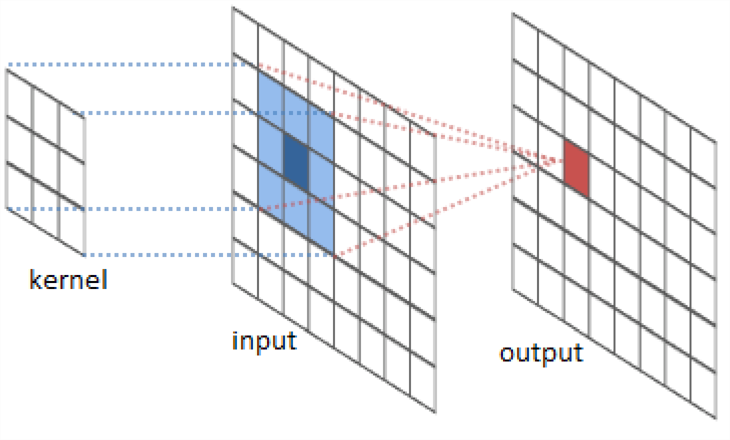
Fully connected networks 1



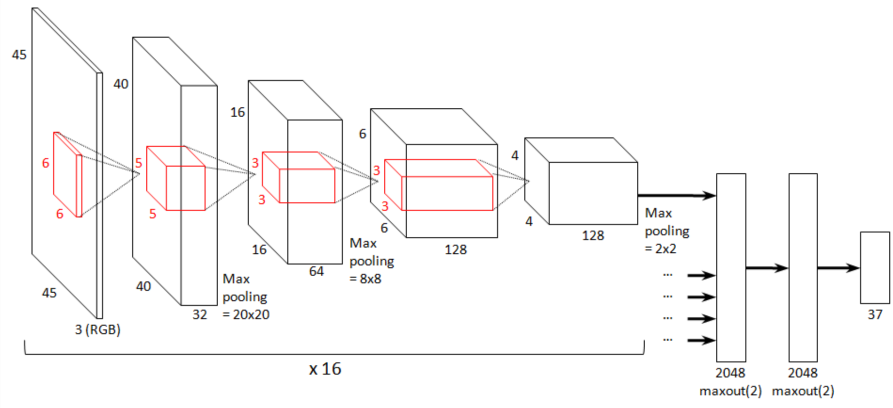
Fully connected networks 2



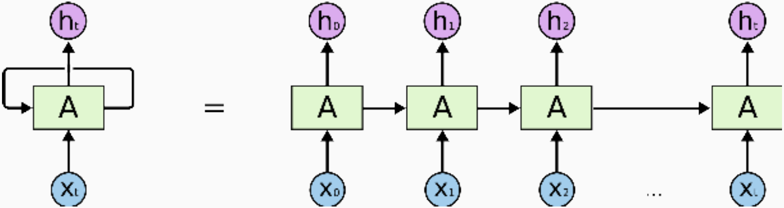
Convolutions



Convolutional neural networks



Recurrent neural networks



MHC:peptide binding affinity prediction

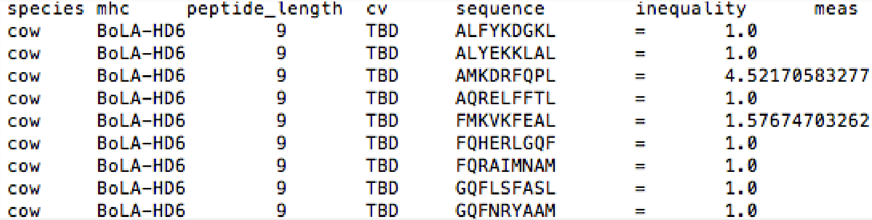
Problem

Prediction of strong / weak binders (immunotherapy, etc.)

