
DTU

Biohackathon 2021

Prediction of TCR-pMHC binding

Who are we?

AI for Immunological Molecules group at DTU (Paolo Marcatili)

Ida Meitil
Master thesis student at DTU

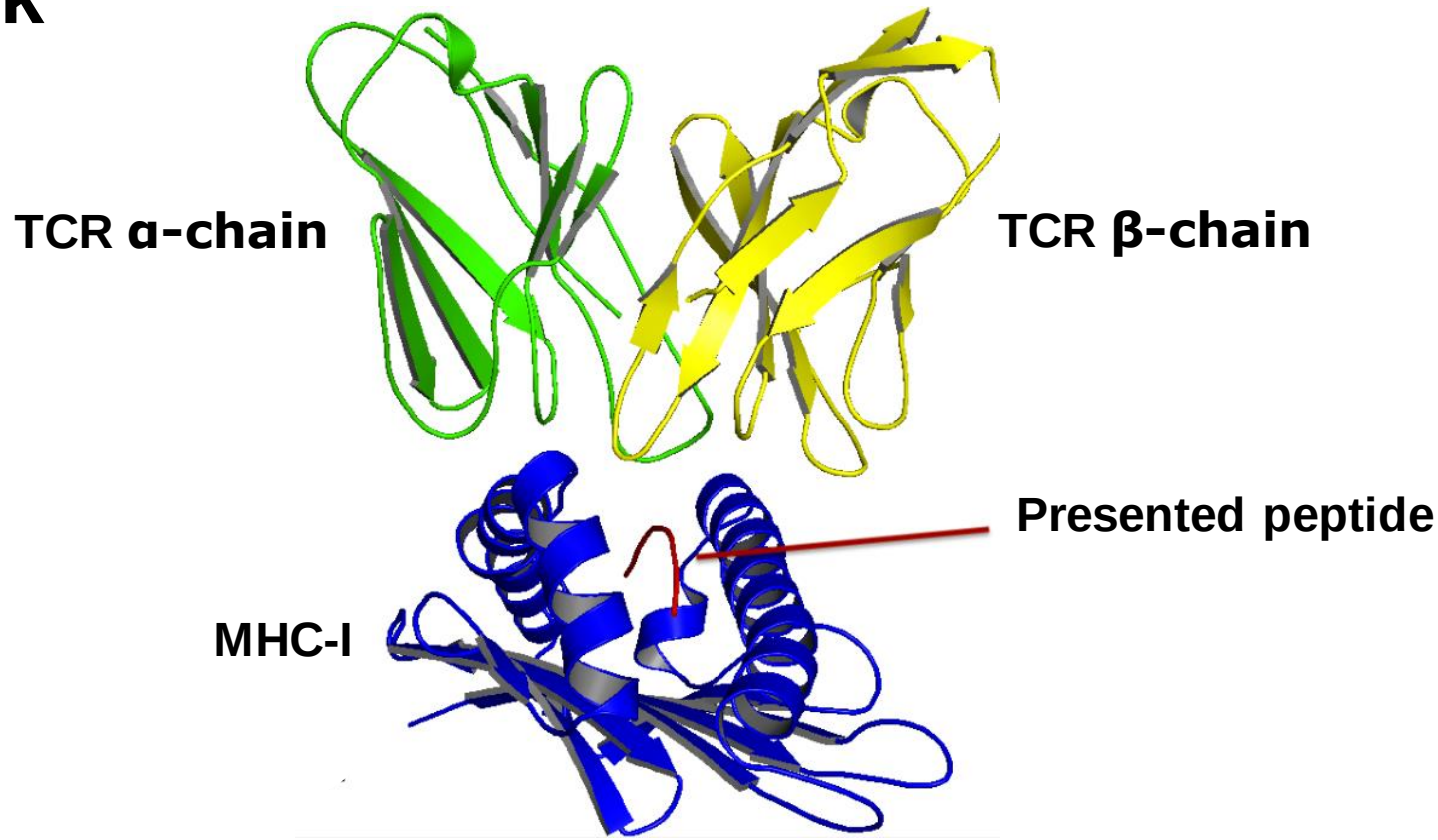
Magnus Høie
Research Assistant at University of Copenhagen

