Bachelor's Degree Final Project
Héctor Puigdomènech Gómez | Bachelor's Degree in Genetics

Analysis of the latest **pan-specific** bioinformatic tools for the discovery of MHC class II binding epitopes

Universitat Autònoma de Barcelona

Final Project Tutor: Daniel Yero Corona

ABSTRACT: Performance of bioinformatic approaches to discover HLA-II binding epitopes is still far from best. HLA-II polymorphism lead to the creation of the so-called pan-specific tools, which are able to predict binding epitopes for MHC alleles without previous affinity data. In fact, this kind of tools do perform slightly worse than tools focused on HLA-I molecules. The aim of this review is to perform a presentation of the fortes as well as the limitations in the latest pan-specific tools in order to give some **enlightenment** for the progress in the field.

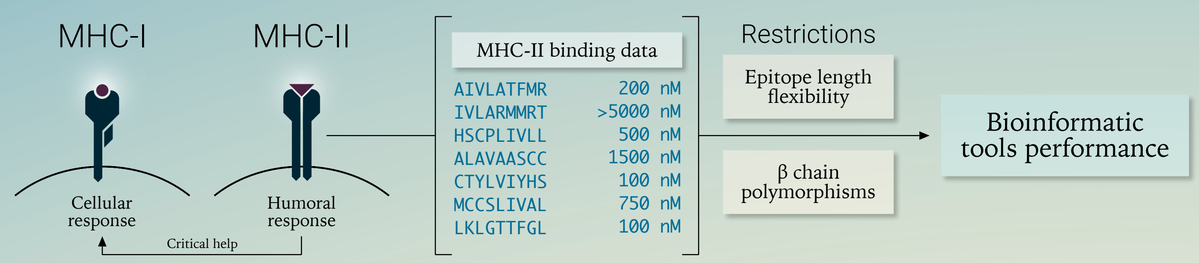


Figure 1. General schematic introduction into bioinformatic tools for the prediction of MHC-II binding epitopes.

Introduction

The aim to improve existing vaccines and to create new ones for remaining infections lead researchers into the use of omics data and bioinformatics to predict which proteins have the best potentiality to induce a big immune response, thus establishing the approach of Reverse Vaccinology.

MHC-II molecules present a higher degree of complexity for the prediction of binding epitopes due to its extremely polymorphic beta chain as well as the length variation binding peptides have because of the openness of the binding groove (Guo, Luo, and Zhu 2013). These polymorphisms are responsible for the development of pan-specific tools, which use information concerning several alleles in order to predict binders not only for known alleles but also for alleles of which there is no binding information (Zhang, Udaka, et al. 2012).

The aim of this review is to analyse important features which represent some of the latest pan-specific bioinformatic tools created to discover new MHC-II binding epitopes.

Methodology

The following conditions were declared to choose which tools would be analysed: selected tools must have been created or updated within the last five years and its only function must be the discovery of MHC class II binding epitopes —so tools which include other functionalities will be discarded—, and they must be pan-specific.

Finally, a PUBMED search was performed using the query (*"HLA-DR" OR "MHC-II") AND "pan-specific"* and the result was restricted to items publicated in the last five years.

Results

TEPITOPEpan

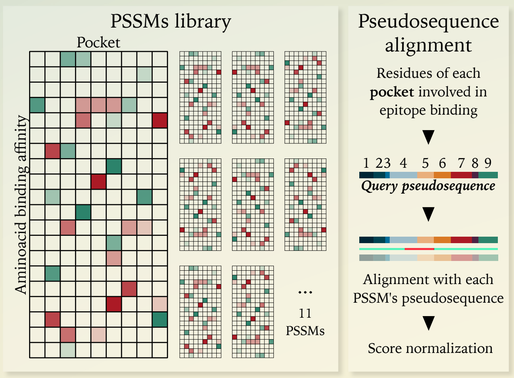


Figure 2. Graphic explanation of TEPITOPEpan's workflow. Once the query allele PSSM is generated, it is used to predict binders.

