# Prediction of RNA localization with fast-training QRNNs


Alessandro Montemurro s171964
Léa Riera s201848
Niels Mølgaard Knudsen s153054



## Introduction

Biological functions of RNAs, including translation of genetic information, cellular signal transduction and transcriptional regulation are determined by their location in cell. As a high-throughput way of determining subcellular localization of RNA is not feasible yet, it remains attractive to use the RNA sequence itself for prediction of its subcellular localization.

Traditionally, LSTM networks are used to model sequential data. The main drawback of LSTMs is that each timestep's computation depends on the previous timestep's output. Convolutional networks are fast and parallelizable but they fail in modelling long-distance dependencies. Quasi-Recurrent Neural Networks (qRNNs) take advantages from both networks: they alternates convolutional layers, which apply in parallel across timesteps, and a minimalist recurrent pooling function that applies in parallel across channels, often speeding up training [1].

## Key points

- Due to the high length of the sequences, only the first and last 600 nucleotides are used for the classification. This choice relies on the assumption that the most informative part of the sequence is located at the beginning and at the end.
- We focused on a sub-problem where we considered only two classes: nucleus and ribosome.
- Up-sampling has been used to deal with classes imbalance
- Dynamic batching is applied to avoid pointless computation on the padded part of the sequences.

## Description of data set

- The data set is a subset of the RNALocate database [2].
- The data is divided in 3 sets, training (70%), validation (10%) and test (20%), Different sequence lengths are equally represented in each set.

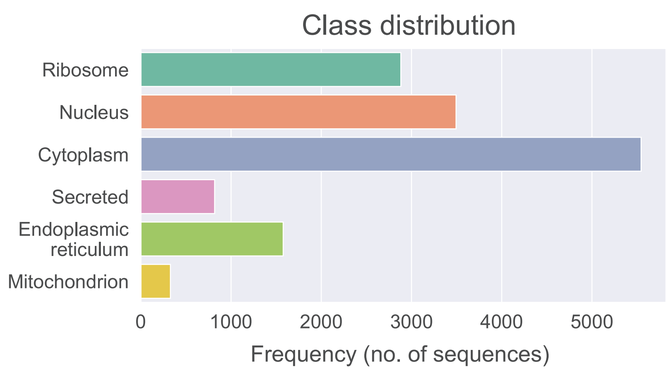


**Figure 1.** Bar plot showing the distribution of subcellular localizations for the samples in the data set.

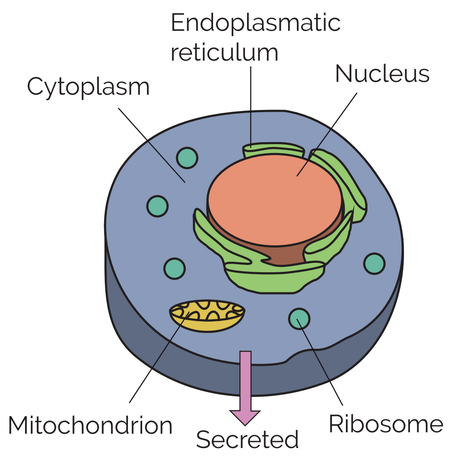


**Figure 3.** Scheme showing the placement of the tracked subcellular compartments in a typical eukaryotic cell.

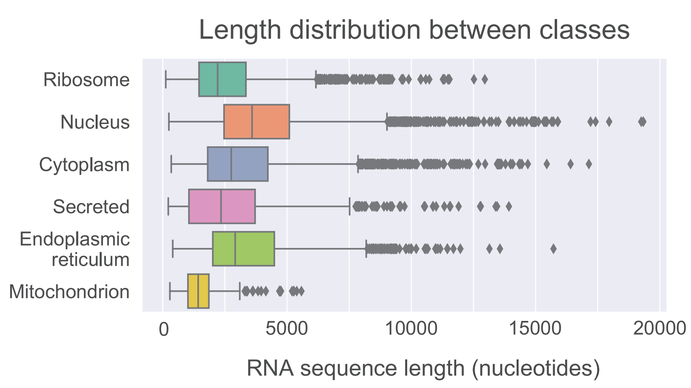


**Figure 2.** Box plots showing the distribution of sequence lengths for each individual subcellular localization (class).

## LSTM vs QRNN

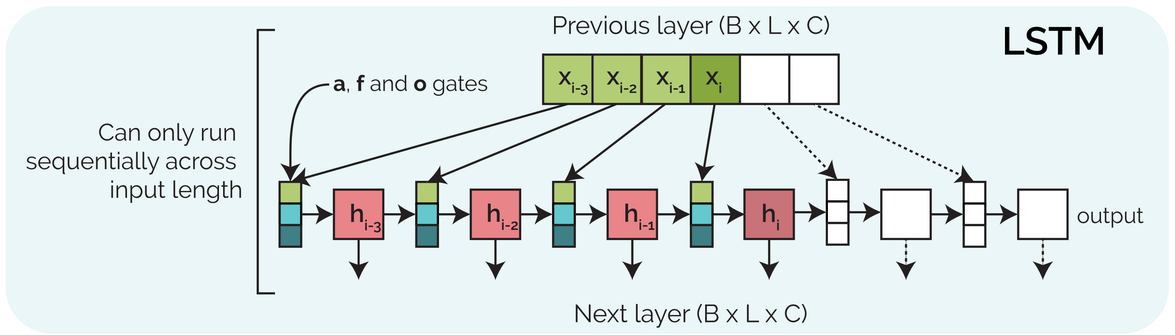


**Figure 4** First plot shows the distribution of sequences across classes, second plot shows the distribution of sequence lengths across classes.