Anna-Lisa Schaap-Johansen
PhD student at DTU

TCR: T cell receptor

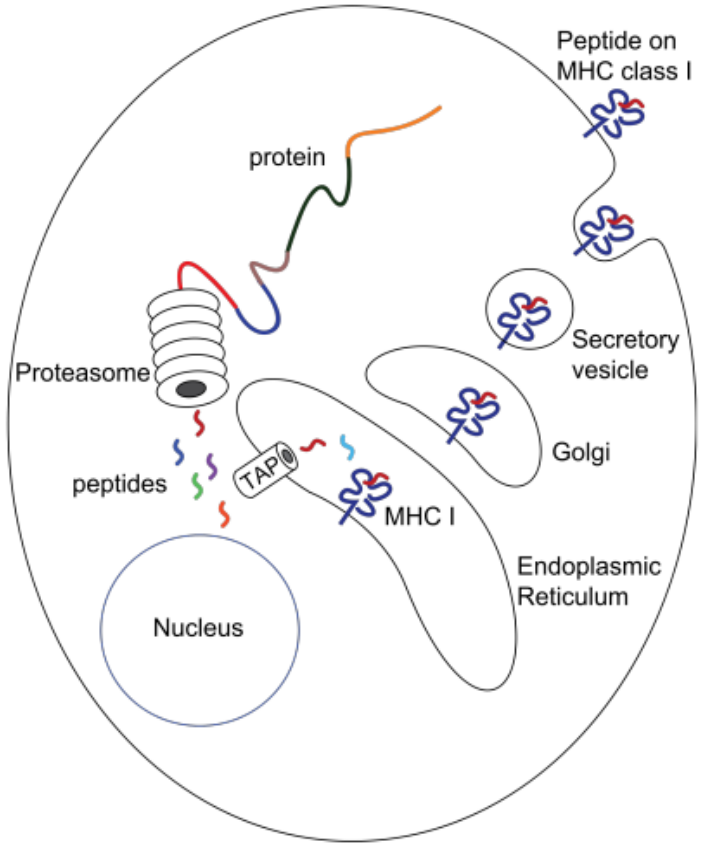
MHC: Major histocompatibility complex

The task



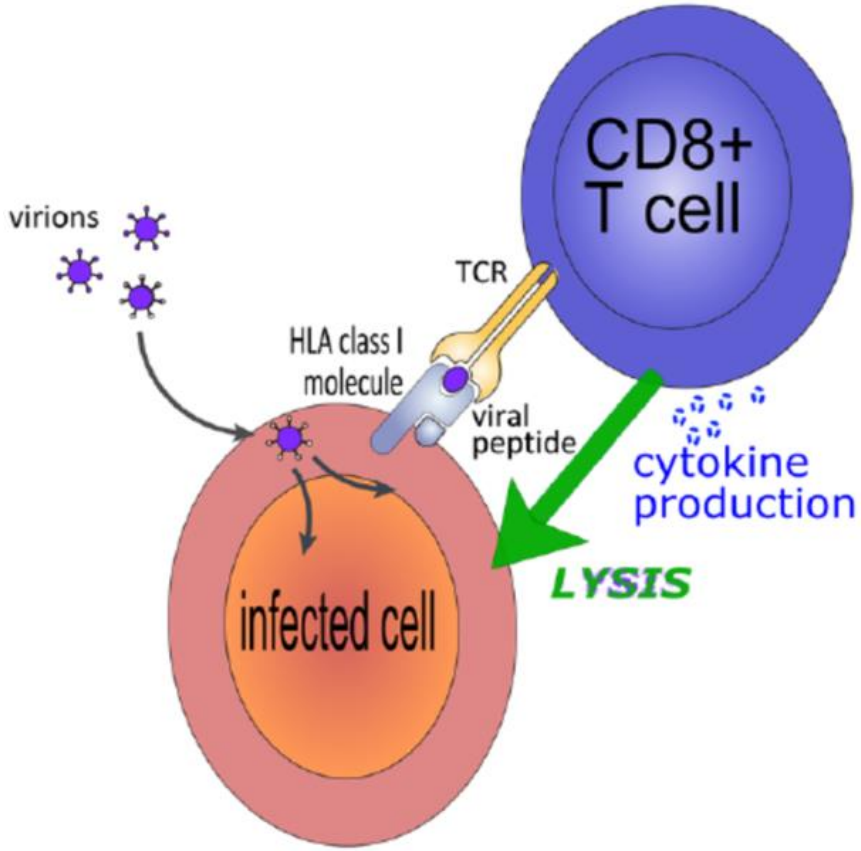
Predict binding of TCR and pMHC

Why is this important



All nucleated cells show peptide fragments on the cell surface

Why is this important



If a T-cell binds to the pMHC, it will kill the cell

* HLA: Human leukocyte antigen (a subclass of MHC)

What can we do with such a model?

- Understand the rules of TCR binding
- Immunotherapy: Find TCRs for cancer epitopes
- Vaccine development: Screening for epitopes

Previous TCRpMHC prediction models

