PREDICTION OF T-CELL EPITOPES INTERACTIONS WITH MAJOR HISTOCOMPATIBILITY COMPLEX II USING THEIR STRUCTURAL FORMULAS

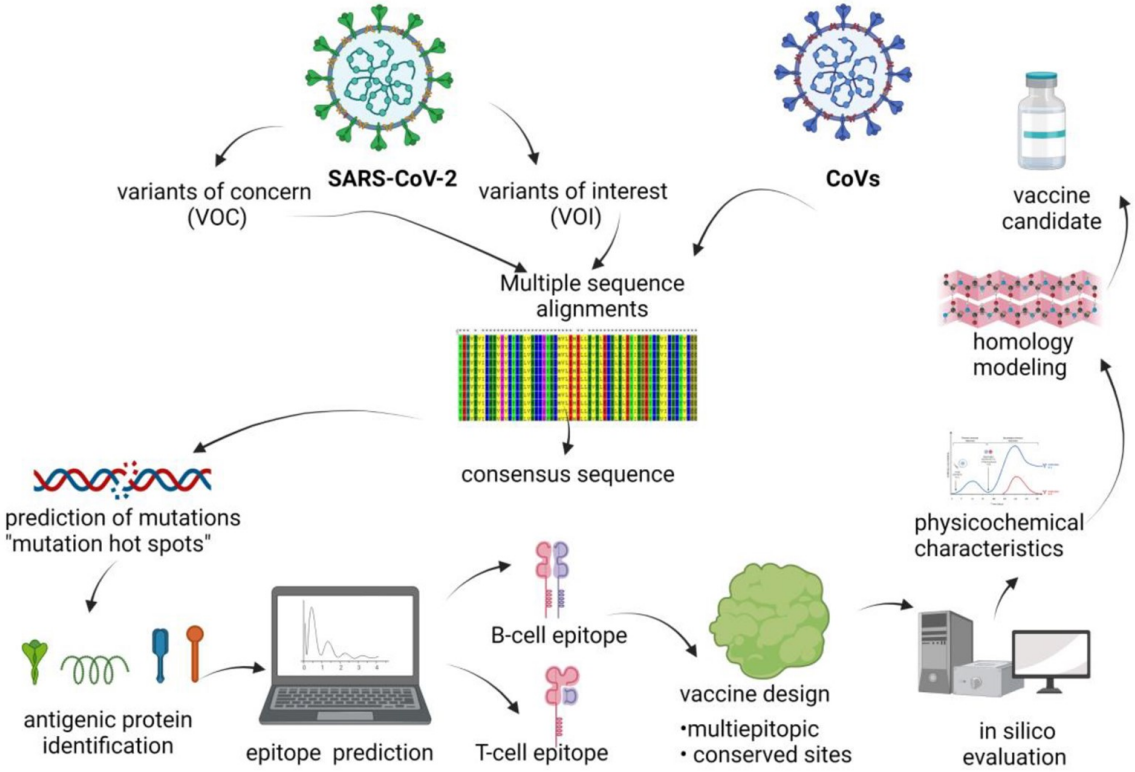
ПИРОГОВСКИЙ УНИВЕРСИТЕТ

Smirnov Anton Sergeevich

Pirogov Russian National Research Medical University

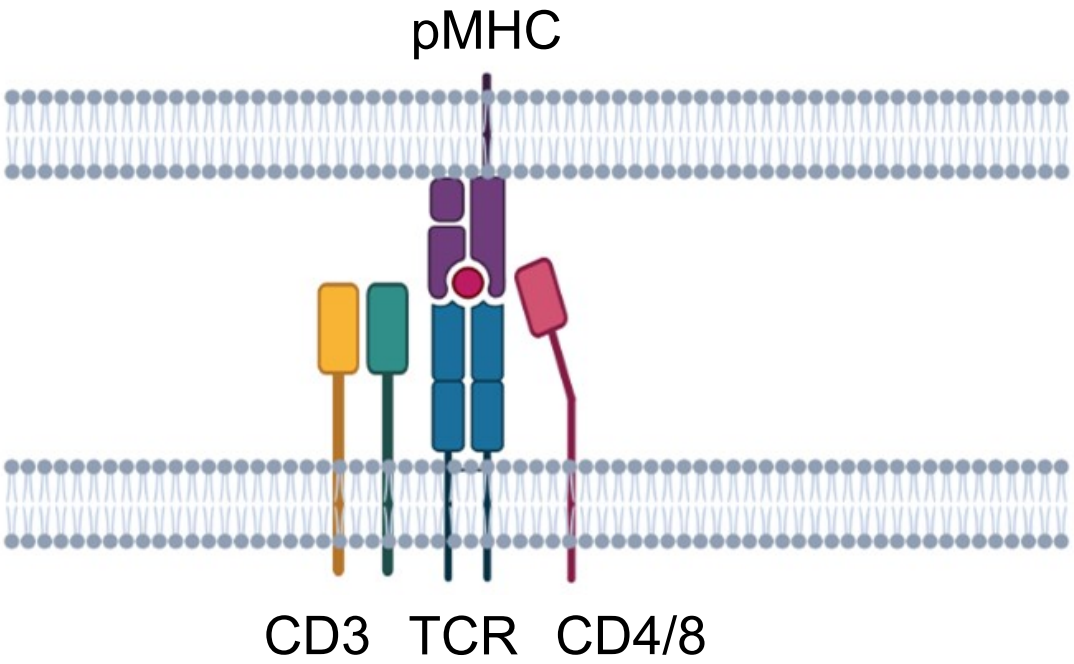
Bochkov Research Centre for Medical Genetics