Data

140,000 pairs of MHC-peptide for training

30,000 pairs of MHC-peptide for testing



species	mhc	peptide_length	cv	sequence	inequality	meas
cow	BoLA-HD6	9	TBD	ALFYKDGKL	=	1.0
cow	BoLA-HD6	9	TBD	ALYEKKLAL	=	1.0
cow	BoLA-HD6	9	TBD	AMKDRFQPL	=	4.52170583277
cow	BoLA-HD6	9	TBD	AQRELFFTL	=	1.0
cow	BoLA-HD6	9	TBD	FMKVKFEAL	=	1.57674703262
cow	BoLA-HD6	9	TBD	FQHERLGQF	=	1.0
cow	BoLA-HD6	9	TBD	FQRAIMNAM	=	1.0
cow	BoLA-HD6	9	TBD	GQFLSFASL	=	1.0
cow	BoLA-HD6	9	TBD	GQFNRYAAM	=	1.0

NetMHCpan

Paper: just google ”netMHCpan paper”

Features:

- Onehot encoding
- Blosum encoding
- Lengths
- Indels

Pseudo-sequences – pan-allele approach

Model: DNN with 60 hidden neurons

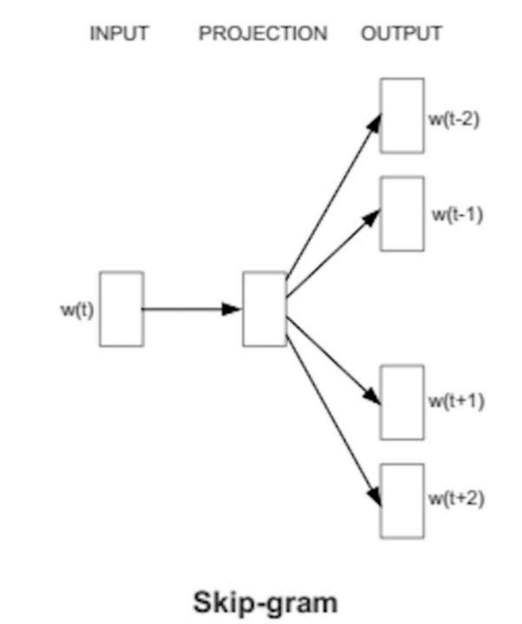
F1 score - 0.8

$$F1 = 2 * precision * recall/(precision + recall)$$

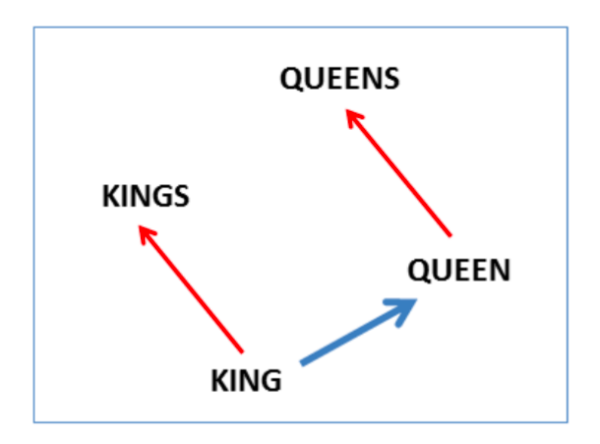
$$precision = TP/(TP + FP)$$

$$recall = TP/(TP + FN)$$

word2vec

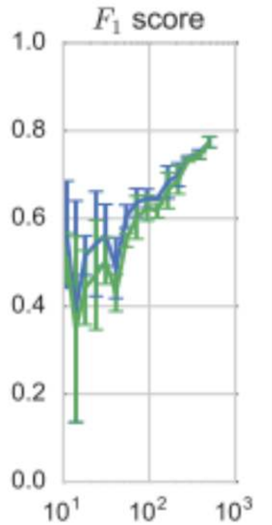


word2vec vectors



Imputation

MICE: average multiple imputations generated using Gibbs sampling from the joint distribution of columns.



mhcflurry

Paper: http://biorxiv.org/content/biorxiv/early/2016/05/22/054775.full.pdf

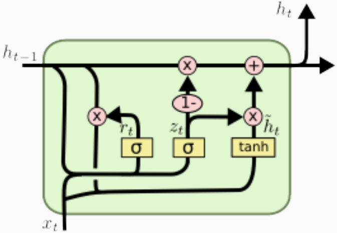
Features:

- Embeddings (per-pseudo-sequence!)