NetTCR (2018)

Cold Spring Harbor Laboratory

bioRχiv

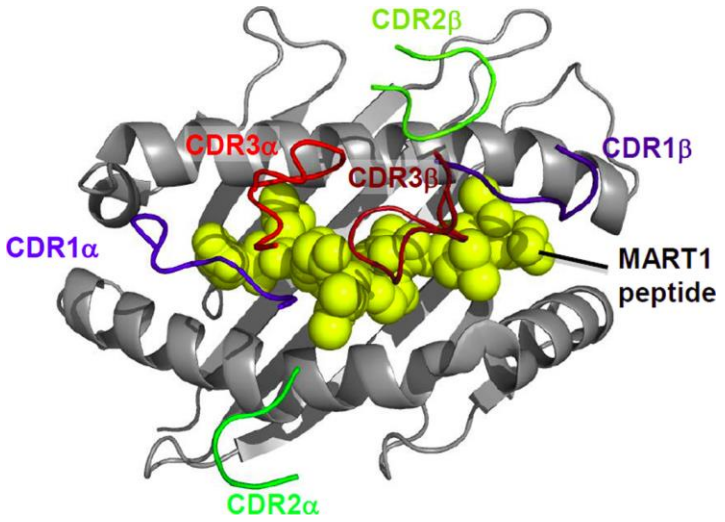
THE PREPRINT SERVER FOR BIOLOGY

NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks

Vanessa Isabell Jurtz, Leon Eyrich Jessen, Amalie Kai Bentzen, Martin Closter Jespersen, Swapnil Mahajan, Randi Vita, Kamilla Kjærgaard Jensen, Paolo Marcatili, Sine Reker Hadrup, Bjoern Peters, Morten Nielsen

doi: https://doi.org/10.1101/433706

- Sequence-based
- Convolutional neural network
- Trains only on TCR CDR3 and peptide
- Struggles to predict on new peptides

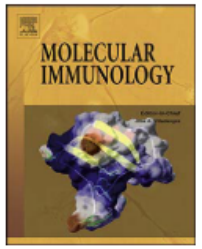


Molecular modeling and force field scoring (2018)

