Statistical learning approaches to modelling T cell response at the molecular level

Barbara Bravi

Department of Mathematics, Imperial College London

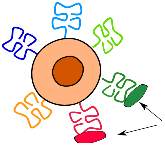
Joint work with: S. Cocco, R. Monasson, T. Mora, A. Walczak (ENS Paris)

Mini-colloque: Information et Biologie, SFP, 3 July 2023

Imperial College
London

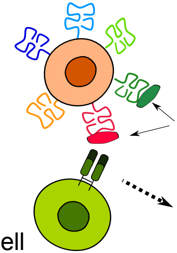
T cell response

Infected cell



Antigens: protein fragments 'foreign' to the organism

T cell response



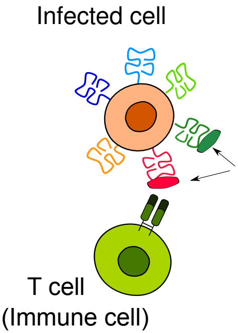
Infected cell

Antigens: protein fragments 'foreign' to the organism

T cell
(Immune cell)

Immune recognition

T cell response



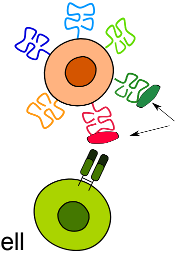
Infected cell

T cell
(Immune cell)

Antigens: protein fragments 'foreign' to the organism

What proteins are recognized?
Targets for Vaccines

T cell response

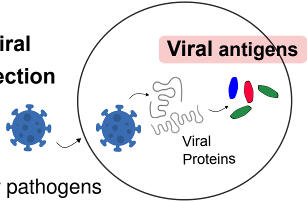


Infected cell

T cell
(Immune cell)

Antigens: protein fragments 'foreign' to the organism

What proteins are recognized?
Targets for Vaccines



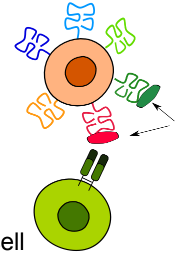
Viral infection

Viral antigens

Viral
Proteins

...or other pathogens

T cell response



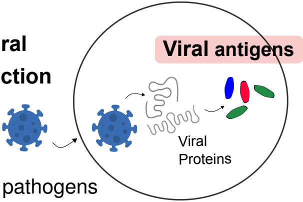
Infected cell

T cell
(Immune cell)

Antigens: protein fragments 'foreign' to the organism

What proteins are recognized?
Targets for Vaccines

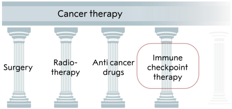
Viral infection



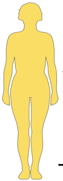
Viral antigens

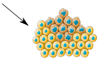
...or other pathogens

Cancer Immunotherapy



From 2018 Nobel Prize in Physiology or Medicine



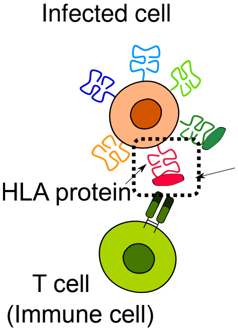


Cancer neoantigens

Tumour tissue
contains mutated proteins

Targeted by anti-cancer vaccines

Mechanisms: protein-protein binding



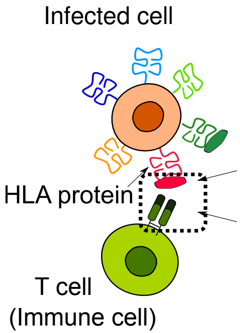
Infected cell

HLA protein

T cell
(Immune cell)

Antigen - HLA protein binding: 'Presentation'

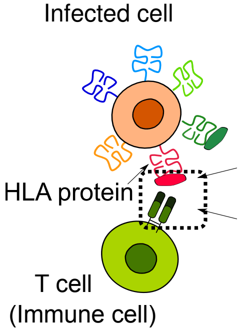
Mechanisms: protein-protein binding



Antigen - HLA protein binding: 'Presentation'

Receptor - antigen binding: recognition

Mechanisms: protein-protein binding



Infected cell

HLA protein

T cell
(Immune cell)

Antigen - HLA protein binding: 'Presentation'