TEPITOPEpan (Zhang, Chen, et al. 2012) works with the PSSMs generated for TEPITOPE (Sturniolo et al. 1999). Each PSSM consists of 20 rows, one for each aminoacid, and 9 columns which correspond to significant pockets within the binding groove with regard to determining peptide binding specificity. TEPITOPEpan calculates the contribution of these matrices to generate a PSSM for an unkown allele based on the similarity between the unkown allele protein sequence and all the alleles for which TEPITOPE generated a PSSM.

Predivac

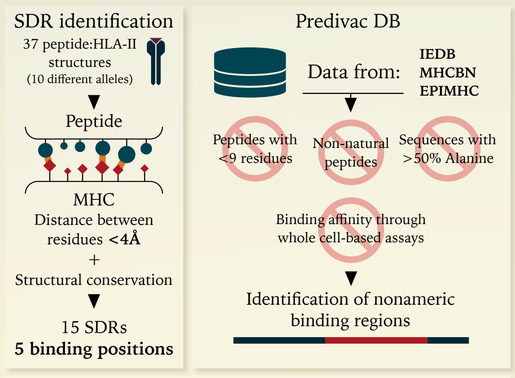


Figure 3. Steps of Predivac's workflow. Using PWMs, which are similar to TEPITOPEpan's PSSMs, Predivac assigns values to peptide nonamers.

Oyarzún et al. (2013) based their tool in the Specificity Determining Residues (SDRs) concept and was developed using high affinity data to infer peptidic properties which are related to epitope promiscuity and immunodominance (Sirskyj et al. 2011). These data were used to build a specific database for Predivac, which is called PredivacDB and, in fact, contains 2695 sequences which cover 29 class II MHC alleles.

MHC2SKpan

$$K((x,y)(x',y')) \triangleq K_{allele}(x,x') \cdot K_{peptide}(y,y')$$

$$K_{allele}(x,x') = K^{\varphi}_{|x|,\sigma_a}(x,x') = exp(-\frac{\sum_{i=1}^{l} \|\varphi(s_i) - \varphi(s'_i)\|^2}{2\sigma_c^2})$$

Aminoacid differences under an encoding scheme

In order to consider different epitope lengths

$$K_{peptide}(x,x') = K_{MHC2SK}(y,y',L,\sigma_c) =$$

$$\sum_{l=L'}^{min(|y|,|y'|)} \sum_{i=1}^{|y|-l+1} \sum_{j=1}^{|y|-l+1} exp(-\frac{\sum_{k=0}^{l-1} \|\varphi(y_{i+k}) - \varphi(y'_{i+k})\|^2}{2\sigma_c^2})$$

L' is the minimum length of compared epitopes

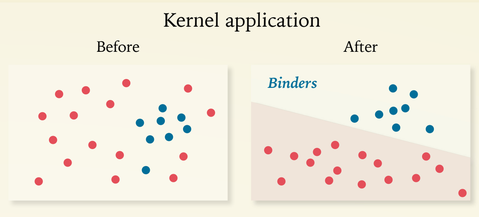


Figure 4. MHC2SKpan's tool basic points.

Developped by Guo, Luo, and Zhu (2013), MHC2SKpan is a kernel-based method that considers variation of epitope lengths because nonamers may not be sufficient to significantly predict class II MHC-binding proteins. Kernels permit a fast classification of data as they use features and transform them into easy intepretable information which permits distinction of binders and non-binders.