Introduction



10.3390/vaccines10111844

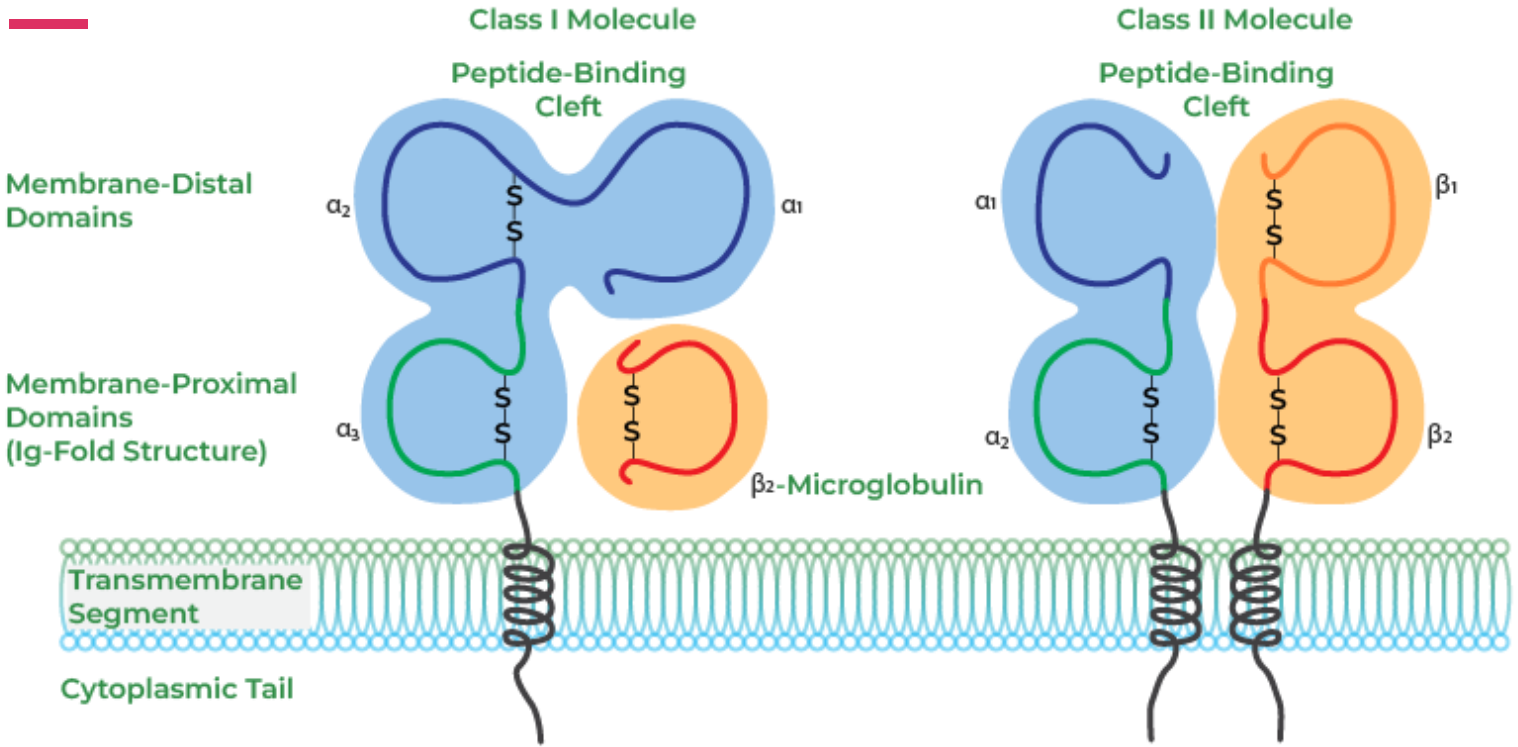
Key questions



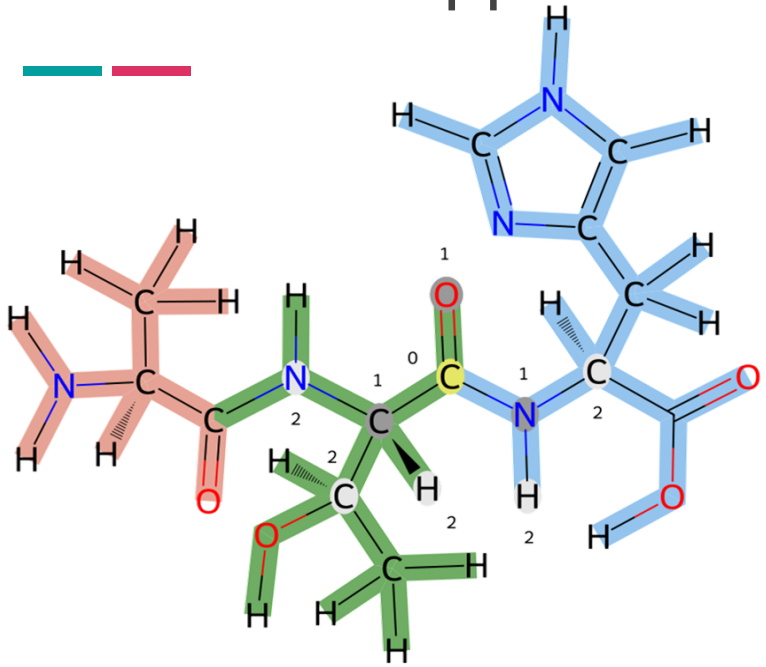
What peptides are presented on MHC?

What T-cell receptor reacts?

MHC Class I vs MHC Class II



MultiPASS approach



$$B_k = (S_k - S_{0k}) / (1 - S_k \cdot S_{0k})$$

$$S_k = Sin\left[\sum_{i=1}^{m} ArcSin\left(2P\left(A_k|D_i\right) - 1\right)/m\right]$$

$$S_{0k} = 2P\left(A_k\right) - 1$$

10.18097/BMCRM00004
10.1111/imm.13641

Peptide: Ala-Thr-His
MNA_0: -C
MNA_1: -C(-C-N-O)
MNA_2: -C(-C(-H-C-C-N)-N(-H-C-C)-O(-C))

