Evaluation of Convolutionary Neural Networks Modeling of DNA Sequences using Ordinal versus one-hot Encoding Method

Allen Chieng Hoon Choong
Faculty of Cognitive Sciences and Human Development
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia
allen.choong@gmail.com

*Nung Kion Lee
Faculty of Cognitive Sciences and Human Development
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia
nklee@unimas.my

***Abstract*—Convolutionary neural network (CNN) is a popular choice for supervised DNA motif prediction due to its excellent performances. To employ CNN, the input DNA sequences are required to be encoded as numerical values and represented as either vectors or multi-dimensional matrices. This paper evaluates a simple and more compact ordinal encoding method versus the popular one-hot encoding for DNA sequences. We compare the performances of both encoding methods using three sets of datasets enriched with DNA motifs. We found that the ordinal encoding performs comparable to the one-hot method but with significant reduction in training time. In addition, the one-hot encoding performances are rather consistent across various datasets but would require suitable CNN configuration to perform well. The ordinal encoding with matrix representation performs best in some of the evaluated datasets. This study implies that the performances of CNN for DNA motif discovery depends on the suitable design of the sequence encoding and representation. The good performances of the ordinal encoding method demonstrates that there are still rooms for improvement for the one-hot encoding method.**

***Index Terms*—DNA sequence encoding, convolutionary neural networks, motif discovery**

I. Introduction

CNN (Convolutional Neural Network) [1], [2] is currently one of the most widely used deep learning methods in machine learning due to its powerful modelling capability on complex and large-scale datasets. Recently, CNN has been widely used for learning DNA sequence datasets related to regulatory regions and other functional landmarks [3]–[6]. The advantage of CNN is its learning can be performed without the need of engineered features. The intrinsic features in the raw dataset are learned through the many layers structure which represents the different abstraction of features. The layers in a CNN consist of convolutionary and pooling layers. A convolutionary layer consists of multiple maps of neurons which are called filters. A filter convolves the inputs from the previous layer to produce a reduced sample. It only connected to a patch of the previous layer, which is named as ”receptive field”. Moreover, all neurons in the filters detect the same features of the previous layer but at different map locations. Different filters might detect different types of features [7]. In a DNA dataset, the features might represent different motifs enriched in the input DNA sequences. In addition, [7] stated that the exact locations and frequency of a feature are unimportant to the learning purpose because the final output of the deep learning is recognition of the input data. On the other hand, the pooling layer summarizes the adjacent neurons by computing their activity. As a result, the model parameters are greatly reduced. After the last pooling layer, it has a fully connected multi-layers perceptron neural networks.

CNN is designed to effectively models multi-dimensional input data. Thus, it is powerful in solving problems related to computer vision and image recognition [2] where the data consists of images. To employ the CNN on DNA datasets, existing works typically encode the nucleotides in DNA sequences by using the one-hot method [3], [5], [6]. That is, each nucleotide is encoded with a binary vector of four bits with one of them is hot (i.e. 1) while others are 0. For instance $A = (1, 0, 0, 0)$, $G = (0, 1, 0, 0)$, $C = (0, 0, 1, 0)$, and $T = (0, 0, 0, 1)$. This sequence encoding method draws similarity to the Position Frequency Matrix [8]. In which, the values in a vector are considered as the probability of finding the four bases at a certain position in a DNA sequence. Once encoded, an input DNA sequence of length l is represented as $4 \times l$ matrix. Or in another word, a ”2D image” with one channel.

Methods for converting biological sequences into numerical values have been existed in numerious past studies [9]. Those encoding methods can be categorized into direct and indirect encoding [9]. Direct methods represent each nucleotides/amino acids with a numerical value or vector of numerical values. They preserved the original order the bases appeared in a biological sequence after the encoding. While the indirect methods engineered a fixed number of features (numerical values) from the biological sequences. The features can be based on frequency counts of various k-mers (short sequence segments of length *k* bp), biological, or biochemical properties.

This study is supported by the Minister of Education Malaysia, Fundamental Research Grant Scheme-FRGS/SG03(01)/1134/2014(01)

978-1-5386-0765-7/17/$31.00 2017 IEEE, *Corresponding author