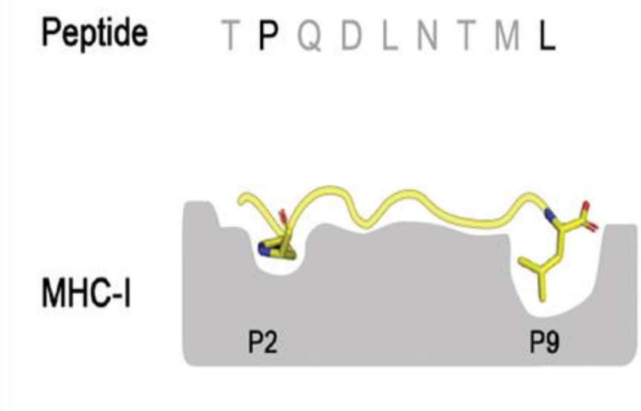
Automated annotation of MHC binding motifs

Credits
Made by : Carrel Ramil & Liu Jiayi
Supervisor : Tadros Daniel

Introduction

The main goal of our project is to automate the identification and annotation of MHC binding motifs using available datasets. By automating the process, we aim to help improve the prediction of MHC epitopes, which is crucial in the study of cancer and infectious diseases.
T cells recognize antigens via MHC epitopes. Antigens, which can be pathogen-derived or neo-antigens from tumor cells, bind selectively to MHC molecules. These antigen-presenting cells (APCs) then interact with T-lymphocytes, triggering an immune response. This mechanism is essential for developing effective treatments and vaccines, and by automating the annotation of these MHC binding motifs, we help to streamline this process.
We are looking at MHC class I epitopes, which are bind with short peptides 8 to 14 amino acids long.

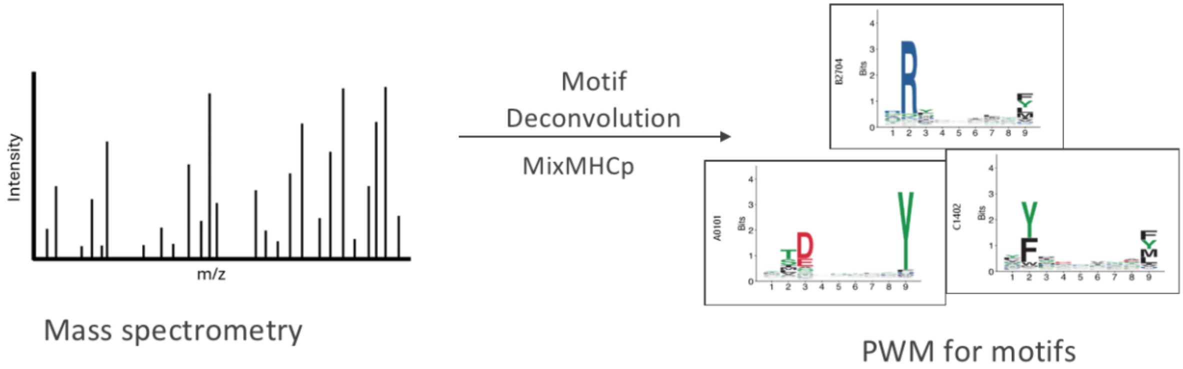


These molecules are called human leukocyte antigens (HLA). Antigen presentation prediction has a vast field of applications in medicine and research. This includes vaccine development, infectious disease research, allergen identification, personalized immunotherapy, and immunogenicity assessment.

Methods

Using 2 algorithms, MixMHCp and MixMHCpred, we can identify, characterize and predict the MHC-I motifs.
MixMHCp is a motif deconvolution algorithm. Motif deconvolution involves analyzing mass spectrometry data to identify and characterize MHC-I motifs. In the later steps of mass spectrometry, the target peptide fragments are eluted, giving us a profile of all the peptides based on their mass-to-charge ratio. Technically, motif deconvolution is the step where we

transform such mass spectrometry data to motif matrices called PWM.



MixMHCpred is a prediction tool, which was written and trained by our supervisor's lab members. This algorithm predicts which ligand is likely to bind to specific MHC-I molecules with different affinities. The researchers have trained the algorithm with a curated set of HLA ligands, with a focus on those that are naturally present.

$$S^{h}(X) = \frac{M^{(h,L)}(X) - C^{(h,L)}}{D^{(h,L)}}$$

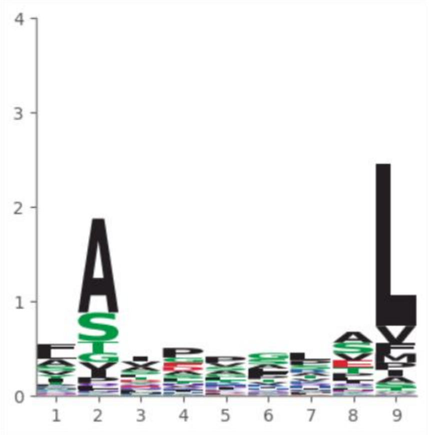
M(h,L): raw score of peptide X given by PWM representing the motif of allele h for L-mers
D(h,L): correction factors, S.D. of the scores of these peptides)
C(h,L): correction factor for adjusting the length distribution

The dataset is taken from the article : Spliced Peptides and Cytokine-Driven Changes in the immunopeptidome of Melanoma by Faridi et al. (2020), which studies about antigen recognition by CD8+ T cells by the pool of peptide antigens presented on the cell surface of HLA class I using, bioinformatic approaches to characterize the immunopeptidome of melanoma cells in the presence or absence of IFNγ. With over 60'000 peptides, we had to do some coding to treat the data and make the position weight matrix (PWM) and sequence logos.

