Mathematical Methods and Models in Biosciences
June 15–20, 2025, Sofia, Bulgaria
https://biomath.math.bas.bg/biomath/index.php/bmcs

Predicting cleavage sites of Cathepsin-V using machine learning models

Nida Ansari, Georgi Momekov, Ivan Dimitrov

Faculty of Pharmacy,
Medical University of Sofia, Bulgaria
n.ansari@pharmfac.mu-sofia.bg
gmomekov@pharmfac.mu-sofia.bg
idimitrov@pharmfac.mu-sofia.bg

Cathepsins are an enzyme family involved in the antigen processing pathway, especially in the protein cleavage in the endo/lysosome. Accurate prediction of protein substrate cleavage sites by the cathepsins could help to determine the potential antigens for processing by the antigen-presenting cells. In this study, advanced machine learning techniques were employed to create models for predicting Cathepsin V cleavage sites in the protein sequences. The models were trained on a dataset of peptides derived by mass spectrometry after the Cathepsin V cleavage of proteins. The peptides were encoded by Z-scale numerical descriptors, representing the physicochemical characteristics of the amino acids within peptide sequences. The models were built upon the peptide dataset after 10-fold cross-validation on the training set and validation on an external test set. Their performance was assessed using binary classification metrics, where the XGBoost, Support Vector Machine, Multilayer Perception, and Quadratic Discriminant Analysis showed superior results with high sensitivity. Feature selection was performed to identify the most important features of the amino acid's key positions within the peptides and to improve model performance. Future research will seek to utilize this approach to other cathepsins and combining in silico modelling with the knowledge in the biochemical processes in cells to expand the domain of antigen-processing research.