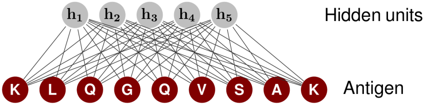
Receptor - antigen binding: recognition

Aim: build models of immune interactions from available protein data

The statistical learning approach

Restricted Boltzmann Machines (RBMs)

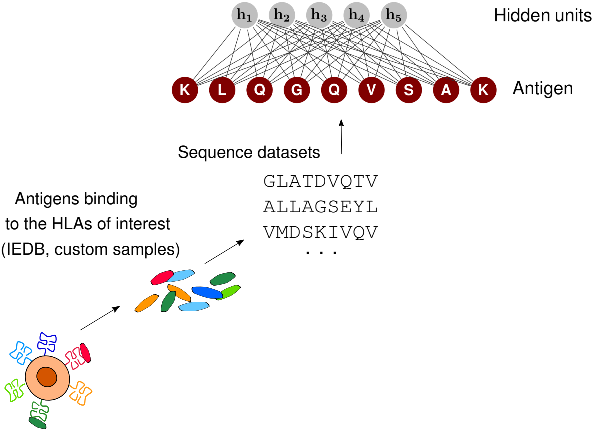
(Smolensky 1986, Hinton 2002, Tubiana et al. 2019)



Predicting antigen presentation

Restricted Boltzmann Machines (RBMs)

(Smolensky 1986, Hinton 2002, Tubiana et al. 2019)



Hidden units

Antigen

Sequence datasets

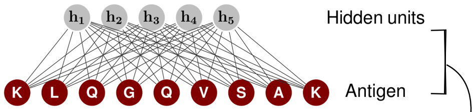
```
GLATDVQTV
ALLAGSEYL
VMDSKIVQV
   . . .
```

Antigens binding
to the HLAs of interest
(IEDB, custom samples)

Predicting antigen presentation

Restricted Boltzmann Machines (RBMs)

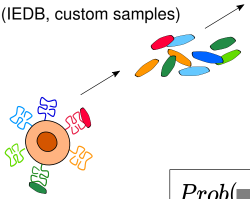
(Smolensky 1986, Hinton 2002, Tubiana et al. 2019)



Sequence datasets

```
GLATDVQTV
ALLAGSEYL
VMDSKIVQV
   . . .
```

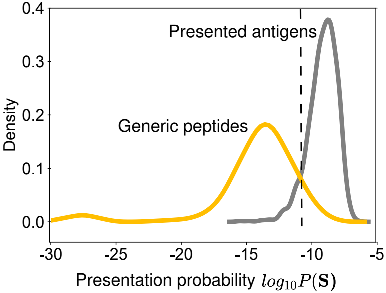
Antigens binding
to the HLAs of interest
(IEDB, custom samples)



Learn from sequence data
a probability of presentation

$Prob($■ $>$ 🟨$)$ = 0.973

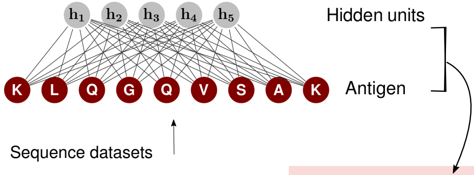
RBM discriminates presented antigens



Predicting antigen presentation

Restricted Boltzmann Machines (RBMs)

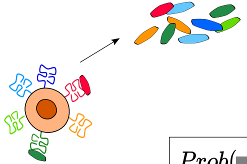
(Smolensky 1986, Hinton 2002, Tubiana et al. 2019)



Sequence datasets

```
GLATDVQTV
ALLAGSEYL
VMDSKIVQV
   . . .
```

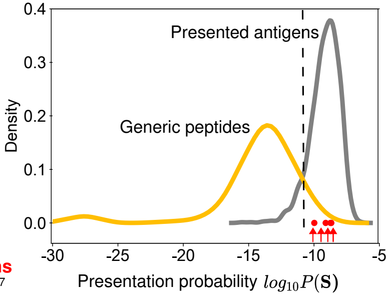
Antigens binding
to the HLAs of interest
(IEDB, custom samples)