Contents lists available at ScienceDirect

Molecular Immunology

journal homepage: www.elsevier.com/locate/molimm



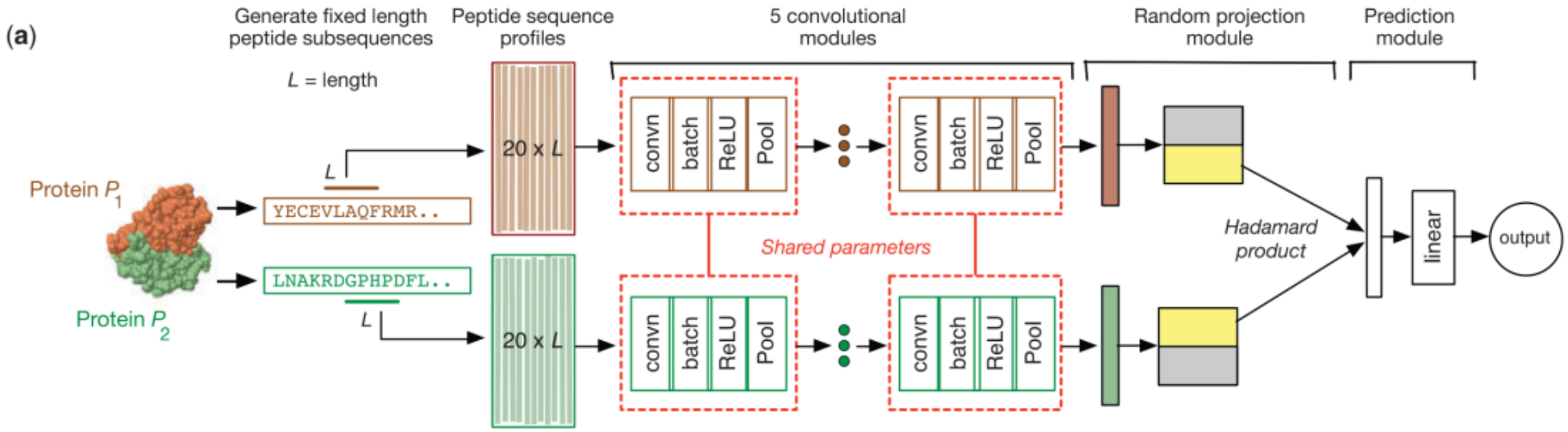
Identification of the cognate peptide-MHC target of T cell receptors using molecular modeling and force field scoring

Esteban Lanzarotti[a], Paolo Marcatili[b], Morten Nielsen[a,b,*]

[a] Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, San Martín, Buenos Aires, Argentina
[b] Department of Bio and Health Informatics, Technical University of Denmark, Building 208, Kemitorvet, 2800 Lyngby, Denmark

- Uses molecular modeling and FoldX and Rosetta energy terms
- Limited performance

Siamese network (2019)



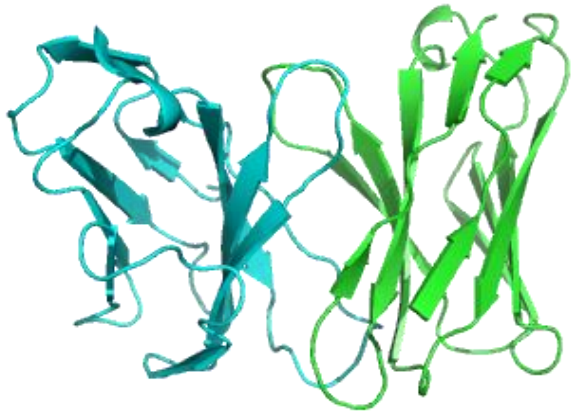
The dataset

The dataset

- Molecular models and energy terms calculated from these
  - TCRs have a very conserved structure => easy to model

- New dataset

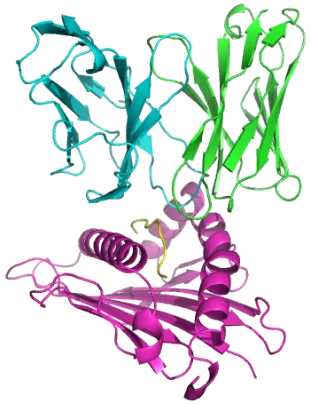
- We are excited to see your results!



Preparation of dataset

Sequences of TCRpMHC complexes

Modeling

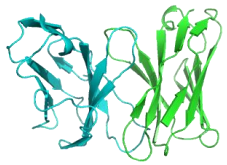


FoldX

Rosetta

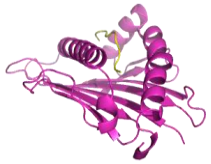
Split

TCR



Rosetta

pMHC



Rosetta

Dataset with sequences, FoldX and Rosetta energy terms

Starting point

- ~ 8,000 TCR-p-MHC-complex sequences
  - 25% positive complexes coming from VDJdb and IEDB
  - 75% negative complexes coming from the 10X genomics dataset and swapped complexes

- Only HLA-02-01
- 18 different peptides. GILGFVFTL (influenza) ~60%, GLCTLVAML (herpesvirus) ~10%, NLVPMVATV (herpesvirus) ~10%

Modeling and energy calculations

- The complexes were modeled using TCRpMHCmodels
- Relaxation and energy calculation using FoldX
- Relaxation and energy calculation using Rosetta (global and per-residue)
  - Electrostatic, h-bond, repulsive, attractive etc...