10.60797/jbg.2026.33.3

The datasets

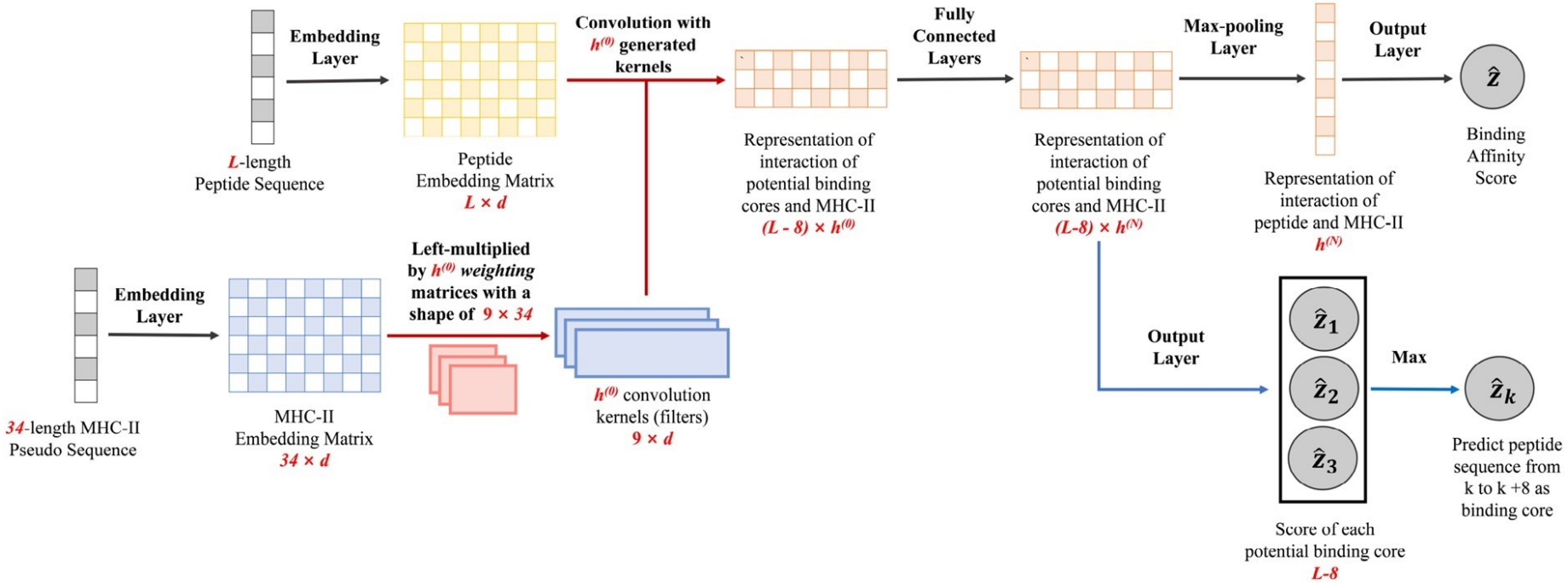
BD2016 dataset (DeepMHCII)

- 134 281 records
- 80 different MHC-II alleles (72 human and 8 mouse alleles)