Learn from sequence data a probability of presentation

Prob(■ > 🟨) = 0.973

RBM discriminates presented antigens



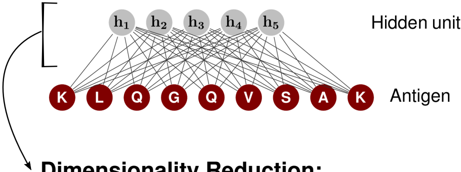
Prediction of cancer neoantigens

Validated neoantigens from Marty et al. 2017

RBM low-dim. representation

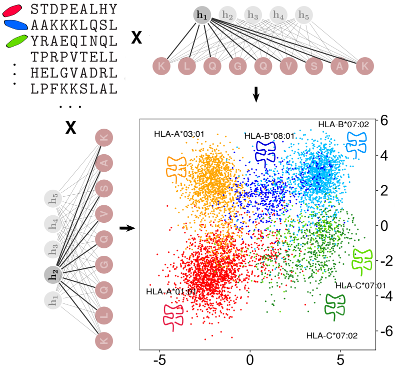
Restricted Boltzmann Machines (RBMs)

(Smolensky 1986, Hinton 2002, Tubiana et al. 2019)



Dimensionality Reduction:

like PCA but non-linear!

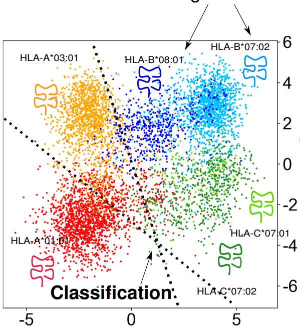

STDPEALHY
AAKKKLQSL
YRAEQINQL
TPRPVTELL
HELGVADRL
LPFKKSLAL
. . .

Low-dim. representation

Clusters =
HLA binding specificity

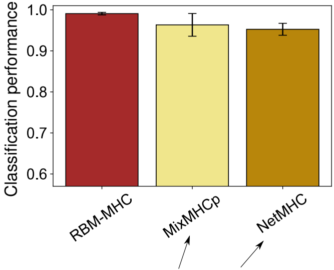
Prediction of HLA specificity

Trained using a few HLA annotations



Classification

Accurate prediction

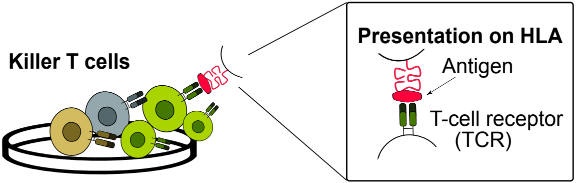


State-of-the-art methods

B. Bravi, J. Tubiana, S. Cocco, R. Monasson, T. Mora, A.M. Walczak, *RBM-MHC: a semi-supervised machine-learning method for sample-specific prediction of antigen presentation by HLA-I alleles*, *Cell Systems* (2021)

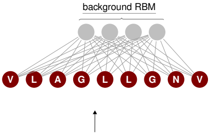
Antigen immunogenicity

Presentation alone is not immunogenicity!



Only a fraction of HLA-presented antigens are immunogenic (promote a T cell response).
Immunogenicity prediction: still low success rate (Wells et al. 2020, Buckley et al. 2022)

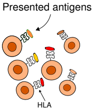
Transfer learning with ‘differential’ units



HLA-A*02:01 presented antigens

```
VLAGLLGNV
GILGFVFTL
FLCLFLLPSL
SLQQELAHM
ALYGVWPLLL
ALAESIRPL
```

Data from IEDB



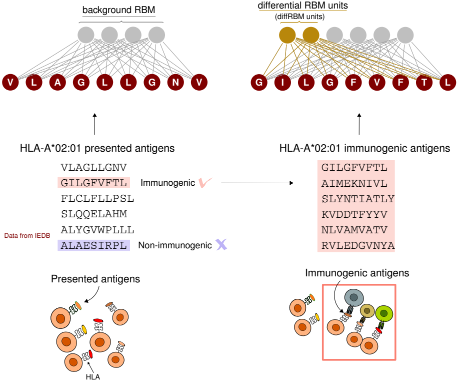
Model of presentation

(RBM-MHC, Bravi et al. 2021)

captures background constraints
(binding affinity to HLA)

B. Bravi, A. Di Gioacchino, J. Fernandez-de-Cossio-Diaz, A.M. Walczak, T. Mora, S. Cocco, R. Monasson, pre-print Biorxiv 2022.12.06.519259v1 (2022)