Model: DNN with 60 neurons

F1 score - 0.79

GRU



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

MHCnuggets

Paper:
http://www.biorxiv.org/content/biorxiv/early/2017/07/27/154757.full.pdf

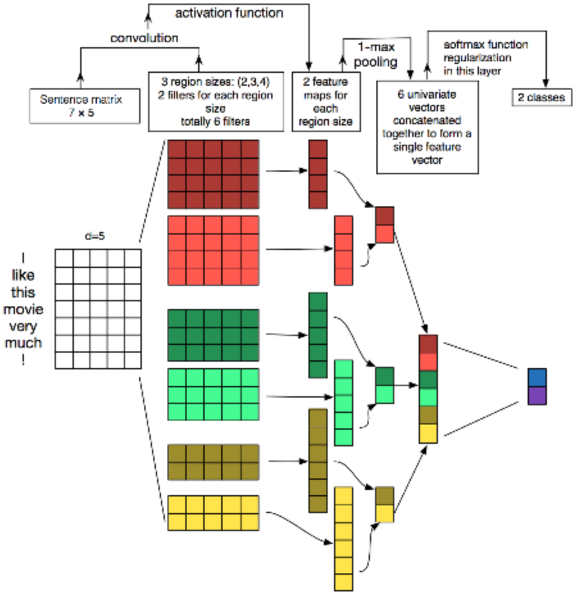
Features:

- 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


activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 x 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
2 classes
d=5
I
like
this
movie
very
much
!

HLA-CNN

Paper:
http://www.biorxiv.org/content/biorxiv/early/2017/07/27/154757.full.pdf

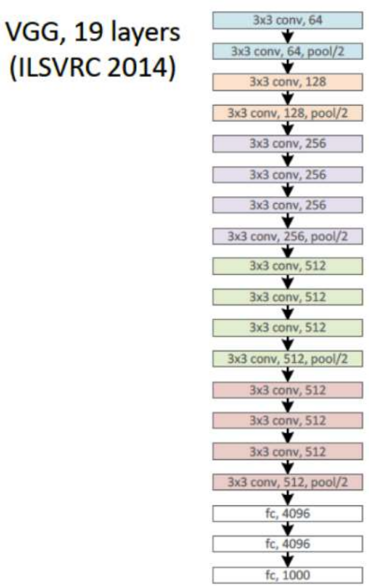
Features:

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

Model: CNN

F1 score - 0.75

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, pool/2

3x3 conv, 512

3x3 conv, 512

3x3 conv, 512

3x3 conv, 512, pool/2

fc, 4096

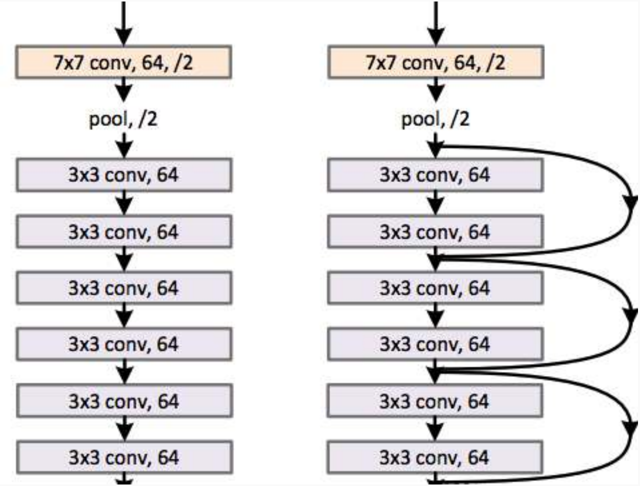
fc, 4096

fc, 1000

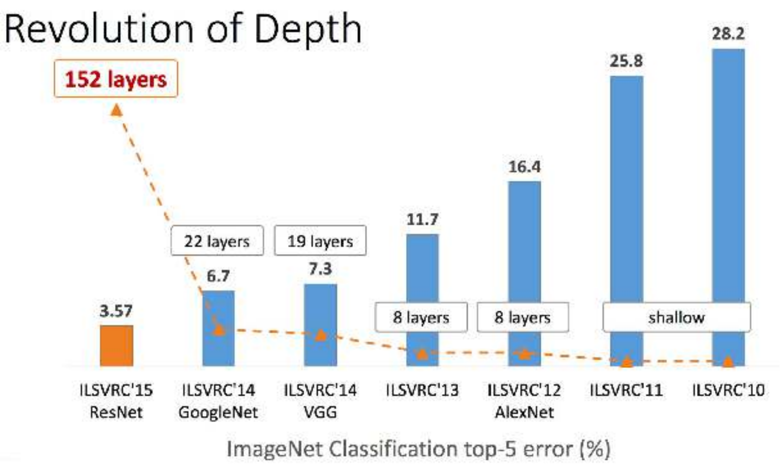
ResNet - old networks' problems

- Gradient vanishing
- Large number of parameters
- Shallowness

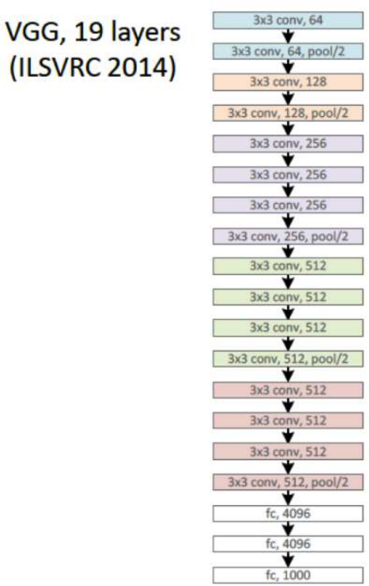
ResNet - proposed model



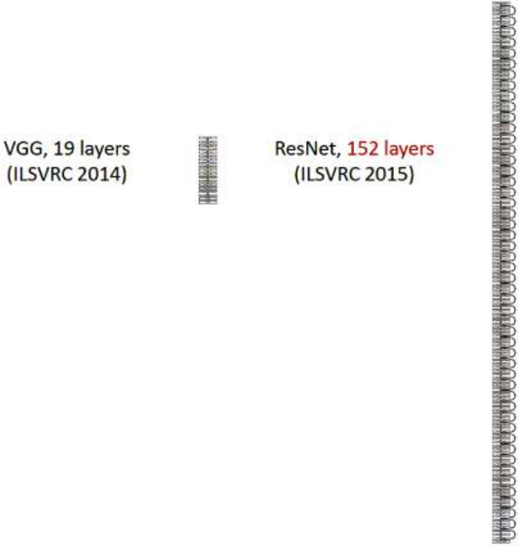
ResNet - results



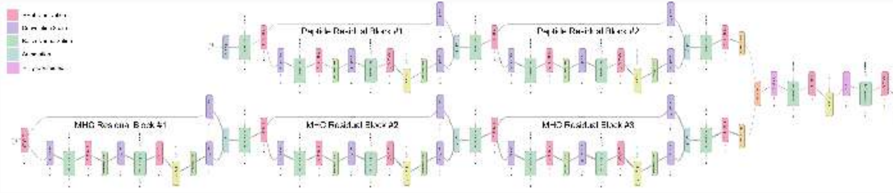
ResNet - old networks



ResNet - current deep networks



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