NetMHCIIpan-3.1

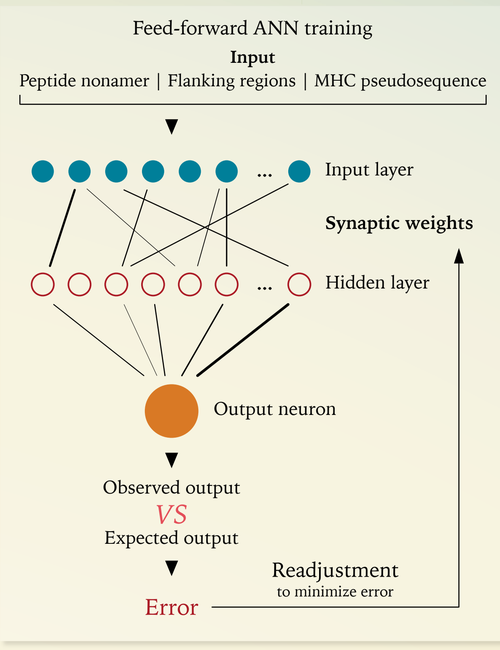


Figure 5. Illustration of NetMHCIIpan's ANN system training.

NetMHCIIpan (Andreatta et al. 2015) uses a feed-forward artificial neural network, which is composed of an input layer, a layer with different amounts of hidden neurons, and a single output neuron. Artificial neural networks are trained with input and output real data, so weights adjust in a way that the calculated output fits the real output value.

Discussion

Database dynamics

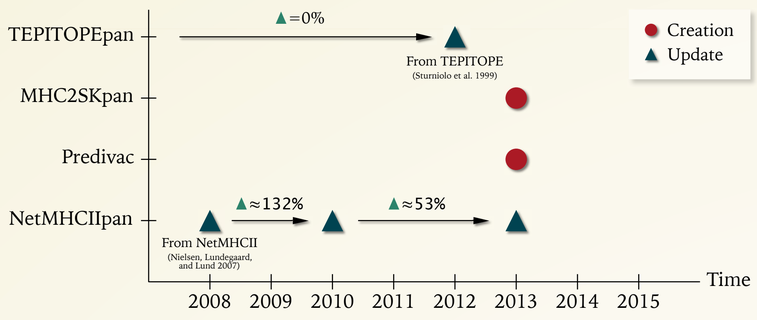


Figure 6. Database updating of each tool analysed. This shows a major drawback as almost all tools present an evident lack of updating, so they cannot improve their accuracy by handling a major amount of data.

As we can see in *figure 6*, NetMHCIIpan's database updates reflect the growth of data available for these prediction tools. If databases created for each tool were updated regularly, accuracy of prediction would probably increae, as more alleles and more diverse epitopes would be taken into consideration.

Pseudosequence generation

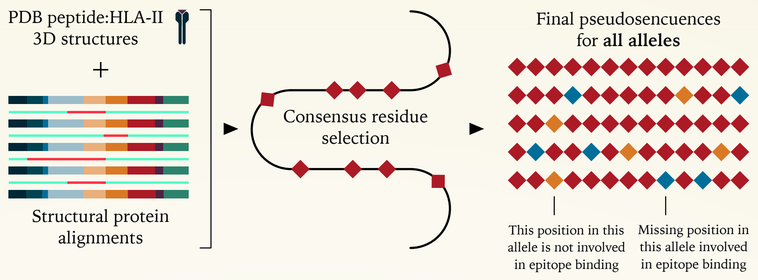


Figure 7. Graphic representation of pseudosequence generation in each of the tools analysed in this review.

Three of the four tools presented in this review use the same rule to create their pseudosequences, which is based on the distance between some peptide-bound MHC-II molecules whose structure was analysed. The fact is that some residues in an unknown allele could be included into the pseudosequence because of their position despite not being contact residues and be considered as such, thus introducing false results. Moreover, this leads to the fact that some contact residues could become absent in the pseudosequence, which implies loss of information.

Promiscuity evaluation

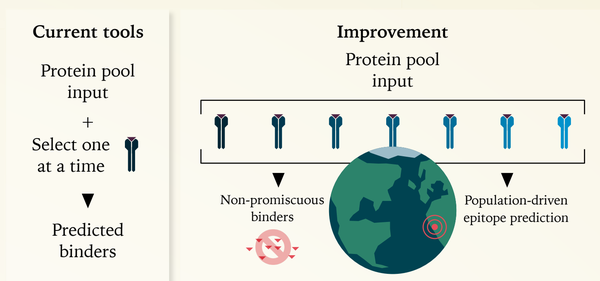


Figure 8. Impact of epitope promiscuity evaluation.

