HLA-EpiCheck : A B-cell epitope prediction tool on HLA antigens using molecular dynamics simulation data

Diego Amaya-Ramirez, Romain Lhotte, Cedric Usureau, Magali Devriese, Malika Smaïl-Tabbone, Jean-Luc Taupin, Devignes Marie-Dominique

HLA-EpiCheck : A B-cell epitope prediction tool on HLA antigens using molecular dynamics simulation data

Amaya-Ramirez Diego[1], Lhotte Romain[2], Usureau Cedric[2], Devriese Magali[2], Smail-Tabbone Malika[1], Taupin Jean-Luc[2] and Devignes Marie-Dominique[1]
[1] Université de Lorraine, CNRS, INRIA, LORIA, Nancy, F-54000, France; [2] INSERM, Hôpital Saint Louis, F-75010 Paris
Contact : diego.amaya-ramirez@inria.fr

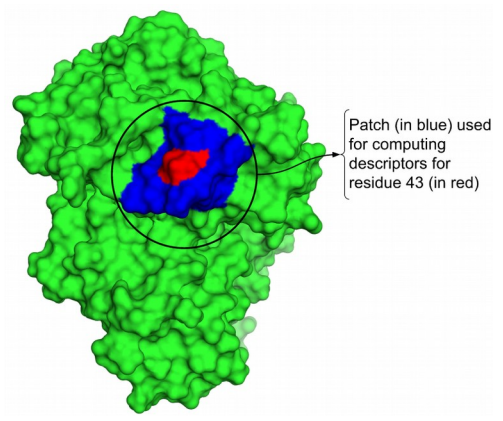
Introduction

In the context of organ transplantation, recipient's antibodies against donor-specific **HLA** antigens are the main reason for transplant loss. The prediction of B-cell (antibody) epitopes on **HLA** antigens is therefore a key challenge on the way to improving the matching step between donor and recipient from a structural point of view. Here, we present **HLA-EpiCheck**, a B-cell epitope prediction tool that relies on an unprecedented dataset of short Molecular Dynamics (**MD**) simulations of 207 **HLA** antigens. We use hydrophobic properties, electrostatic charges, flexibility and solvent accessibility as descriptors calculated on patches sampled from **MD** trajectories. Then, we train an Extremely Randomized Trees machine learning model. This model outperforms the state-of-the-art **DiscoTope 3.0 tool** [1] for B-cell epitope prediction on HLA antigens.

Dataset generation

MD data :

- **207 HLA antigens** (the most prevalent in the European population).
- Initial templates for **MD simulations** were obtained from **PDB** [3] or **Alphafold2** [4].
- 10ns MD simulations performed for each HLA antigen and 500 frames extracted from the last 5ns of the trajectory.
- **NAMD3 software [5], CHARMM36 [6]** force field and **TIP3P** explicit solvent model were used for MD runs.



Patch Descriptors Dictionary (18)				
Name	Description (per patch)	Descriptor type		Value range
		Static	Dynamic	
H_central	Hydrophobicity of central residue	x		[-4.5, 4.5]±
H_patch_min	Minimum residue hydrophobicity	x		[-4.5, 4.5]±
H_patch_max	Maximum residue hydrophobicity	x		[-4.5, 4.5]±
H_patch_avg	Average of all residue hydrophobicities	x		[-4.5, 4.5]±
Pos_central	Positive charge of central residue	x		[0, 1]
Pos_patch	Sum of positive charges of all residues	x		[0, 1]
Neg_central	Negative charge of central residue	x		[-1, 0]
Neg_patch	Sum of negative charges of all residues	x		[-1, 0]
S_central_min	Minimum RSASA* of central residue over MD		x	[0, 91] Å^2
S_central_max	Maximum RSASA of central residue over MD		x	[21, 138] Å^2
S_central_avg	Average RSASA of central residue over MD		x	[10, 116] Å^2
S_patch_min	Weighted** average of all residue minimum RSASAs over MD		x	[2, 49] Å^2
S_patch_max	Weighted average of all residue maximum RSASAs over MD		x	[9, 90] Å^2
S_patch_avg	Weighted average of all residue average RSASAs over MD		x	[6, 65] Å^2
F_central	N-RMSF*** of central residue		x	[-2, 13]
F_patch_min	Minimum residue N-RMSF		x	[-2, 3]
F_patch_max	Maximum residue N-RMSF		x	[-2, 15]
F_patch_Avg	Weighted average of all residue N-RMSFs		x	[-2, 6]

* RSASA : Relative Accessible Surface Area.
** Weights (of residue values) correspond to the frequency of presence of each residue in the patch over MD (% frames)
*** N-RMSF : Normalized-Root Mean Square Fluctuation during MD [7].
± Kyte-Doolittle hydrophobicity scale was used.

