Title: A DEEP LEARNING MODEL FOR DNA SEQUENCE PREDICTION OF PROKARYOTES USING AN ARTIFICIAL NEURAL NETWORK

Authors: Kenneth Flloyd D. Lemente Glynn Noel T. Pupa Aprille Shane A. Tiedra

ABSTRACT

Artificial Neural Network is broadly utilized in both academia and industry. It is a deep learning model with a capability to perceive a data's sequential characteristics and utilize patterns to predict the next likely scenario. Over the last few years, many types of research show how Neural Networks can be used for computational biology applications such as DNA sequence classification. A difficult problem that continuously remains in the wide field of biological community, is the correct gene prediction. A powerful predictive model can be a step towards developing more reliable gene prediction methods for DNA which would give a great advantage for the scientific community and researchers. Datasets released by the NCBI-NIH Gene Bank and Encyclopedia of DNA Elements (ENCODE) are acquired and used in this study as input data. The researchers used the encoding method of one hot vector in representing the sequences as input data to the algorithm where the non-coding variant DNA sequence will be encoded in a vector of integer, to transform our categorical labels into a vector of zeros and ones. The length of these vectors is equal to the numbers of the category that our model is expected to classify then the training follows. During the training phase, the embedded numbers will be mapped into substrings, also known as the regulatory motif, and are concatenated after.

Keywords: Deep Learning, DNA Sequence Classification, Prokaryotes, Regulatory Motifs, Machine Learning, Artificial Neural Network