A predictive alignment-free model based on a new logistic regression-based method for feature selection in complete and partial sequences of Senecavirus A

Tatiana Flávia Pinheiro de Oliveira[1], Marcos Augusto dos Santos[2], Marcelo Fernandes Camargos[3], Antônio Augusto Fonseca Júnior[3], Aristóteles Góes-Neto[4], Edel Figueiredo Barbosa Stancioli[4],

1 Departamento de Microbiologia, Instituto de Ciências Biológicas, UFMG, Ministério da Agricultura, Pecuária e Abastecimento
2 Departamento da Ciência da Computação, Instituto de Ciências Exatas, UFMG
3 Ministério da Agricultura, Pecuária e Abastecimento
4 Departamento de Microbiologia, Instituto de Ciências Biológicas, UFMG

Abstract

In 2015, there was an outbreak involving pig farms in six Brazilian states, whose single agent found and described for the first time in the country was the Senecavirus A (SVA), a virus belonging to the genus Senecavirus (Picornavirales, Picornaviridae). This viral family also houses the genus Aphtovirus, whose species type is the Foot-and-mouth disease virus (FMDV), agent of Foot-and-mouth disease, a highly infectious disease notifiable under the strict control of the Ministry of Agriculture, Livestock and Supply (MAPA) and the World Organisation for Animal Health (OIE). In the past few years, there has been a growing interest in the application of methods of linear algebra and statistics in data mining, social networks, machine learning, bioinformatics and information retrieval. Among these methods, logistic regression approach draws some special interest as it is a standard method for data classification using genome data and is the most frequently used method for disease prediction. We introduce a model that represents sequences as 6-nucletotide frequency vectors in R4096 and 3- amino acids frequency vectors in R800 and uses information of SVA and FMDV from the complete genome or amino acid sequences of the polyprotein of theses viruses. In addition, partial sequences of nucleotides / amino acids of structural proteins (VP1, VP2, VP3 and VP4) of these viruses were used to build a new logistic regression-based method for classification. This new model allowed the assignment of values to parameters ai* that are associated with the frequency of a certain hexanucleotide or triplets codons of amino acids. Scrutinizing these parameters ai* unveiled that the most positive value may be related to important target sites of key virus proteins. Thus, this methodology was able to predict key regions in Senecavirus A, which can be important in studies of viral replication mechanism or in the development of diagnostic kits.

Funding: LANAGRO-MG ; FAPEMIG; CNPq