Non-Homology-Based Prediction of Protein Target Regions by Logistic Regression

Gustavo Santos de Oliveira[1], Marcos Augusto dos Santos[1], Vasco Ariston de Carvalho Azevedo[1]

[1]*Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte – MG, Brazil.*

Influenza A is a RNA virus responsible for multiple types of health problems in different hosts, such as chickens, pigs and humans. The two proteins involved in its pathogenicity are hemagglutinin and neuraminidase. These are proteins, which triggers the virus' entrance inside the cell. Treatment options has been a challenge due the high evolutionary rate of the virus proteins. The prediction of potential target sites of key virus proteins has been a difficult task due lack of homology between sequences. In this sense, here it is proposed a new methodology to identify conserved residues associated with viruses' proteins by means of logistic regression. Due to its known activity related to the virus internalization, hemagglutinin protein sequences of Influenza A H5N1 and H3N2 were selected as a positive highly pathogenic and less pathogenic model, respectively. From UNIPROT, 5466 sequences for H5N1 and 259 for H2N3 were obtained. The model was built using MATLAB®, in an aproach where a sliding window was designed to count all possible triplet residues for each sequence. The matrix of all possible triplets for all sequences were submitted to an initial Singular Value Decomposition (SVD) for sampling homogeneization, and logistic regression, aiming to find triplets associated with the enhanced pathogenicity effect for H5N1 hemaglutinin compared to H2N3. The present aproach was able to detect critical regions associated with the high pathogenic hemaglutinin activity. That is the case of RRKKR, which is located within a connector loop between HA1 and HA2 subunits and is known as a target for proteolytic cleavage responsible for hemagglutinin activation. Hemagglutinin allows virus internalization thanks to its ability to bind the host membrane and its capacity to bend itself in low pH, permitting the virus to get inside the cell. The methodology detected, also, the triplet SII, located within the alpha-helix of HA2 subunit, in a region that acts as a hinge that tights the interaction between the two subunits after conformational change due pH decrease. Other sections detected were SNEQG, a terminal HA2 region, responsible for the formation of host membrane pores, and KIA, located in a beta-sheet region in HA1 and associated with the protein stabilization in acidic environments. In conclusion, this methodology was able to predict key regions for the hemagglutinin mechanisms of action and to corroborate with some authors that highlight the entrance mechanisms as more significant than the host cells recognition mechanisms by hemagglutinin, for the virus pathogenicity.