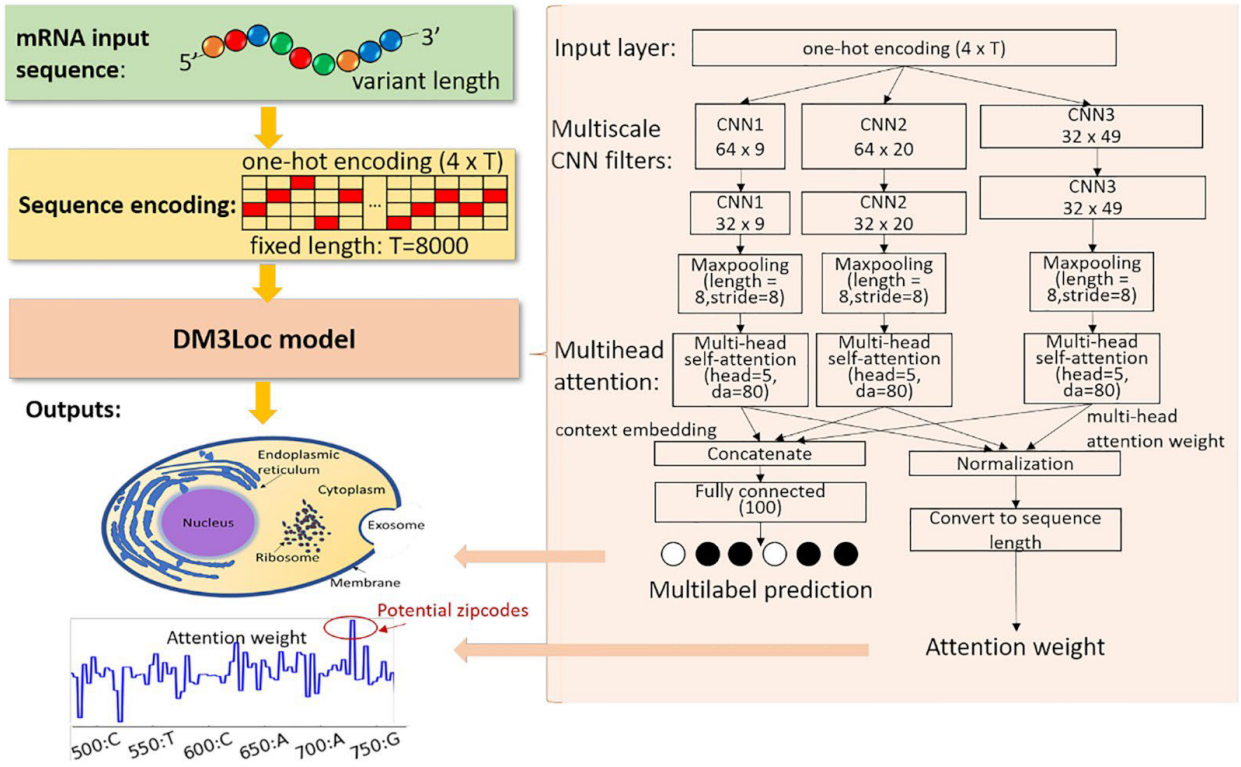
Duolin Wang, Zhaoyue Zhang, Yuexu Jiang, Ziting Mao, Dong Wang, Hao Lin, Dong Xu. DM3Loc: Multi-label mRNA Subcellular Localization Prediction and Analysis Based on Multi-head Self-attention Mechanism. Nucleic Acids Research. 49: e46, https://doi.org/10.1093/nar/gkab016, 2021.

Subcellular localization of messenger RNAs (mRNAs), as a prevalent mechanism, gives precise and efficient control for the translation process. There is mounting evidence for the important roles of this process in a variety of cellular events. Computational methods for mRNA subcellular localization prediction provide a useful approach for studying mRNA functions. However, few computational methods were designed for mRNA subcellular localization prediction and their performance have room for improvement. Especially, there is still no available tool to predict for mRNAs that have multiple localization annotations. In this paper, we propose a multi-head self-attention method, DM3Loc, for multi-label mRNA subcellular localization prediction. Evaluation results show that DM3Loc outperforms existing methods and tools in general. Furthermore, DM3Loc has the interpretation ability to analyze RNA-binding protein motifs and key signals on mRNAs for subcellular localization. Our analyses found hundreds of instances of mRNA isoform-specific subcellular localizations and many significantly enriched gene functions for mRNAs in different subcellular localizations.



DM3Loc framework. The input of the model is an mRNA sequence of any length; then, it is one-hot encoded into a fixed-length matrix. There are two outputs of the DM3Loc model; one is the prediction scores for six localization categories and the other is the attention weights, which have the same length as the input sequence. The regions with high attention weights indicate the zipcode regions. The DM3Loc model mainly consists of multiscale CNN filters and multi-head self-attention layers. The multiscale CNN is used to capture the localization signals at variant lengths, and the multi-head self-attention is used to make the model attend more to the important sequence regions for localization.