Applying Machine Learning to Classify Gene Types in Bacteria using DNA Sequences

INSTITUTE of TECHNOLOGY CARLOW
Institiúid Teicneolaíochta Cheatharlach

Anjuman Ara Kali
MSc in Data Science
Supervisor: Paul Barry

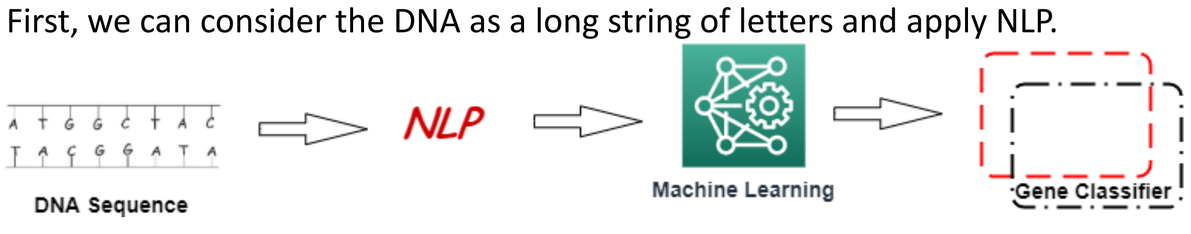
Introduction

Genes are identified and classified using bio-informatics tools (e.g., BLAST) where alignment algorithms (e.g., Needleman-Wunsch algorithm) are used along with a huge database of already classified DNA sequences. Afterwards, some wet lab experiments are done by biologists to validate the classification. As it is a huge task, the question rises on how we can build an alternative classification model based on machine learning by being agnostic to those alignment algorithms and bio-informatics knowledge, at the same time, not use the whole corpus of DNA sequence. There will be many times where we may simply need to identify the genes when the scope is very limited. For example, in a scenario where it is identified that the genes are surely responsive to toxic environment or the genes show some resistance to survive in the toxic environment. Then can it be identified whether the genes are biocide or heavy metal responsive using the existing knowledge base. It is redundant to go through the awful lot of data to make a search in this case which will be very irrelevant anyway. In short, in the case of having a small dataset, we simply want to analyse how machine learning can be helpful to do some faster classifications with limited scope. Furthermore, DNA structure is responsible for its behaviour. Understanding the probable protein structure from the DNA and estimating its behaviour, we want to know whether we can be agnostic to that kind of analysis and let the machine learning figure it out.

Literature review

To build a machine learning classifier, we need to extract features from the DNA sequence. Based on our literature review, we target to explore following three ways to deal with feature extraction.

First, we can consider the DNA as a long string of letters and apply NLP.



Second, Converting DNA as a signal and apply Signal processing methods (FFT, DWT etc.) for feature extraction.

Finally, if CNN based deep learning is considered then it can be left for the CNN (1-D Convolution) to do the feature extractions indirectly for us.

Research Question

- Can Machine learning help us in identifying gene types when limited dataset is available while traditional alignment based algorithm will be avoided and avoiding irrelevant search is preferred?
- What are the classification details that can be achieved through machine learning algorithm?

Technologies

Python, TensorFlow, Keras, scikit learn

Data Collection

Data Sources:

- BacMet: Antibacterial biocide and metal resistance genes database
- NCBI: National Center for Biotechnology Information

Data Collection Technique:

- Web scrapping using python script
- NCBI API for data collection

The collected data will be labelled according to their responsiveness (toxic metal or biocide) described in BacMet.

Methodology

- Using the extracted features common supervised ML algorithms e.g., Support Vector Machine, Naive Bayes, Random Forest, K-Nearest Neighbours and Deep learning etc. will be applied to create and test a model to classify the types of the genes.
- CNN based deep learning will be applied to leave the feature extraction up to the model.
- Feature selection techniques will be applied to eliminate unnecessary features and reduce the space to represent the genes.
- Metrics such as accuracy, precision, recall metrics etc. will be measured to compare the outcome of these various ML algorithms.

Early Development and Next Step

- The data collection from BacMet source has been started.
- Next, feature selection and ML model creation will start.
- We will start with the classification of genes in responsive to biocide and heavy metal (two broad categories). Then we can gradually try do go little deeper by classifying genes in responsive to heavy metals in details e.g., Nickle, Iron, Aluminium, Arsenic etc. as labels.
- Finally, we have a stretch goal, which is to use a bio informatics algorithm and apply it to our tiny dataset (tiny compared to the enormous corpus used by the bio informatics tools) for classifying the genes and then compare the result with our machine learning techniques. Hence our technique can be justified that when the data is very limited, machine learning technique can be used to get better results.

REFERENCES

[1]] National Center for Biotechnology Information (NCBI). 2019. Basic Local Alignment Search Tool (BLAST). https://blast.ncbi.nlm.nih.gov/Blast.cgi

[2] C. Pal, J. Bengtsson-Palme, C. Rensing, E. Kristiansson, and D. G. J. Larsson, “BacMet: Antibacterial biocide and metal resistance genes database,” *Nucleic Acids Research*, vol. 42, no. D1. Jan. 01, 2014.

[3] Z. Lv, H. Ding, L. Wang, and Q. Zou, “A Convolutional Neural Network Using Dinucleotide One-hot Encoder for identifying DNA N6-Methyladenine Sites in the Rice Genome,” *Neurocomputing*, vol. 422, pp. 214–221, Jan. 2021.

[4] D. W. Liu *et al.*, “Automated detection of cancerous genomic sequences using genomic signal processing and machine learning,” *Futur. Gener. Comput. Syst.*, vol. 98, pp. 233–237, Sep. 2019.

[5] C. Caragea, A. Silvescu, and P. Mitra, “Protein sequence classification using feature hashing,” in *Proteome Science*, Jun. 2012.