R script to HLA epitope predictor based in matrix frequency: training and performance comparisons

Alessandra Lima da Silva[1], Leandro Martins de Freitas[2]

[1]Institute of Biological Sciences at UFMG and [2]Multidisciplinary Institute of Health at UFBA

Epitope prediction assists in the identification of candidate proteins to cause a greater immune response, selecting potential targets for studies and applications in the prevention, treatment, and diagnosis of diseases. The aim of this study was to evaluate the performance of a R script; trained with epitope retrieved from the Immune Epitope Database and Analysis Resource (IEDB) and compared with the NETMHCcons predictor. Peptides with 9 amino acid that binds to the MHC I supertype HLA-A * 02: 01 were selected from IEDB. The search in the IEDB returned 5964 epitopes validated in the database, but only 1022 were established according to criteria. Ligand matrix frequency was prepared using IEDB epitopes and not ligand matrix frequency was prepared using amino acid composition in the UniProt data bank, working as background probability. Using the in-house R script it was possible to select potential targets ligand to the MHC I. 1244 protein sequences of the hybrid strain *T. cruzi CL Brener* were obtained to use as target. Both predictions (R script and NetMHCcons), resulted in the same amount of peptides (633,005). Only peptides with strong binding prediction (cutoff of 0.84 in R script) were selected from the results generated by R script, resulting in 1589 peptides. The comparison with the results generated by NetMHCcons showed that approximately 56% of the peptides showed the same prediction compared with the script. Comparing the strong binding epitopes (0.84 or less) predicted with R script and randomly peptides among 633,005 returned only 2% shared peptides. The above findings are the result of the peptide prediction comparison of a new script in R language with a well-established server, NETMHCcons. The R language script is based on probability matrices, a simple and less sensitive analysis. Even so, 56% of the results were similar to the ones generated by NETMHCcons. The results may be tested further for their effectiveness of stimulating an immune response in both *in vivo* and *in vitro* experiments to support the *in silico* findings.