Transfer learning with ‘differential’ units



HLA-A*02:01 presented antigens

VLAGLLGNV
GILGFVFTL Immunogenic
FLCLFLLPSL
SLQQELAHM
ALYGVWPLLL
ALAESIRPL Non-immunogenic

HLA-A*02:01 immunogenic antigens

GILGFVFTL
AIMEKNIVL
SLYNTIATLY
KVDDTFYYV
NLVAMVATV
RVLEDGVNYA

Model of presentation

(RBM-MHC, Bravi et al. 2021)

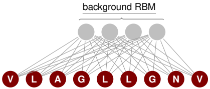
captures background constraints
(binding affinity to HLA)

Model of immunogenicity

captures *differences*
(patterns of immunogenicity)

B. Bravi, A. Di Gioacchino, J. Fernandez-de-Cossio-Diaz, A.M. Walczak, T. Mora, S. Cocco, R. Monasson, pre-print Biorxiv 2022.12.06.519259v1 (2022)

Transfer learning with ‘differential’ units

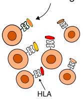


HLA-A*02:01 presented antigens

VLAGLLGNV
GILGFVFTL Immunogenic ✓
FLCLFLLPSL
SLQQELAHM
ALYGVWPLLL
ALAESIRPL Non-immunogenic ✗

Data from IEDB

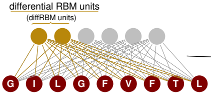
Presented antigens



Model of presentation

(RBM-MHC, Bravi et al. 2021)

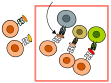
captures background constraints
(binding affinity to HLA)



HLA-A*02:01 immunogenic antigens

GILGFVFTL
AIMEKNIVL
SLYNTIATLY
KVDDTFYYV
NLVAMVATV
RVLEDGVNYA

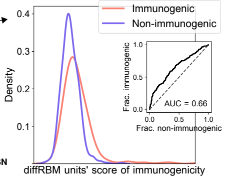
Immunogenic antigens



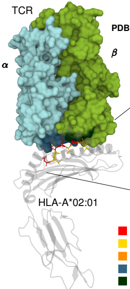
Model of immunogenicity

captures *differences*
(patterns of immunogenicity)

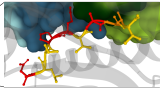
Discriminate immunogenic vs non-immunogenic



Contact prediction



PDB 3GSN

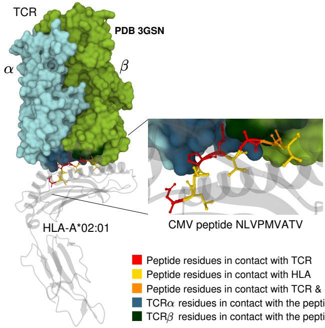


CMV peptide NLVPMVATV

🟥 Peptide residues in contact with TCR
🟨 Peptide residues in contact with HLA
🟧 Peptide residues in contact with TCR & HLA
🟦 TCRα residues in contact with the peptide
🟩 TCRβ residues in contact with the peptide

B. Bravi, A. Di Gioacchino, J. Fernandez-de-Cossio-Diaz, A.M. Walczak, T. Mora, S. Cocco, R. Monasson, pre-print Biorxiv 2022.12.06.519259v1 (2022)

Contact positions in resolved structures:

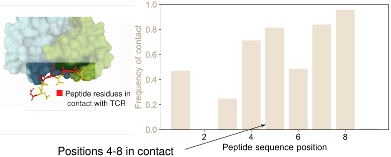


CMV peptide NLVPMVATV

- 🟥 Peptide residues in contact with TCR
- 🟨 Peptide residues in contact with HLA
- 🟧 Peptide residues in contact with TCR & HLA
- 🟦 TCRα residues in contact with the peptide
- 🟩 TCRβ residues in contact with the peptide

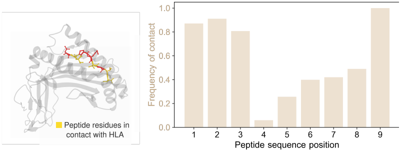
HLA-A*02:01-specific peptides

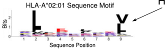
Peptide-TCR contacts



Positions 4-8 in contact
(Schmidt et al 2021, Milighetti et al. 2021)

