Research Project on Sequence Classification via Graph

Project Proposal

CSC 6850 Machine Learning

SPRING 2024

Alenka Tang, Omar Madjitov

❖ Introduction

➢ **Project Type:** research project in sequence classification using graph models

➢ **Background and Motivation**

Classification is not a new topic in bioinformatics. In the analysis of gene expression data, genes obtained from microarray data are clustered and genes in the same cluster are considered to trigger the same function. Sequence clustering and classification is an important task in bioinformatics to produce stable, quick, and accurate results.

➢ **Problem Definition**

This project proposed a method using graphical models for sequence classification. We will take existing models such as de-Bruijn graph[1], OverlapGraphs[5], Node2vec, and Graph2vec models with Support Vector Machine (SVM) and Neural Network (NN) classifiers and test them with different parameters values to find the combination with optimal results for sequence classification.

❖ Methodology

➢ **Datasets**

This project will use the Human DNA (HD) sequence dataset and the Cov–S-Protein-Seq (CPS) dataset which is publicly available. The HD sequence dataset consists of 4380 DNA sequences with each DNA sequence corresponding to a specific gene family which can be used as labels for training and testing the model, with a total of seven labels. The CPS dataset consists of 1238 spike protein sequences from 67 different coronavirus species (CoV). The dataset provides information on both the coronavirus species (CVS) and their host species (CHS). The Cov species are grouped into seven categories and the host species are grouped into six categories which we can use as two separate sets of labels for prediction using the spike protein sequences.

➢ Graph Construction

This project will use the de-Bruijn graph model, as well as Overlap Graphs to convert the sequence data to graphs with each sequence data entry being one graph. For the construction of the de-Bruijn graph for each data entry, we will try different k-mer values and break the sequence into overlapping k-mers. The graph will be constructed by creating nodes for each k-mer and connecting overlapping k-mers with direct edges. Lastly, we will simplify the graph by merging nodes with common overlaps with each sequence data in the dataset being an individual graph.

To construct the Overlap Graph we are going to use Overlap-Layout-Consensus [6] (OLC) algorithm, which consists of the following steps. We are going to identify all possible overlaps among the reads using the Suffix Trees Algorithm, where the suffix of one read matches the prefix of another read. Each read becomes a node in the graph. We will add edges between nodes to represent the overlaps between reads. The direction of the edge is going to indicate the direction of the overlap from the suffix of one read to the prefix of another. In the “Consensus” step we aim to infer the most likely sequence that could have produced the reads.

➢ Classification

After constructing the graph models for the dataset, we will apply machine learning models such as Node2vec and Graph2vec where the Node2vec model will convert each graph into a matrix with each node being a vector and the graph2vec will convert each graph into a singular vector. We will then utilize SVM and NN for classification. For the SVM classifier, we will test different kernels and different C and gamma values for the parameters. For using NN for classification, we will be testing using different NN such as convolution NN, feed-forward NN as well as different activation functions.

➢ **Analysis**

At the end of the project, we will compare the results for different combinations with graph construction and classifiers and determine if there is an optimal model for sequence classification via graph models.

❖ Timeline

- ➢ **Phase 1** - Literature review on de-Bruijn graph model, Node2vec and Graph2vec embeddings as well as SVM and NN classifiers.
- ➢ **Phase 2** - Construction of the de-Bruijn graph with different kmer values for both datasets and construction of the Overlap Graph using the Overlap-Layout-Consensus algorithm
- ➢ **Phase 3** - Implementation of Node2vec and Graph2vec model with SVM, NN for both datasets and testing different parameters.
- ➢ **Phase 4** (4/7 - 14) - Trying the combination of de-Bruijn graph construction and models with classifiers for final comparison.

❖ Supplement Materials

[1] Compeau, P., Pevzner, P. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* 29, 987–991 (2011). https://doi.org/10.1038/nbt.2023

[2] Nagesh Singh Chauhan. [n. d.]. Demystify DNA Sequencing with Machine Learning|https://www.kaggle.com/code/nageshsingh/demystify-dna-sequencing-with-machine-learning/notebook

[3] Ali, Sarwan, Babatunde Bello, Prakash Chourasia, Ria Thazhe Punathil, Yijing Zhou, and Murray Patterson. 2022. "PWM2Vec: An Efficient Embedding Approach for Viral Host Specification from Coronavirus Spike Sequences" *Biology* 11, no. 3: 418. https://doi.org/10.3390/biology11030418

[4] Kiril Kuzmin, Ayotomiwa Ezekiel Adeniyi, Arthur Kevin DaSouza, Deuk Lim, Huyen Nguyen, Nuria Ramirez Molina, Lanqiao Xiong, Irene T. Weber, and Robert W. Harrison. 2020. Machine learning methods accurately predict the host specificity of coronaviruses based on spike sequences alone. Biochemical and Biophysical Research Communications 533 (12 2020), 553–558. Issue 3.https://doi.org/10.1016/j.bbrc.2020.09.010

[5] Rizzi, R., Beretta, S., Patterson, M. et al. Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era. Quant Biol 7, 278–292 (2019). https://doi.org/10.1007/s40484-019-0181-x

[6] Zhenyu Li, Yanxiang Chen, Desheng Mu, Jianying Yuan, Yujian Shi, Hao Zhang, Jun Gan, Nan Li, Xuesong Hu, Binghang Liu, Bicheng Yang, Wei Fan, Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph, *Briefings in Functional Genomics*, Volume 11, Issue 1, January 2012, Pages 25–37, https://doi.org/10.1093/bfgp/elr035