All of reviewed methods analyse binding epitopes for only one allele at once. Appropriate epitopes must be promiscuous and thus they must bind to different MHC alleles. It makes sense to consider that epitope-predicting tools should intrinsicly feature this characteristic in order to eliminate non-promiscuous predictions if needed or to simply know which MHC alleles could bind an epitope to, for example, select epitopes that would bind to the great majority of a determined population.

Algorhythm critics

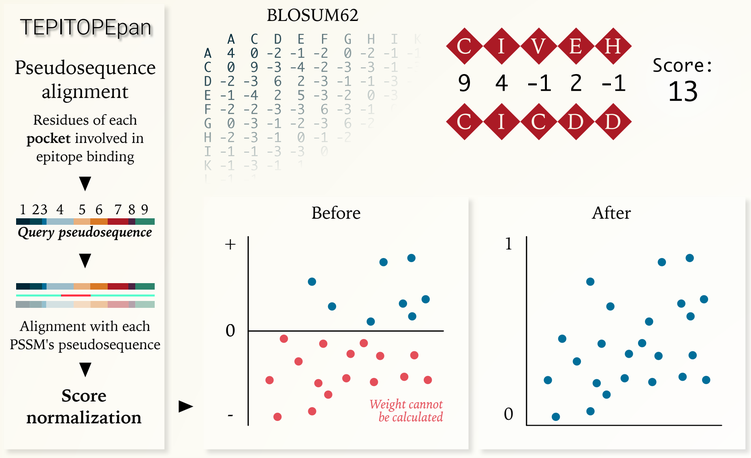
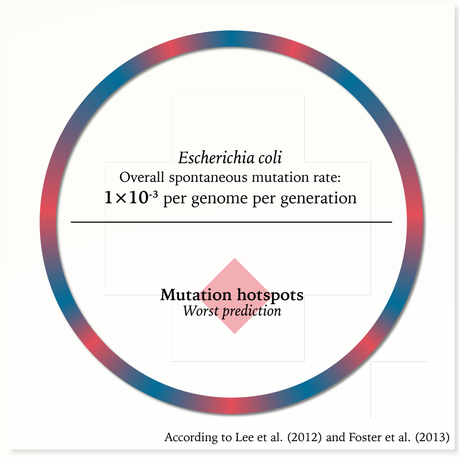


Figure 9. Score assignment and score normalization illustrated processes.

TEPITOPEpan method has been previously criticized by Shen, Zhang, and Wong (2013) because of the utilization of BLOSUM62 in order to calculate similarity between to sequences to establish the contribution of each allele to the PSSM for an unkown allele. BLOSUM62 similarity score can easily be negative, so the weight of negative-scoring comparisons could not be computed in TEPITOPEpan. In fact, Zhang et al. (2012) did consider this issue because their tool normalizes similarity scores in order to have only positive values.

Mutation as an escape way



According to Lee et al. (2012) and Foster et al. (2013)

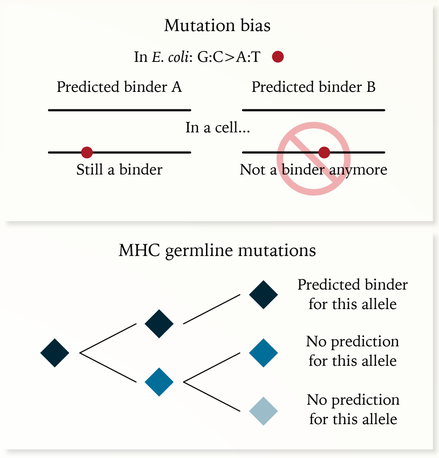


Figure 10. Considerations related to base substitutions in both epitopes and MHC molecules.