- Splitted for 5-fold cross-validation

BD2020 dataset (MHCAttnNet)

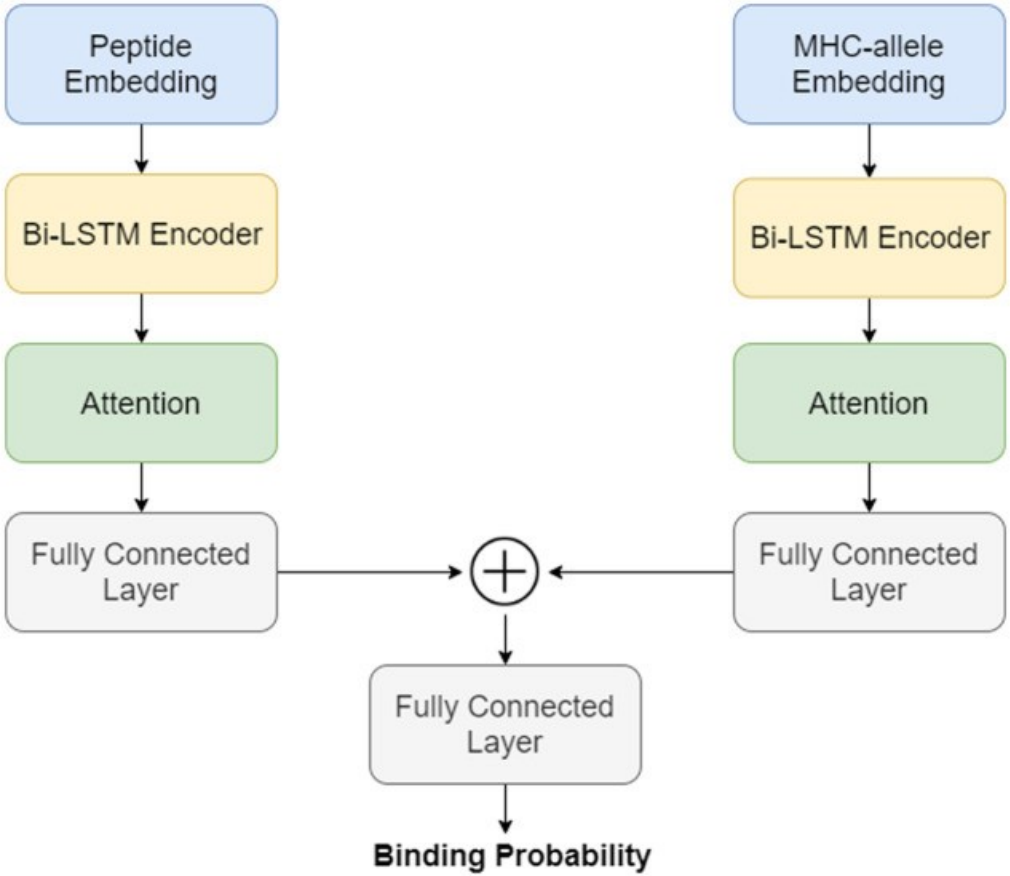
- 65 954 records
- 49 HLA-DRB alleles
- Splitted for 5-fold cross-validation

