Combining Structure and Sequence Data to Predict Peptide-HLA Binding Affinity by Anja Conev

Abstract: Class I Human Leukocyte Antigen (HLA) alleles play a key role in immune response through presentation of peptides to T-cell lymphocytes. For that reason, the HLA pathway is very important for vaccinology efforts. We are currently working on the implementation of a computational infrastructure for identifying highly conserved peptides within SARS-CoV-2 proteins that could be used as targets for a broad-spectrum peptide vaccine in the context of specific HLAs. Here, we show one of the steps of this pipeline that involves the implementation of a scoring function that considers the pHLA structure, not only the sequence of the peptide. Our group developed a method using random forest classifier trained on features extracted from modeled structures of the pHLA complexes, which has shown competitive results compared to sequence-based methods. In this context, we combined the available structural and sequence data to build new models and use the data on 82,000 pHLA structures across 30 HLA alleles. We mapped the structures to pHLA binding affinity values and used these values as labels for our regression models. We engineered the features to be comprised of a structural distance matrix along with the chemical properties of the peptide sequence. The dataset was split into training/test set (20% of the data for each HLA was left out of the training phase) and trained a random forest regressor on the features and labels for each of the HLA alleles. Data distribution is such that each model is trained on at least 1,000 structures. Finally, we tuned the parameters of random forest regressors in a 5-fold validation setting. The coefficient of determination (R2) scored on the test set ranged from 0.74 to 0.99. Also, an increase of scores up to 1% was observed by lowering the minimum number of samples per leaf and minimum number of samples per split. Models trained on the combination of structure and sequence features of pHLA complexes highlights the power of leveraging sequence and structural data.
This work was funded by the National Science Foundation (NSF) (award number 2033262) and Rice University funds.