A subunit vaccine could result in failure because of a spontaneus aminoacid substitution in its target. A nucleotide mutation rate should be taken into consideration, as mutation rates in different parts of a genome tend to be different in a determined pattern (Foster et al. 2013). In addition to this, different microorganisms usually have different biases for certain types of mutations (Lee et al. 2012) that could be useful to predict those immunodominant and promiscuous epitopes whose mutations do not affect their immunodominancy. New MHC alleles generated by germline mutations should be taken into consideration to increase vaccine effectiveness.

Final remarks

TEPITOPEpan
AUC 0.6-0.85

Predivac
AUC ≈0.8

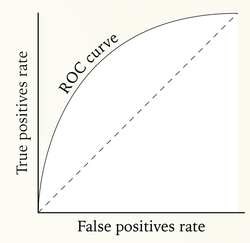
MHC2SKpan
AUC ≈0.8

NetMHCIIpan-3.1
AUC 0.8-0.9

AUC
Area under the receiver operating characteristic curve

AUC = 0.5
Random assignment

AUC = 1
No false positives



One of the most important conclusions of this review is the fact that there is a need of bioinformaticians who are able to handle and manipulate raw data in order to obtain more specific data for the training and to understand more about immune response and rational vaccine development.

Key references

Andreatta, Massimo, Edita Karosiene, Michael Rasmussen, Anette Stryhn, Søren Buus, and Morten Nielsen. 'Accurate Pan-Specific Prediction of Peptide-MHC Class II Binding Affinity with Improved Binding Core Identification.' *Immunogenetics* 67, no. 11–12 (November 2015): 641–50. doi:10.1007/s00251-015-0873-y.

Foster, P. L., A. J. Hanson, H. Lee, E. M. Popodi, and H. Tang. 'On the Mutational Topology of the Bacterial Genome.' *G3 Genes|Genomes|Genetics* 3, no. 3 (27 February 2013): 399–407. doi:10.1534/g3.112.005355.

Guo, Linyuan, Cheng Luo, and Shanfeng Zhu. 'MHC2SKpan: A Novel Kernel Based Approach for Pan-Specific MHC Class II Peptide Binding Prediction.' *BMC Genomics* 14, no. 5 (2013): 1.

Lee, H., E. Popodi, H. Tang, and P. L. Foster. 'Rate and Molecular Spectrum of Spontaneous Mutations in the Bacterium Escherichia Coli as Determined by Whole-Genome Sequencing.' *Proceedings of the National Academy of Sciences* 109, no. 41 (9 October 2012): E2774–83. doi:10.1073/pnas.1210309109.

Oyarzún, Patricio, Jonathan J Ellis, Mikael Bodén, and Boštjan Kobe. 'PREDIVAC: CD4+ T-Cell Epitope Prediction for Vaccine Design That Covers 95% of HLA Class II DR Protein Diversity.' *BMC Bioinformatics* 14 (14 February 2013): 52. doi:10.1186/1471-2105-14-52.

Shen, Wen-Jun, Shaohong Zhang, and Hau-San Wong. 'An Effective and Effecient Peptide Binding Prediction Approach for a Broad Set of HLA-DR Molecules Based on Ordered Weighted Averaging of Binding Pocket Profiles.' *Proteome Science* 11, no. Suppl 1 (7 November 2013): S15. doi:10.1186/1477-5956-11-S1-S15.

Sirskyj, Danylo, Francisco Diaz-Mitoma, Ashkan Golshani, Ashok Kumar, and Ali Azizi. 'Innovative Bioinformatic Approaches for Developing Peptide-Based Vaccines against Hypervariable Viruses.' *Immunology and Cell Biology* 89, no. 1 (January 2011): 81–89. doi:10.1038/icb.2010.65.

Zhang, Lianming, K. Udaka, H. Mamitsuka, and S. Zhu. 'Toward More Accurate Pan-Specific MHC-Peptide Binding Prediction: A Review of Current Methods and Tools.' *Briefings in Bioinformatics* 13, no. 3 (1 May 2012): 350–64. doi:10.1093/bib/bbr060.