DeepMHCII



10.1093/bioinformatics/btac225

MHCAttnNet



10.1093/bioinformatics/btaa479

Choosing training mode

Species	PASS-AS	PASS-MA	NetMHCIIpan-3.2	PUFFIN	DeepSeqPan II	DeepMHCII
H2	0,773	**0,859**	0,792	0,770	0,662	0,782
HLA	0,748	0,783	0,861	0,863	0,778	**0,874**

Allele-Specific (AS)

1. DeepMHCII – 40
2. PUFFIN – 8
3. NetMHCIIpan-3.2 – 8
4. PASS – 3

Multi-Allele (MA)

1. DeepMHCII – 38
2. PASS – 14
3. PUFFIN – 5
4. NetMHCIIpan-3.2 – 4

DeepMHCII comparision

metric	AUPRC		AUROC		BA		F1		MCC	
tool	Deep MHCII	Multi PASS	Deep MHCII	Multi PASS	Deep MHCII	Multi PASS	Deep MHCII	Multi PASS	Deep MHCII	Multi PASS
value	0.764	0.679	0.843	0.778	0.710	0.731	0.787	0.730	0.452	0.442

Excluded

- H2-EA*d/EB*d
- HLA-DRB1*04:02
- HLA-DQA1*03:01/DQB1*03:01

MHCAttnNet comparision

metric	AUPRC		AUROC		BA		F1		MCC	
tool	MHC Attn Net	Multi PASS	MHC Attn Net	Multi PASS	MHC Attn Net	Multi PASS	MHC Attn Net	Multi PASS	MHC Attn Net	Multi PASS
All	0.627	0.830	0.616	0.839	0.583	0.799	0.577	0.795	0.210	0.592
1-2	0.824	0.842	0.843	0.850	0.804	0.807	0.804	0.798	0.604	0.602

Excluded

HLA-DRB1*01:02
HLA-DRB1*03:03
HLA-DRB1*08:01
HLA-DRB1*12:02
HLA-DRB1*14:01
HLA-DRB1*16:02
HLA-DRB3*03:01
HLA-DRB5*01:02

With charged descriptors: BD2016

Allele	neutral	charged
H2-AA*u/AB*u	0,976583	0,976583
HLA-DPA1*01:03/ DPB1*06:01	0,91504	0,912537
HLA-DQA1*05:01/ DQB1*03:01	0,857604	0,858132
H2-AA*s/AB*s	0,855782	0,850218
HLA-DPA1*02:01/ DPB1*14:01	0,854902	0,855095

With charged descriptors: BD2020

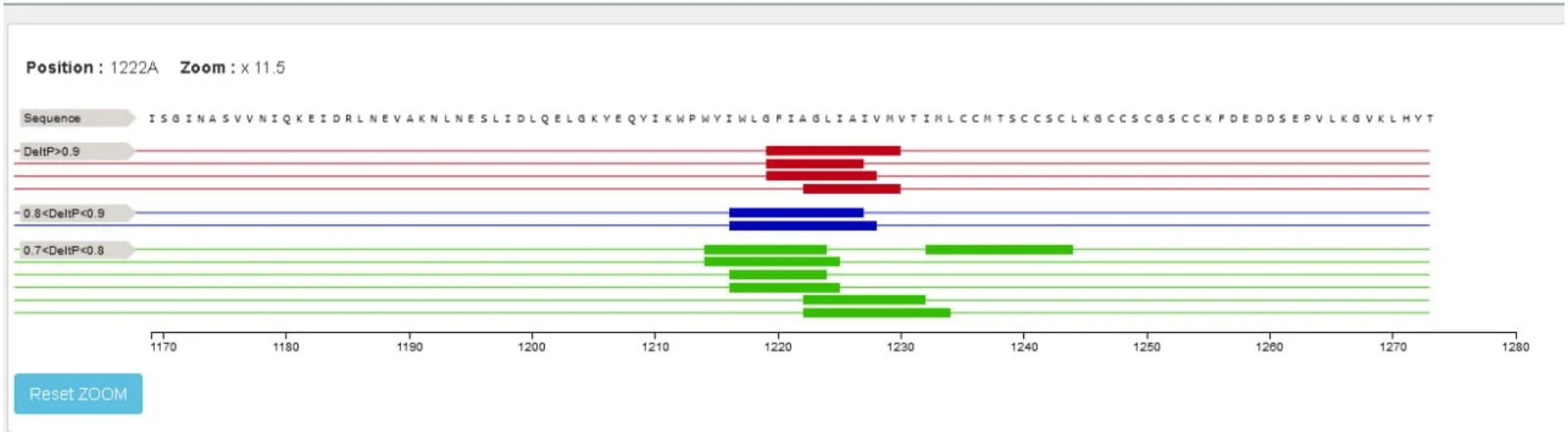
Allele	neutral	charged
HLA-DRB1*10:01	0,961422	0,961829
HLA-DRB1*15:02	0,940108	0,939954
HLA-DRB1*03:02	0,919903	0,921986
HLA-DRB1*04:05	0,915943	0,915947
HLA-DRB1*16:01	0,907500	0,907500

