Proceedings of the World Congress on New Technologies (NewTech 2015)
Barcelona, Spain – July 15 - 17, 2015
Paper No. 164

Use of Quadratic Discriminant Analysis in Viral Host Prediction

Wojciech Gałan
Department of Computational Biophysics and Bioinformatics, Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University
Gronostajowa 7, 30-387 Kraków, Poland
wojciech.galan@gmail.com

Maciej Bąk
Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University
Gronostajowa 7, 30-387 Kraków, Poland
bakmaciej@wp.pl

Extended Abstract

The advent of new sequencing techniques allows cheap and rapid sequencing of environmental samples and the availability of the sequence data poses new challenges for scientists. For example, we do not know anything about the potential host or hosts of the newly discovered virus isolated from ocean water or sewage. First approach to infer the phylum of a viral host was presented by Kapoor et al. (2010). The group, representing each viral sequence by its mono- and dinucleotide frequencies, successfully employed discriminant analysis to distinguish between the Picorna-like viruses infecting mammals, plants or insects.

Here we report a similar analysis based on quadratic discriminant analysis (QDA) method. In our study we aimed to obtain a classifier, which would be able to predict whether a given virus infects eukaryotes or *Bacteria*/*Archaea*. This prediction should be accomplished based only on the viral sequence, without any additional information. In our research we have used 560 complete viral reference genomes (280 infecting *Eukaryota* and 280 other viruses) each of which was represented by a vector of biochemical information about the sequences (the number of strands and nucleic acid type) as well as some factors computed using sequence itself (strand length, mono- and dinucleotide frequencies, relative di- and trinucleotide frequencies). To achieve the best predictive abilities of our classifier, subsets of the sequence features were selected using genetic algorithms. Classifiers were evaluated using 10-fold stratified cross-validation. Applying this approach, we obtained a model which was able to properly ascribe ~90% viruses to one of the two groups: the viruses infecting *Eukaryota* and those infecting *Bacteria* or *Archaea*, achieving Matthew's correlation coefficient above 0.80. Our analysis was not restricted to any taxonomic group of viruses, but covers all viral sequences.

Biological samples (like ocean water or sewage), could potentially contain viruses infecting a wide range of hosts, from bacteriophages to humans. Similar problems may be encountered while studying feces or respiratory fluids, since viruses found in these samples might have been ingested or inhaled earlier. We believe that our discovery may facilitate an analysis of viral sequences found in such samples and give an approximate answer to the question: 'Could the virus be potentially harmful to a human?'

Research reported in this publication was supported by National Science Centre under grant number DEC-2012/07/N/NZ2/00108

Kapoor, A., & Simmonds, P. (2010). Use of Nucleotide Composition Analysis to Infer Hosts for Three Novel Picorna-Like Viruses. *J. Virol.*, 84(19), 10322-8.