World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:15, No:01, 2021

Towards End-To-End Disease Prediction from Raw Metagenomic Data

Authors : Maxence Queyrel, Edi Prifti, Alexandre Templier, Jean-Daniel Zucker

Abstract : Analysis of the human microbiome using metagenomic sequencing data has demonstrated high ability in discriminating various human diseases. Raw metagenomic sequencing data require multiple complex and computationally heavy bioinformatics steps prior to data analysis. Such data contain millions of short sequences read from the fragmented DNA sequences and stored as fastq files. Conventional processing pipelines consist in multiple steps including quality control, filtering, alignment of sequences against genomic catalogs (genes, species, taxonomic levels, functional pathways, etc.). These pipelines are complex to use, time consuming and rely on a large number of parameters that often provide variability and impact the estimation of the microbiome elements. Training Deep Neural Networks directly from raw sequencing data is a promising approach to bypass some of the challenges associated with mainstream bioinformatics pipelines. Most of these methods use the concept of word and sentence embeddings that create a meaningful and numerical representation of DNA sequences, while extracting features and reducing the dimensionality of the data. In this paper we present an end-to-end approach that classifies patients into disease groups directly from raw metagenomic reads: metagenome2vec. This approach is composed of four steps (i) generating a vocabulary of k-mers and learning their numerical embeddings; (ii) learning DNA sequence (read) embeddings; (iii) identifying the genome from which the sequence is most likely to come and (iv) training a multiple instance learning classifier which predicts the phenotype based on the vector representation of the raw data. An attention mechanism is applied in the network so that the model can be interpreted, assigning a weight to the influence of the prediction for each genome. Using two public real-life data-sets as well a simulated one, we demonstrated that this original approach reaches high performance, comparable with the state-of-the-art methods applied directly on processed data though mainstream bioinformatics workflows. These results are encouraging for this proof of concept work. We believe that with further dedication, the DNN models have the potential to surpass mainstream bioinformatics workflows in disease classification tasks.

Keywords : deep learning, disease prediction, end-to-end machine learning, metagenomics, multiple instance learning, precision medicine
Conference Title : ICMLB 2021 : International Conference on Machine Learning and Bioinformatics
Conference Location : Bangkok, Thailand
Conference Dates : January 18-19, 2021