Peptide-HLA contacts

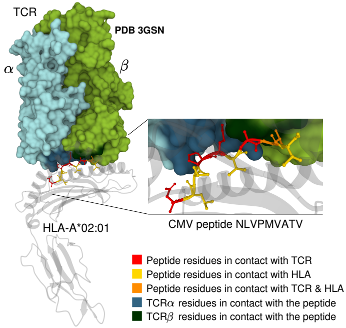




High frequency amino acids
at anchor positions (2,9)

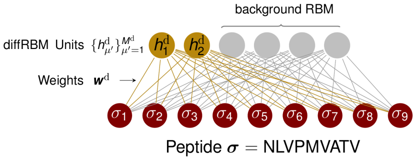
Constraints on statistics of immunogenic antigens should reflect contacts

Contact positions in resolved structures:


TCR
PDB 3GSN
α
β
HLA-A*02:01
CMV peptide NLVPMVATV
🟥 Peptide residues in contact with TCR
🟨 Peptide residues in contact with HLA
🟧 Peptide residues in contact with TCR & HLA
🟦 TCRα residues in contact with the peptide
🟩 TCRβ residues in contact with the peptide

Model prediction:

diffRBM architecture


background RBM
diffRBM Units $\{h^{\rm d}_{\mu'}\}_{\mu'=1}^{M^{\rm d}}$
$h^{\rm d}_1$
$h^{\rm d}_2$
Weights $\boldsymbol{w}^{\rm d}$ →
σ_1
σ_2
σ_3
σ_4
σ_5
σ_6
σ_7
σ_8
σ_9
Peptide $\boldsymbol{\sigma}$ = NLVPMVATV

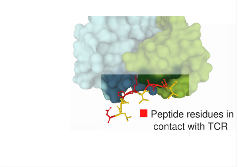
Single-site importance factors

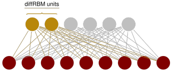
$$T_i(\sigma_i) = \underbrace{g_i^{\rm d}(\sigma_i)}_{\text{related to amino acid frequency difference between immunogenic and presented}} + \sum_{\mu'=1}^{M^{\rm d}} \underbrace{w^{\rm d}_{i\mu'}(\sigma_i)\langle h_{\mu'}|\boldsymbol{\sigma}\rangle}_{\text{captures correlations between positions}}$$

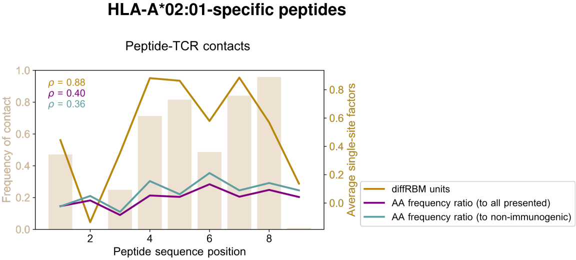
$\langle h_{\mu'}|\boldsymbol{\sigma}\rangle$: from $P(h_{\mu'}|I_{\mu'}(\boldsymbol{\sigma}))$, where $I_{\mu'}(\boldsymbol{\sigma}) = \sum_i w^{\rm d}_{i\mu'}(\sigma_i)$

We hypothesize that sites
at high $T_i(\sigma_i)$ are potential contacts

Structural interpretation





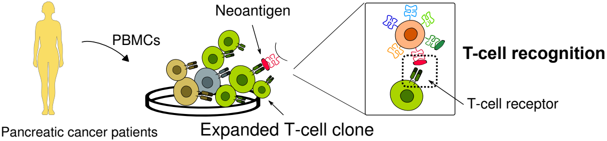


DiffRBM identifies positions 4-8 as the most relevant to immunogenicity without restricting a priori the input sequences to a subset of positions

Comparison: independent-site models
based purely amino acid (AA) frequency

T cell response specificity

Bravi et al., PLoS Comput. Biol. (2021)



PBMCs

Neoantigen

Expanded T-cell clone

T-cell recognition

T-cell receptor

Pancreatic cancer patients

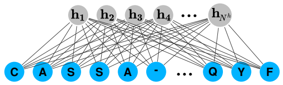
from Balachandran et al. 2017

Incorporate T-cell expansion

Receptor abundance	CDR3 Sequences
7920	CASSLNNQPQHF CASSLNNQPQHF CASSLNNQPQHF
1	CASSVEGATEAFF
40	CASSAGASGAARETQYF CASSAGASGAARETQYF