## Network architectures

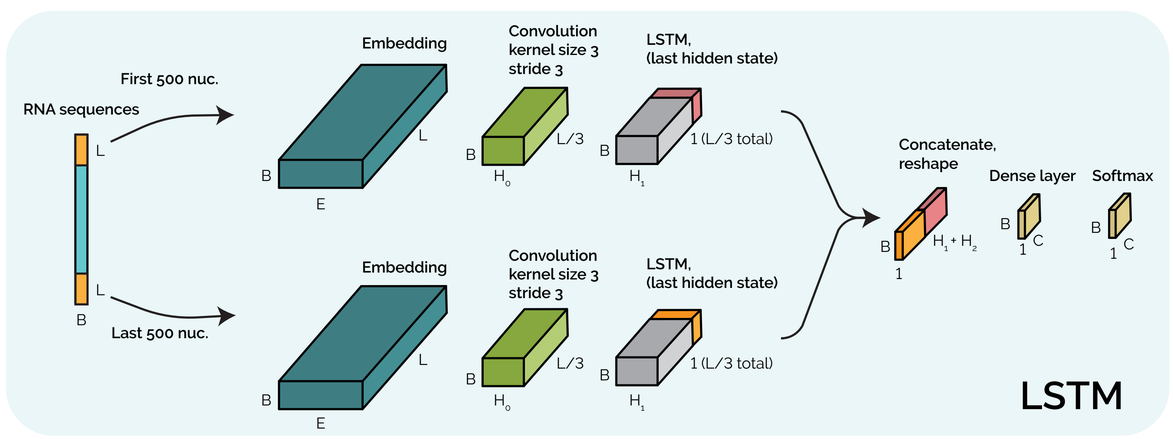


**Figure 5**. Network architectures overview.

## Results

- With a small batch size of 2, the QRNN trains at least 10x faster per epoch.
QRNN epoch: 127 seconds.
LSTM epoch: >20 minutes
- Similar accuracies are accomplished by the two networks.
Beat random guessing (53%).
- QRNN network had no problem handling the entire (length > 20k) sequences.

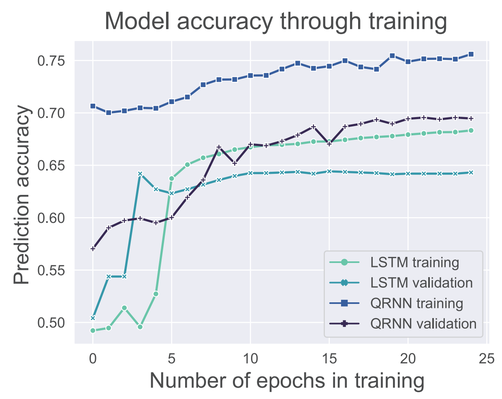


**Figure 6.** Graph showing model accuracy for each network during training.

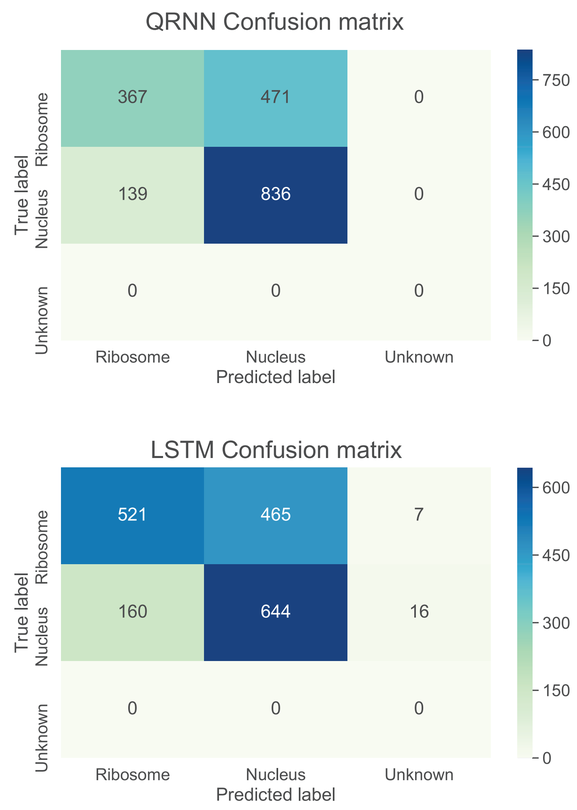


**Figure 7.** Confusion matrices for both networks traned on the two-class data set.


**References**

[1] James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher,"Quasi-recurrent neural networks", arXiv preprint arXiv:1611.01576, 2016.
[2] Zhang T, Tan P, Wang L, Jin N, Li Y, Zhang L, Yang H, Hu Z, Zhang L, Hu C, Li C. "RNALocate: a resource for RNA subcellular localizations", Nucleic acids research, 2016