Dataset :

- **Patch definition**
  - Calculated for each frame of each MD run of each HLA antigen
  - Centered on every surface accessible residue (RSASA* $\geq 20\%$)
  - Composed of the surface accessible residues within a radius of 7Å from the central residue (centers of mass where used to measure distances).
- **18 descriptors:** (8 static and 10 dynamic).
- **2 classes:** epitope and non-epitope.
  - Epitope labels = confirmed eplets in the **HLA Eplet Registry database** [2].
  - Non epitope = neither confirmed nor putative eplet.
  - Patches composed solely of conserved residues (with respect to the locus) were discarded.

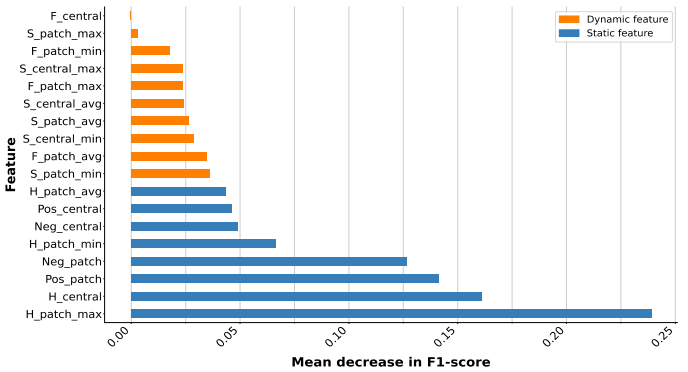
		A	B	C	DP	DQ	DR	Total
# antigens		34	59	17	31	30	36	**207**
Dataset	# epitope class	435	542	174	179	464	308	**2102**
	# non-epitope class	1302	3072	793	535	1646	626	**7974**
Training set (80%)	# epitope samples	349	434	133	137	377	243	**1673**
	# non-epitope class	1047	2458	605	432	1335	472	**6349**
Test set (20%)	# epitope class	81	101	44	47	88	66	**427**
	# non-epitope class	260	621	185	98	310	153	**1627**

Results

Model selection (on Training set)

Algorithm	F1-score for 10 repetitions of 10-fold cross-validations			
	With Dynamic features		Only static features	
	Class epitope	Class non-epitope	Class epitope	Class non-epitope
Logistic regression	0.17 +/- 0.05	0.88 +/- 0.01	0.05 +/- 0.01	0.76 +/- 0.002
Decision trees	0.57 +/- 0.08	0.88 +/- 0.03	0.48 +/- 0.06	0.74 +/- 0.07
Random Forest	0.64 +/- 0.11	0.92 +/- 0.03	0.50 +/- 0.06	0.83 +/- 0.05
Extremely Randomized Trees (Extra Trees) [8]	0.74 +/- 0.10	0.93 +/- 0.03	0.67 +/- 0.08	0.92 +/- 0.04

Feature Importance for the Extra Trees model



Comparison with DiscoTope 3.0 (on Test set)

	HLA-EpiCheck		DiscoTope	
	Class epitope	Class non-epitope	Class epitope	Class non-epitope
Precision	0.94	0.94	0.41	0.79
Recall	0.78	0.99	0.05	0.98
F1-score	0.85	0.97	0.09	0.88

Conclusion and Perspectives

- ✓ Modest contribution of MD descriptors on prediction performance.
- ✓ Poor performance of DiscoTope 3.0 : no HLA antigen in the dataset used to train DiscoTope.
- ✓ Find better descriptors for MD : analyse false negative epitopes.
- ✓ Try instance-based predictors such as K-NN.

References

[1] M. Høie *et al.* DiscoTope-3.0 - bioRxiv, 2023.
[2] R. Duquesnoy *et al.* 16th IHIW. Int. J. Immunogenetics, 2017.
[3] H. M. Berman et al. PDB. Nucleic Acids Research, 2000.
[4] J. Jumper et al. AlphaFold2. Nature, 2021.
[5] J. Phillips. *et al.* NAMD3. Journal of Chemical Physics, 2020.
[6] J. Huang *et al.* CHARMM36m.Nat Methods, 2017.
[7] K. Dong-Gun *et al.* J. of Chem. Information and Modeling 2021.
[8] P. Geurts *et al.*. Extremely randomized trees. Machine Learning, 2006.

Acknowledgements

DAR is recipient of an Inria-Inserm doctoral fellowship. This project is funded by ANR EPI-HLA N° ANR-22-CE15-0036-03.
Experiments presented in this poster were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr), as well as the MBI-DS4H platform hosted by Inria/Loria and funded by CPER IT2MP (Contrat Plan État Région, Innovations, Technologiques, Modélisation & Médicine Personalisée), including a FEDER co-funding.