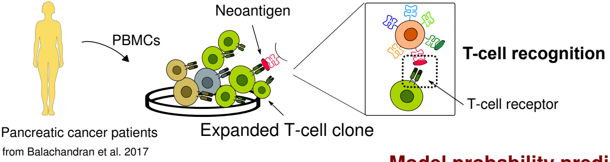
Infer probability of T-cell recognition

Restricted Boltzmann Machine



T cell response specificity

Bravi et al., PLoS Comput. Biol. (2021)



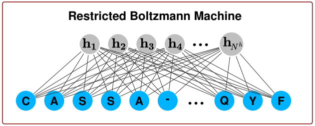
Pancreatic cancer patients

from Balachandran et al. 2017

Incorporate T-cell expansion

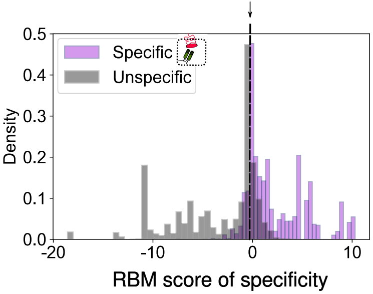
Receptor abundance	CDR3 Sequences
7920	CASSLNNQPQHF CASSLNNQPQHF CASSLNNQPQHF
1	CASSVEGATEAFF
40	CASSAGASGAARETQYF CASSAGASGAARETQYF

Infer probability of T-cell recognition



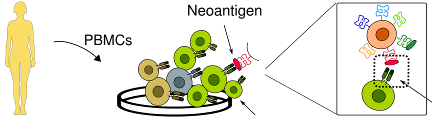
Model probability predicts specificity of response

RBM discriminates specific vs unspecific



T cell response specificity

Bravi et al., PLoS Comput. Biol. (2021)



Pancreatic cancer patients
from Balachandran et al. 2017

Expanded T-cell clone

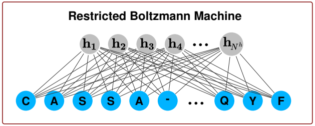
T-cell recognition

Uncover sequence patterns determining specificity

Incorporate T-cell expansion

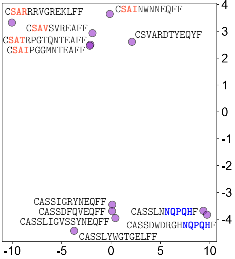
Receptor abundance	CDR3 Sequences
7920	CASSLNNQPQHF CASSLNNQPQHF CASSLNNQPQHF
1	CASSVEGATEAFF
40	CASSAGASGAARETQYF CASSAGASGAARETQYF

Infer probability of T-cell recognition

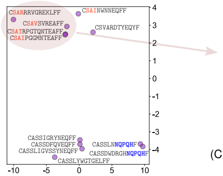


RBM representation space

Responding T cells



T cell response specificity



Bravi et al., PLoS Comput. Biol. (2021)

We want a summary metric of these convergence in amino acid patterns

Dissimilarity Index:

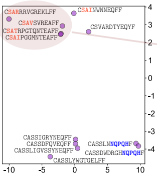
$$DI = \frac{1}{f} \qquad f = \frac{1}{T}\sum_{i<j} e^{-\left(\frac{d(\sigma_i,\sigma_j)}{\delta}\right)^2}$$

Distance between CDR3 σ_i and σ_j

(CDR3-only version of TCRdist from Dash et al 2017)

T cell response specificity

Bravi et al., PLoS Comput. Biol. (2021)



We want a summary metric of these convergence in amino acid patterns

Dissimilarity Index:

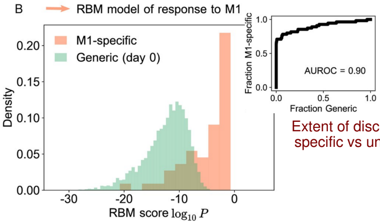
$$DI = \frac{1}{f} \qquad f = \frac{1}{T}\sum_{i<j} e^{-\left(\frac{d(\sigma_i,\sigma_j)}{\delta}\right)^2}$$

Distance between CDR3 σ_i and σ_j

(CDR3-only version of TCRdist from Dash et al 2017)