The features

Feature	Columns	Feature encoding
Amino acid	1-20	One-hot encoded
Rosetta per-residue energy terms	21-27	
FoldX energies	28-33	Constant, one value
Rosetta global energy complex	34-40	Constant, one value
Rosetta global energy TCR	41-47	Constant, one value
Rosetta global energy pMHC	48-54	Constant, one value

CSV

The files

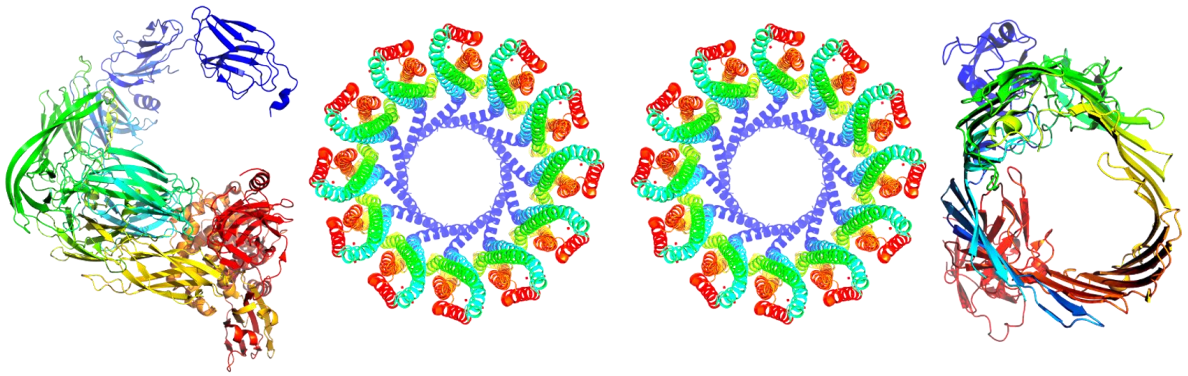
- *P1_input.npz, P2_input.npz …*
  - The input data
  - A 3D-array. (sample, position, features)
- *P1_labels.npz, P2_labels.npz …*
  - The labels
  - A 1D-array. 0 meaning negative, 1 meaning positive

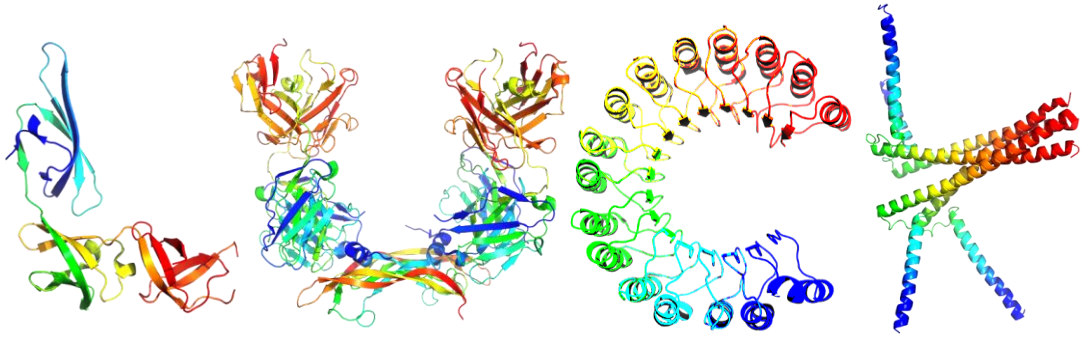
- Sequence have been padded in order to make up for the differences in length

- [0:179] MHC
- [179:192] peptide
- [192:] TCR

Practicals

- Ask questions in the Slack channel
- Check-in Saturday at 17





M. Howarth, Nat. Struct. Mol. Biol. 2015