Accurate Prediction of Major Histocompatibility Complex Class II Epitopes by Sparse Representation via ℓ_1-Minimization

Clemente Aguilar-Bonavides[1], Reinaldo Sanchez-Arias[2], Cristina Lanzas[3]
[1] National Institute for Mathematical and Biological Synthesis, University of Tennessee, Knoxville, TN, US 37996-3410
clemen@nimbios.org
[2] Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX, US 79902
rsanchezarias@utep.edu
[3] Department of Biomedical and Diagnostic Sciences, University of Tennessee, Knoxville, TN, US 37996-3410
clanzas@utk.edu

Keywords: Major Histocompatibility Complex, Immunoinformatics, Epitope Prediction, Machine Learning.

The major histocompatibility complex (MHC) is responsible for presenting antigens (epitopes) on the surface of antigen-presenting cells (APCs). When pathogen-derived epitopes are presented by MHC class II on an APC surface, T cells may be able to trigger an immune response. Prediction of MHC-II epitopes is particularly challenging because the open binding cleft of the MHC-II molecule allows epitopes to bind beyond the peptide binding groove; then, the molecule is capable of accommodating peptides of variable length. We propose a novel classification algorithm to predict MHC-II called sparse representation via ℓ_1-minimization.

We obtained a collection of experimentally confirmed MHC-II epitopes from the Immune Epitope Database and Analysis Resource (IEDB) and applied our ℓ_1-minimization algorithm. To benchmark the performance of our algorithm, we compared our predictions against a SVM classifier. We measured sensitivity, specificity and accuracy; and used Receiver Operating Characteristic (ROC) analysis to evaluate our method's performance. The prediction performance of MHC-II epitopes of the ℓ_1-minimization algorithm was generally comparable and, in some cases, superior to the standard SVM classification method and overcame the lack of robustness of other methods with respect to outliers. While our method consistently favored DPPS encoding with the alleles tested, SVM showed a slightly better accuracy when "11-factor" encoding was used.