Deep feature learning for virus detection using a Convolutional Neural Network

Calvo, Diego ; de la Torre, Isabel ; Franco, Manuel Angel; Brunak, Søren; Gonzalez-Izarzugaza, Jose Maria

Published in:
Book of Abstracts, Sustain 2017

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Calvo, D., de la Torre, I., Franco, M. A., Brunak, S., & Gonzalez-Izarzugaza, J. M. (2017). Deep feature learning for virus detection using a Convolutional Neural Network. In Book of Abstracts, Sustain 2017 [H-6] Technical University of Denmark (DTU).

Deep feature learning for virus detection using a Convolutional Neural Network

Diego Calvo[1, 2], Isabel de la Torre[2], Manuel Angel Franco [2], Søren Brunak[1], José M.G. Izarzugaza*[1]

1: Department of Bioinformatics, Technical University of Denmark, Kgs. Lyngby, Denmark.

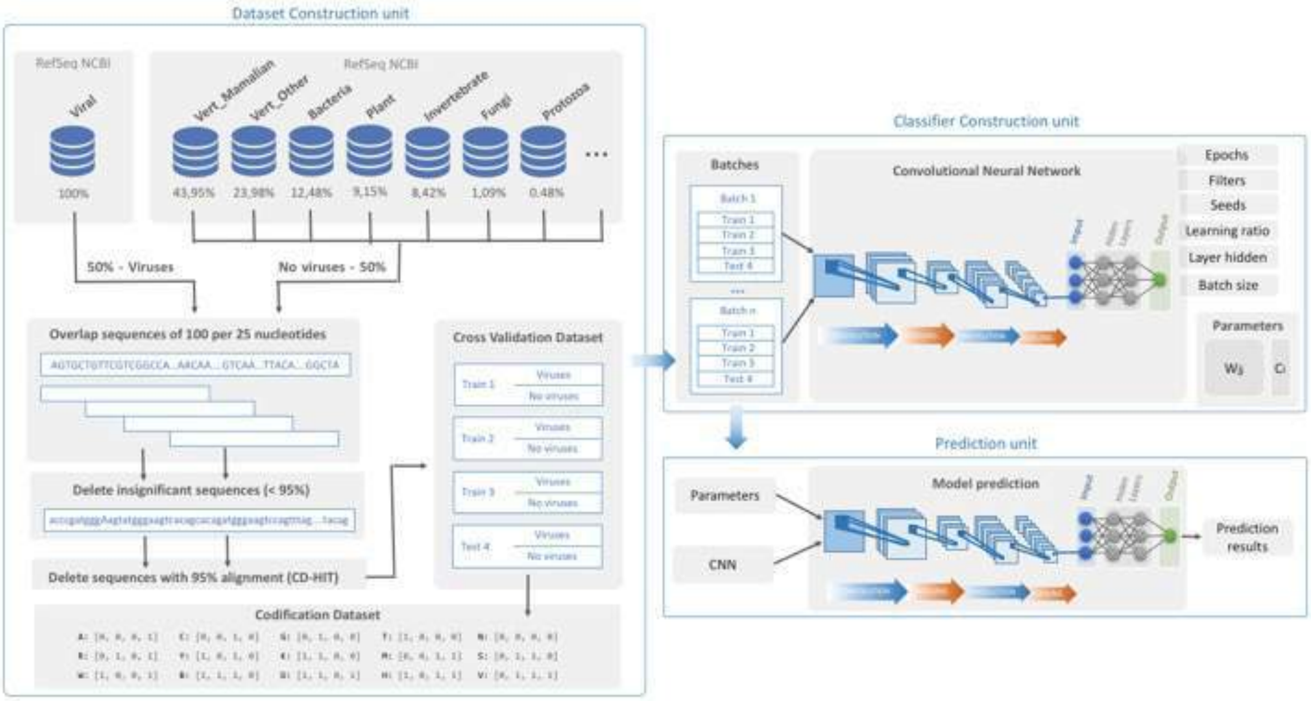
2: Department of Signal Theory and Communications, University of Valladolid, Valladolid, Spain.

*Corresponding author e-mail: txema@bioinformatics.dtu.dk

This study is focused on the development of a technology to identify characteristics in nucleotide sequences using deep learning provided by Convolutional Neural Networks. In order to demonstrate the effectiveness of this technology, a classifier has been developed to identify viruses in sequencing reads of 100 nucleotides, a proxy for a real NGS scenario. This classifier is able to search for known virus characteristics and identify potential new viruses that are currently undetected. As it is not necessary to read the complete sequences to recognize a virus, we manage to reduce the time and costs of virus identification.

The used Convolutional Neural Network to develop the classifier has been trained with RefSeq data. The training set was made up of two subsets. The first subset (positive set) includes all the nucleotides sequences of found viruses in the database and the second subset (negative set) is composed by a random selection of all the nucleotide sequences of non-viruses respecting the existing proportion of each found specie.

This training group undergoes is partitioning, overlapping and data cleaning transformations and it has resulted in a training set of 39.807.052 elements of approximately 2.2Gb of storage.


Dataset Construction unit
RefSeq NCBI
Viral
100%
RefSeq NCBI
Vert_Mamalian
Vert_Other
Bacteria
Plant
Invertebrate
Fungi
Protozoa
43,95%
23,98%
12,48%
9,15%
8,42%
1,09%
0.48%
...
50% - Viruses
No viruses - 50%
Overlap sequences of 100 per 25 nucleotides
AGTGCTGTTCGTCGGCCA...AACAA...GTCAA...TTACA...GGCTA
Delete insignificant sequences (< 95%)
Delete sequences with 95% alignment (CD-HIT)
Cross Validation Dataset
Train 1
Viruses
No viruses
Train 2
Viruses
No viruses
Train 3
Viruses
No viruses
Test 4
Viruses
No viruses
Codification Dataset
Classifier Construction unit
Batches
Batch 1
Train 1
Train 2
Train 3
Test 4
...
Batch n
Train 1
Train 2
Train 3
Test 4
Convolutional Neural Network
Epochs
Filters
Seeds
Learning ratio
Layer hidden
Batch size
Parameters
W3
Ci
Prediction unit
Parameters
CNN
Model prediction
Prediction results