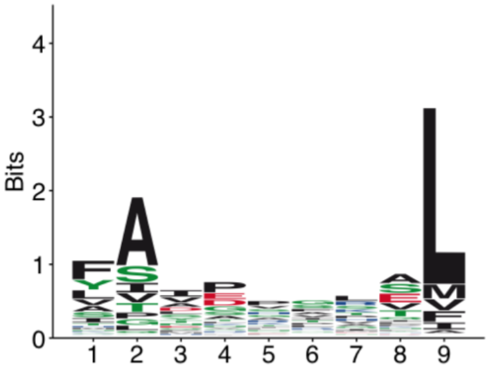
Results

For each allele, we obtain a sequence logo that we compare to the sequence logo of the MHCmotifatlas.org to see if our prediction are correct.

Results : Allele C0304

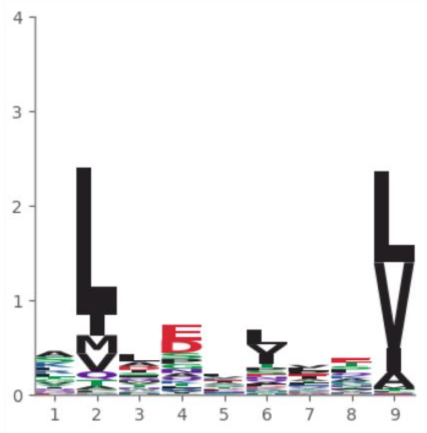


Our sequence logo

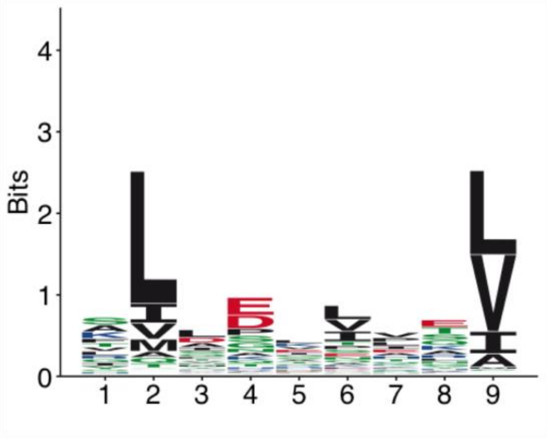


Sequence logo of MHCmotifatlas database

Results : Allele A0201

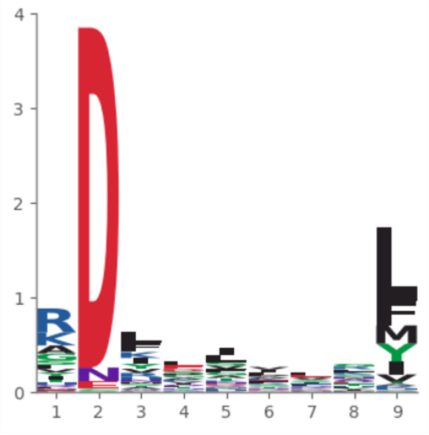


Our sequence logo

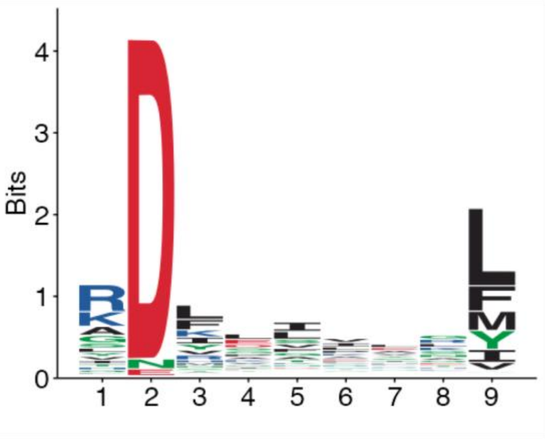


Sequence logo of MHCmotifatlas database

Results : Allele B4701

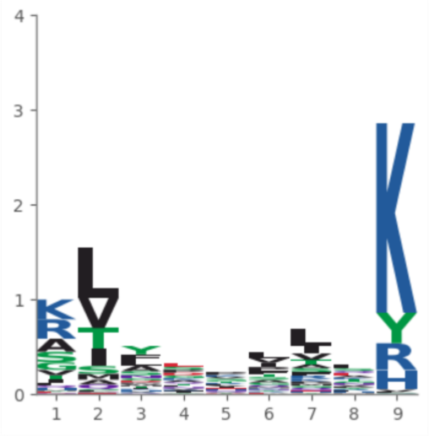


Our sequence logo

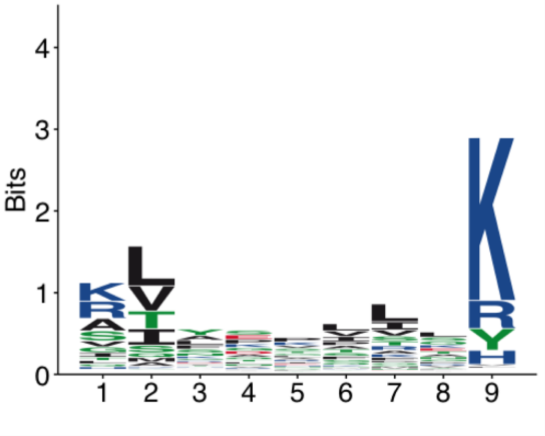


Sequence logo of MHCmotifatlas database

Results : Allele A0301

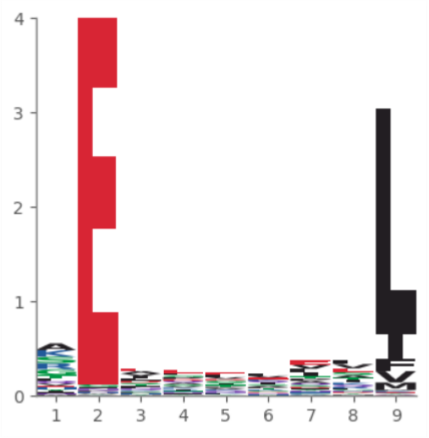


Our sequence logo

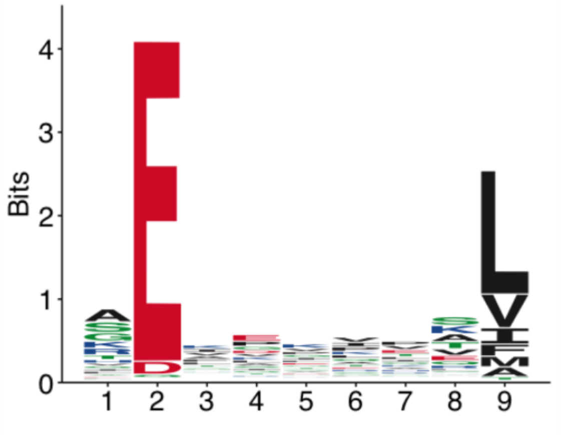


Sequence logo of MHCmotifatlas database

Results : Allele B4002

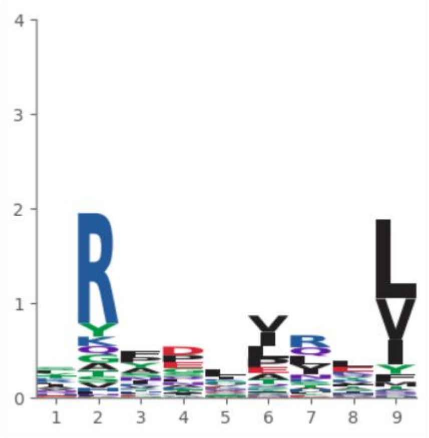


Our sequence logo

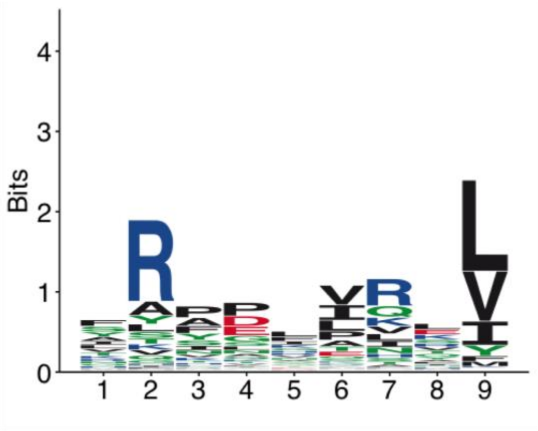


Sequence logo of MHCmotifatlas database

Results : Allele C0602



Our sequence logo



Sequence logo of MHCmotifatlas database

Conclusion

We have very similar results between our sequences logos and the sequence logos of MHCmotifatlas for ligands prediction, we only see some slight difference of certain amino acids. Here we had the ideal situation due to the limit of time we had for this project. We had 6 motifs expressed for 6 clustering in our algorithms.

But if this is not the case what do we do ? For example if we had 6 alleles expressed but we only found 3-5 clustering resulting in our algorithm. Then we would have done the clustering but we do not know which motif is which clustering so we would do the annotation by ourselves. Thus we will do everything step by step : which means motif by motif we would look which one is ideal depending on the number of cluster from 1 to 9. Also what we could do is to look at others dataset and get more relevant results.

Course Feedback (Module)

What we like about the course :

- Courses are usually followed by an exercise the same day with accessible assistance
- Timeline for the project: enough time to work on

What could have been improved :

- Details about the evaluation components are unclear: projects, presentations, etc.
- Schedule sometimes gets too busy: Can not digest 4 lectures in one day