

Development of a method of numerical representation of peptides using machine learning to screen compounds of interest in combating viral infections.

Paula Lopes Cascabulho ^{1,2}, Manuela Leal da Silva ^{1,3}, Maria Fernanda Ribeiro Dias ^{1,4}

1 National Institute of Metrology, Quality and Technology (INMETRO), Brazil

2 Catholic University of Petropolis (UCP), Brazil

3 Federal University of Rio de Janeiro (UFRJ), Brazil

4 State Secretary of Espírito Santo (SEDU), Brazil

Abstract- The rapid mutation rate of the human immunodeficiency virus 1 (HIV-1) gives it the ability to quickly adapt to the host's immune responses and acquire resistance to antiretroviral therapy. These are factors that place Brazil at the center of one of the largest HIV-1 epidemics in the Western world and with a great diversity of virus subtypes circulating in the country. The proteolytic processing of gag and gag-pol polyproteins by viral protease is crucial for viral evolution. The importance of this highly ordered and complete processing for the maturation and infectivity of HIV-1 becomes a promising target for the discovery of new antiretrovirals. In this context, we use computational biology techniques to develop tools capable of selecting and extracting information necessary for our problem. This approach allows for a dialogue between biology and computation. We propose a method that applies machine learning tools on natural substrates of viral proteases, as a strategy to train a search engine based on their molecular characteristics. This makes it possible to screen for new compounds with a focus on inhibiting infectious processes. The developed method can be applicable in the planning of drugs against other viral infections. The project uses the polyproteins Gag and Gag-pol, natural substrates of the HIV-1 protease, in training systems based on machine learning methods. Polyproteins were fragmented into peptides with a length of 3 to 8 amino acids, and represented using the AAindex database. AAindex uses numerical indices that represent physicochemical properties of amino acid residues. The numerical representation was elaborated through a matrix of objects in which the lines represent the peptides and the columns the attributes (physicalchemical properties). The search and choice of attributes was performed using the pyaaisc library, implemented in python language, using keywords that align with the correlated properties. Subsequently, these attributes were grouped using unsupervised machine learning algorithms (K-means and hierarchical clustering). Initially we obtained 5 groups of physicochemical properties - Volume, Hydrophobicity, Weight, Polarity and Van der Waals interactions. These presented groups of: 12, 33, 62, 3 and 3 related attributes, respectively. Some inconsistencies, such as outliers were observed in the formation of these groups. These are factors that need to be analyzed with greater caution, due to the inference of noise in the representation system. Thus, statistical functions are being analyzed to decide on the most adequate numerical representation. In addition, clustering techniques are being employed aiming at a clustering structure where the attributes belonging to each cluster share relevant characteristics for the problem under study. In light of these results, our proposal to feed a training set with natural enzyme substrates is an innovative proposal and stands as a new strategy for finding new drug prototypes.

Keywords: HIV infections, viral infections, data science, computational biology, physicochemical properties, virtual drug screening.