AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING
Computer and Systems Engineering

Metagenomic data analysis using deep learning

A Thesis submitted in partial fulfillment of the requirements of
Master of Science in Electrical Engineering
(Computer and Systems Engineering)

by

Aly O. Abdelkareem
Bachelor of Science in Electrical Engineering
(Computer and Systems Engineering)
Faculty of Engineering, Ain Shams University, 2016

Supervised By

Prof. Hazem M. Abbas

Dr. Mahmoud I. Khalil

Cairo, 2018

AIN SHAMS UNIVERSITY

FACULTY OF ENGINEERING

Computer and Systems Engineering

Metagenomic data analysis using deep learning

by

Aly O. Abdelkareem

Bachelor of Science in Electrical Engineering

(Computer and Systems Engineering)

Faculty of Engineering, Ain Shams University, 2016

Examiners' Committee

Name and affiliation **Signature**

Prof. Ayman M. Eldeib
Systems and Biomedical Engineering ....
Faculty of Engineering, Cairo University.

Prof. Hoda Korashy Mohamed
Computer and Systems Engineering ....
Faculty of Engineering, Ain Shams University.

Prof. Hazem M. Abbas
Computer and Systems Engineering ....
Faculty of Engineering, Ain Shams University.

Dr. Mahmoud I. Khalil
Computer and Systems Engineering ....
Faculty of Engineering, Ain Shams University.

Date: / / 2019

Statement

This thesis is submitted as a partial fulfillment of Master of Science in Electrical Engineering, Faculty of Engineering, Ain Shams University. The author carried out the work included in this thesis, and no part of it has been submitted for a degree or a qualification at any other scientific entity.

Aly O. Abdelkareem

Signature

. . . .

Date: 25 Dec 2018

Researcher Data

Name: Aly Osama Aly Ibrahim Abdelkareem (Aly O. Abdelkareem)

Date of Birth: 29/08/1992

Place of Birth: Cairo, Egypt

Last academic degree: Bachelor of Science in Electrical Engineering

Field of specialization: Computer and Systems Engineering (Credit Hours)

University issued the degree : Ain Shams University

Date of issued degree : 20/7/2016

Current job : Teaching Assistant at Faculty of Engineering, Ain Shams University

Abstract

Faculty of Engineering – Ain Shams University
Computer and Systems Engineering Department

Thesis title: **”Metagenomic data analysis using deep learning”**
Submitted by: **Aly O. Abdelkareem**
Degree: **Master of Science**

Abstract

Metagenomics holds a great promise for a better understanding of the function and diversity of viral and microbial communities because only a minority of viruses and microorganisms can be isolated in pure culture. Viral sequence identification is considered one of the essential steps in analyzing metagenomic data. Although various methods use homology and statistical methods to identify viral sequences, these methods encounter many limitations because of the limited genomic databases and the high viral genome diversity. In this thesis, an attention deep neural network model was used for identifying viral reads among metagenomic data. This method is used to purify mixed metagenomic data from viral contamination. The proposed neural network model is able to outperform state-of-the-art tools of viral identification from high throughput sequences on the same testing data. According to these results, our model would help to understand viruses in various microbial communities and discovering new viruses.

Thesis Summary

Summary

Several studies show different methods to identify viruses in mixed metagenomic data and phages in host genomes, using homology and statistical techniques. These techniques have many limitations because of the limited genomic databases and viral genome diversity.

In this work, a text processing approach for text featurization is presented by treating DNA similar to natural language. It reveals the importance of using the text feature extraction pipeline to transform DNA base pairs into a set of characters with a term frequency and inverse document frequency featurization technique. Various machine learning classification algorithms are applied to viral identification tasks such as logistic regression and multi-layer perceptron.

For testing purpose, fragments of viruses and bacteria were generated from RefSeq genomes with different lengths to find the best hyperparameters of deep neural network model. Then, microbiome and virome high throughput data were simulated from our test genome dataset with the aim of validating our approach.

This thesis introduces a deep neural network model for analyzing the metagenomic data. Furthermore, this model was able to identify viral reads in mixed metagenomic data. The proposed neural network model was compared to a recent state-of-the-art statistical tool for viral reads identification and achieved better results regarding accuracy and speed on the same testing data.

Thesis is divided into six chapters as listed below, along with a list of figures, list of tables, and a bibliography.

Chapter 1 is a brief overview about microorganisms and metagenomic data. It also shows some of the challenges facing viral sequence identification from metagenomic data. Finally, the chapter explains the motivation beyond applying machine learning and deep learning algorithms with a demonstration of the main thesis contribution.

Chapter 2 is a survey of the related solutions proposed recently for metagenomic analysis and viral sequence identification. First, the chapter presents a review of various sequence alignment tools used in metagenomic reads annotation and specific tools for viral reads identification. Then, it reviews the tools based on k-mer features and statistical models. Finally, it emphasizes the differences between the currently used tools and the one proposed in this thesis.

Chapter 3 is an explanation to how we formulate DNA sequences as a text classification problem where the natural language processing techniques can be used, such as engineered feature extraction and machine learning algorithms for the viral read identification. Then, a short explanation for supervised learning algorithms are used in the comparisons, followed by different model evaluation techniques. Finally, feature importance was explained and applied to our data in order to find a list of contigs contributed to the viral classification logistic regression model.

Chapter 4 describes the proposed method. First, it explains the feed-forward neural networks and an experiment showing the experimental work with simple multilayer perceptron architecture with the engineered feature extraction. Then, a deep learning approach is proposed using the attention mechanism with a recurrent neural network. Finally, several experiments were carried out on a small sample of the testing data in order to find the best parameters for the proposed deep learning model.

Chapter 5 presents a set of experiments with the proposed method to measure accuracy and the speed of our solution comparing it with recent tools. First, a detailed description of the data generation pipeline for training and testing dataset is explained. Then, virome and microbiome simulation and a description for the case study data are presented. Finally, the results of these data are displayed and discussed.

Chapter 6 shows a brief conclusion of our work and multiple proposals for the possible future work extending this work.

Keywords: attention mechanism, bioinformatics, classification, deep learning, virus, metagenomics, microbiomes, next generation sequencing, recurrent neural networks, viromes

Acknowledgment

Aly O. Abdelkareem
Computer and Systems Engineering
Faculty of Engineering
Ain Shams University
Cairo, Egypt
Dec 2018

It is my pleasure to express my deepest gratitude to my supervisors, Prof. Hazem Abbas, and Dr. Mahmoud Khalil for their continued guidance and effort. Besides my supervisors, I would like to thank Dr. Ali Elbehery for his insightful comments and encouragement.

Last but not the least, I would like to express my sincere gratitude to my mother and my father for the continuous support of my study, my wife Mayada for her patience and motivation, and my daughter Aisha as she brought happiness to my life.

Contents