Previously…

- BCADD2024 - Creation of sar models for prediction of t-cell epitopes with human leukocyte antigens based on protein structural formulas

- BGRS2026 - Prediction of T-cell epitopes interactions with major histocompatibility complex I using their structural formulas

EpiSpectrum



TCR-Pred interface

ASSLVGPPGEL

Make prediction Examples:ASSLVGPPGEL

This resource is about prediction of epitopes and HLA types for alpha and beta chains of CDR3 T-cell receptors (TCR) amino acid sequences based on classification structure-activity relationships (SAR) models. SAR models were created based on the data from VDJdb, McPAS-TCR and IEDB databases using Bayesian-like approach and atom centric substructural Multilevel Neighborhoods of Atoms (MNA) descriptors. During the prediction the letter text representation of CDR3 TCR amino acid sequences are transformed to MOL V3000 format of structural formula representation.

Prediction is based on special version of PASS (Prediction of Activity Spectra for Substances) technology (http://www.way2drug.com/PASSonline) which was modified for generation up to 15 levels of MNA descriptors. After creation of SAR models with different levels of MNA descriptors we selected level of MNA descriptors leading to the most accurate SAR models. The SAR models for prediction of epitopes were created based on 9st (for alpha chain of CDR3 TCR) or 11st (for beta chain of CDR3 TCR) level of MNA descriptors. The SAR models for prediction of MHC types were created based on 11st (for beta chain of CDR3 TCR) or 8st (for alpha chain of CDR3 TCR) level of MNA descriptors. Lists of predicted epitopes and MHC types with accuracy of prediction are available at description of Training Sets.

Please cite us: Smirnov AS, Rudik AV, Filimonov DA, Lagunin AA. TCR-Pred: A new web-application for prediction of epitope and MHC specificity for CDR3 TCR sequences using molecular fragment descriptors. Immunology. 2023, 169(4):447-453. doi: 10.1111/imm.13641 https://onlinelibrary.wiley.com/doi/10.1111/imm.13641

Number of unique sequences		v.2023				v.2026			
		Human		Mouse		Human		Mouse	
		α	β	α	β	α	β	α	β
CDR3 – epitopes	CDR3	35616	141500	1604	3014	72630	134026	3401	6156
	epitopes	799	1257	96	123	1865	2052	148	171
CDR3 – MHC	CDR3	31066	47639	1462	2550	72630	134026	3401	6156
	MHC	76	88	11	12	121	146	11	14

Article

Web-service

https://way2drug.com/TCR-pred/index.php

Summary

1. MultiPASS with structural representation has comparable performance and can win on deep neural networks with letter representation, but DeepMHCII significantly outperforms MultiPASS
2. The charged MNA descriptors give no wins
3. The models will be published soon at platform Way2Drug. Stay tuned!

Acknowledgements

- The study is supported by the grant of Russian Science Foundation No. 25-25-01171, https://rscf.ru/project/25-25-01171/
- To Institute of Translational Medicine, Pirogov Russian National Research Medical University and Aldan3 cluster administrators
- To my scientific advisor, Alexey A. Lagunin for comprehensive expertise and moral support and Dmitry A. Filimonov for MultiPASS program.

Contacts

- Смирнов Антон Сергеевич (Smirnov Anton Sergeevich)
- anton.smirnov.9910@gmail.com
- smirnov_as@rsmu.ru
- Tg: @SmirnygaTotoshka