We consider: neoantigen-specific TCRs from PBMCs,

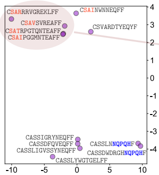
tetramer-sorted TCRs specific to viral epitopes (e.g. M1)



Extent of discrimination specific vs unspecific

T cell response specificity

Bravi et al., PLoS Comput. Biol. (2021)



We want a summary metric of these convergence in amino acid patterns

Dissimilarity Index:

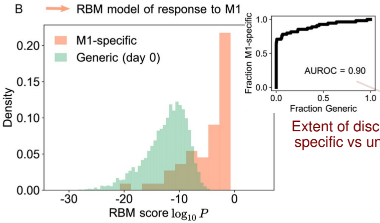
$$DI = \frac{1}{f} \qquad f = \frac{1}{T}\sum_{i<j} e^{-\left(\frac{d(\sigma_i,\sigma_j)}{\delta}\right)^2}$$

Distance between CDR3 σ_i and σ_j

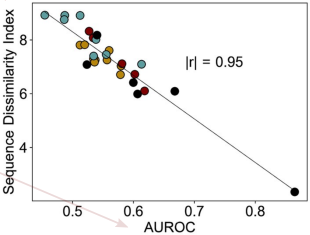
(CDR3-only version of TCRdist from Dash et al 2017)

We consider: neoantigen-specific TCRs from PBMCs,

tetramer-sorted TCRs specific to viral epitopes (e.g. M1)



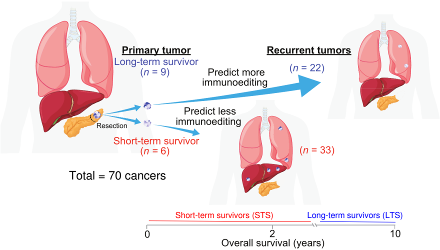
Extent of discrimination specific vs unspecific



Immunoediting in pancreatic cancer

Large-scale study on immunoediting in PDAC

conducted by Vinod Balachandran, Benjamin Greenbaum, Marta Łuksza, Zachary Sethna, Luis Rojas, Jayon Lihm and others

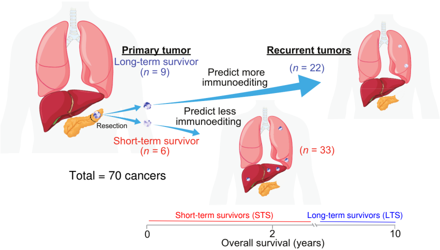


Łuksza* , Sethna*, Rojas*, Lihm, Bravi et al., Nature (2022)

Immunoediting in pancreatic cancer

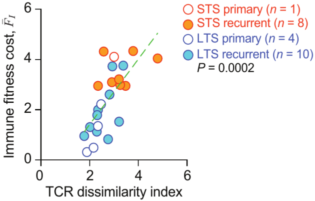
Large-scale study on immunoediting in PDAC

conducted by Vinod Balachandran, Benjamin Greenbaum,
Marta Łuksza, Zachary Sethna, Luis Rojas, Jayon Lihm and others



Our T-cell specificity analysis provides additional statistical evidence of immunoediting

+ Immunoediting



+ T-cell specificity

Łuksza* , Sethna*, Rojas*, Lihm, Bravi et al., Nature (2022)

Summary

Statistical learning approach based on Restricted Boltzmann Machines

- Amino acid patterns $\rightarrow$ scores of molecular specificity (Antigen presentation, immunogenicity, T cell response)

Summary

Statistical learning approach based on Restricted Boltzmann Machines

- Amino acid patterns → scores of molecular specificity (Antigen presentation, immunogenicity, T cell response)
- Transfer-learning approach extracts biologically interpretable features on immunogenicity

ACKNOWLEDGEMENTS

Antigen presentation and immunogenicity
S. Cocco, R. Monasson, T. Mora, A. Walczak, A. Di Gioacchino, J. Fernandez-De-Cossio-Diaz (ENS Paris), J. Tubiana (Tel Aviv University)

T cell response specificity
S. Cocco, R. Monasson, T. Mora, A. Walczak (ENS Paris)
V. P. Balachandran, B. D. Greenbaum, Y. Elhanati, L. Rojas Exposito, Z. Sethna, J. Lihm, D. Hoyos (MSKCC, New York)
M. Łuksza (Mount Sinai, New York)

Thank you for your attention!