A Deep Learning Based Method for Predicting DNA N6-Methyladenine (6mA) Sites in *Eukaryotes*

L.H. Roland* and C.T. Wannige

Department of Computer Science, University of Ruhuna, Matara, Srilanka.

DNA N6-methyladenine (6mA) is an epigenetic modification, which is involved in many biological regulation processes like DNA replication, DNA repair, transcription, and gene expression regulation. The widespread presence of this 6mA modification in *eukaryotes* has been unclear until recently. Therefore, for *eukaryotes*, the study of DNA 6mA is insufficient. Accurate identification of 6mA sites genome-wide provides a deeper understanding of the epigenetic modification process and the biological processes it involves. Existing experimental techniques are time-consuming and computational machine learning methods have room for performance improvement. DNA N6-methyladenine prediction in cross-species shows low performance. Hence, there is a need for a highly accurate, time-efficient method to predict the distribution of 6mA sites in *eukaryotes*. Deep learning models have shown higher accuracy in many experiments in bioinformatics. In this regard, we develop a customized VGG16 based model using convolution neural networks. We introduce a novel 3-dimensional encoding mechanism extending the one-hot encoding method for the given DNA sequences of length 41bp to support the VGG16 model input. Specifically, the 10-fold cross-validation on the benchmark datasets for the proposed model achieves higher accuracies for cross-species, Rice, and *M. musculus* genomes. The cross-species data set was prepared by integrating the benchmark datasets of Rice, and *M. musculus*. This model outperforms the existing computational tools SNNRice6mA, ilM-CNN with a current validation accuracy of 97% for the prediction of 6mA sites. The model trained with cross-species data predicts 6mA sites of other species *Arabidopsis Thaliana*, *Rosa Chinensis*, *Drosophila*, and Yeast with a prediction accuracy over 70%. Thus, this model can be used for the genome-wide prediction of 6mA sites in *eukaryotes*.

Keywords: DNA Sequence encoding method, Deep learning, Epigenetics, Bioinformatics, DNA N6-Methyladenine