Mathematical Methods and Models in Biosciences
June 18-23, 2023, Pomorie, Bulgaria
https://biomath.math.bas.bg/biomath/index.php/bmcs

In silico immunogenicity prediction of viral proteins

Nikolet Doneva, Ivan Dimitrov

Drug Design and Bioinformatics Lab, Faculty of Pharmacy,
Medical University of Sofia, Bulgaria
ndoneva@pharmfac.mu-sofia.bg
idimitrov@pharmfac.mu-sofia.bg

Infectious diseases are primarily caused by viral proteins, making prevention crucial for effective disease control. Vaccines can help protect against many infectious diseases and reduce their spread. The first step in the modern vaccine design and development is the application of bioinformatics tools and computational techniques to identify potential vaccine targets, usually proteins. The identification of protective immunogens is the most important and vigorous initial step in the long-lasting and expensive process of vaccine design and development.

This study aims to derive machine learning models for immunogenicity prediction of viral proteins in order to update the VaxiJen server which was developed ten years ago in our lab. The study has three main phases: data collection and database creation, implementation of machine learning algorithms to find the best prediction models, update the VaxiJen webserver.

A dataset of viral immunogens acting as protective antigens in humans was created, based on exhaustive literature searches and validated sources such as PubMed, UniProt, NCBI, IEDB, and ClinicalTrials.gov. This dataset was implemented in a database, which includes information for 1782 viral antigens, T- and B-cell epitopes, on-going and completed clinical trials with viral immunogens from 31 viruses.

To create a negative set of non-immunogenic proteins from the same viral species, the immunogenicity of the proteins from the proteomes of each virus in the dataset was assessed using the VaxiJen 2.0 webserver. This resulted in a dataset of 468 non-immunogenic proteins. The dataset of viral immunogens and non-immunogens was divided into training and test sets. The primary structures of proteins were encoded by E-descriptors and transformed into uniform vectors by auto- and cross-covariance (ACC) calculations. Various machine learning algorithms were applied to the training set to derive models using the Weka software. The derived models were then validated using the test set. The best

predictive performance was observed with models derived from Xgboost (accuracy 89.85%), Random Forest (accuracy 87.15%), and Multilayer perceptron algorithms (accuracy 89.5%). The gain/ratio algorithm was used for attribute selection, resulting in a reduction in the number of attributes from 125 to 108 and an approximately 2% increase in the specificity of the selected algorithms. The last step to be fulfilled is to update the VaxiJen webserver.

Overall, this study provides an updated machine learning-based approach to predict the immunogenicity of viral proteins, which can aid in the development of effective vaccines.

Keywords: immunogenicity prediction, in silico methods, machine learning algorithms, viral proteins