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

Assessment of human proteins for potential tumour immunogenicity by *in silico* models

Stanislav Sotirov, I. Dimitrov

Faculty of Pharmacy, Medical University of Sofia, Bulgaria

113660@students.mu-sofia.bg
idimitrov@pharmfac.mu-sofia.bg

Cancer is a major cause of death globally and has a major impact on societies across the world. Currently, existing cancer treatments possess several disadvantages, mainly associated with a lack of distinction between healthy and cancerous tissues. Thus, a more specific treatment approach is required. Cancer vaccines are a form of targeted immunotherapy that aims to prevent the occurrence or threat of existing cancer by educating the immune system about what cancer cells look like.

Immunogenic tumour proteins are predominantly mutated self-proteins recognized by the immune system and thus elicit strong, specific antitumor immune responses. However, spontaneous immune recognition of these mutations is inefficient. Because of that, immunogenic tumour proteins are promising candidates as personalized vaccines in the treatment of cancer. Ten years ago, a server for immunogenicity prediction of proteins of tumour origin, named VaxiJen, was developed in our lab. The models for immunogenicity prediction were derived by partial least square (PLS) discriminant analysis on sets of known immunogenic and non-immunogenic proteins. The primary structures of proteins were encoded by z-descriptors and transformed into uniform vectors by auto- and cross-covariance (ACC) calculations.

Our study aimed to collect a comprehensive dataset of human tumour antigens, which is to be used for deriving in silico models for the assessment of the immunogenicity of tumour proteins.

We manually searched the literature for human tumour antigens and looked for their protein sequences in the UniProt protein database. As a result, a set of 5199 antigens and 546 protein sequences was collected. We also collected a mirror set of 547 non-immunogenic tumour proteins using BLAST search of proteins of the human proteome against the collected dataset with tumour antigens and a subsequent check with the VaxiJen 2.0 web server for their tumour immunogenicity. The sets of immunogenic and non-immunogenic proteins were combined and randomly split into training and test sets in a ratio of 4:1.

The properties of each amino acid in the protein's primary structure were described by E-descriptors and the protein sequences were transformed into arrays with different lengths. An auto-cross covariance transformation of the protein arrays into uniform numerical vectors was applied to form a data matrix ready for modelling. We applied different machine learning algorithms using the Weka software on the data matrix of the training set and validated the derived models with the test set. The performance of the derived models was further evaluated.

The study found that the Quadratic Discriminant analysis, Random Forest, and Radial Basis Function classifier algorithms show the best predictive performance. The selected models will be implemented in a new version of the VaxiJen web server for the assessment of tumour immunogenicity of proteins and the discovery of potential candidates for cancer vaccines. This is an important step towards the development of effective cancer vaccines and personalized cancer treatment.