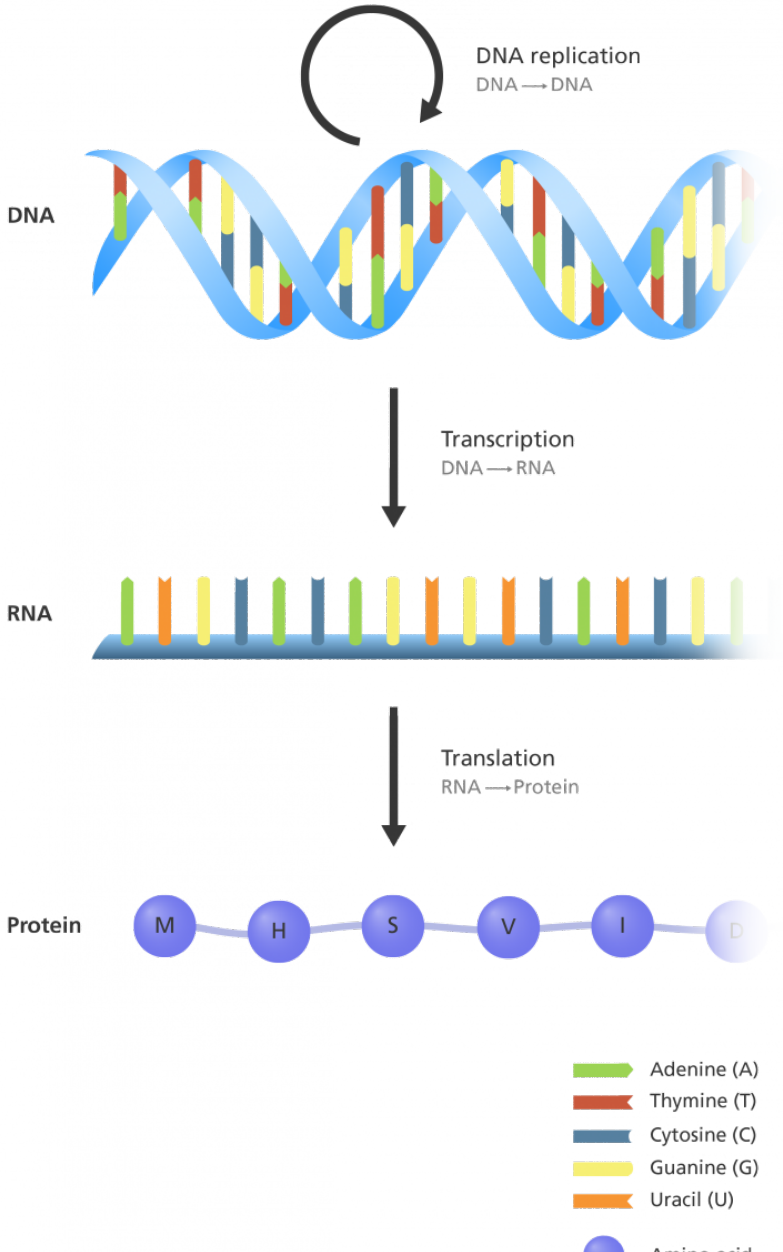
Marking protein coding boundaries on the genome using RNNs

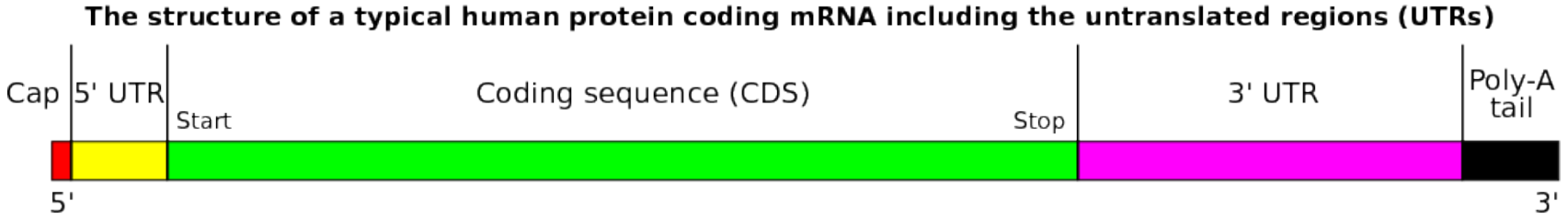
Saket Choudhary

November 6, 2017

Central Dogma of Biology

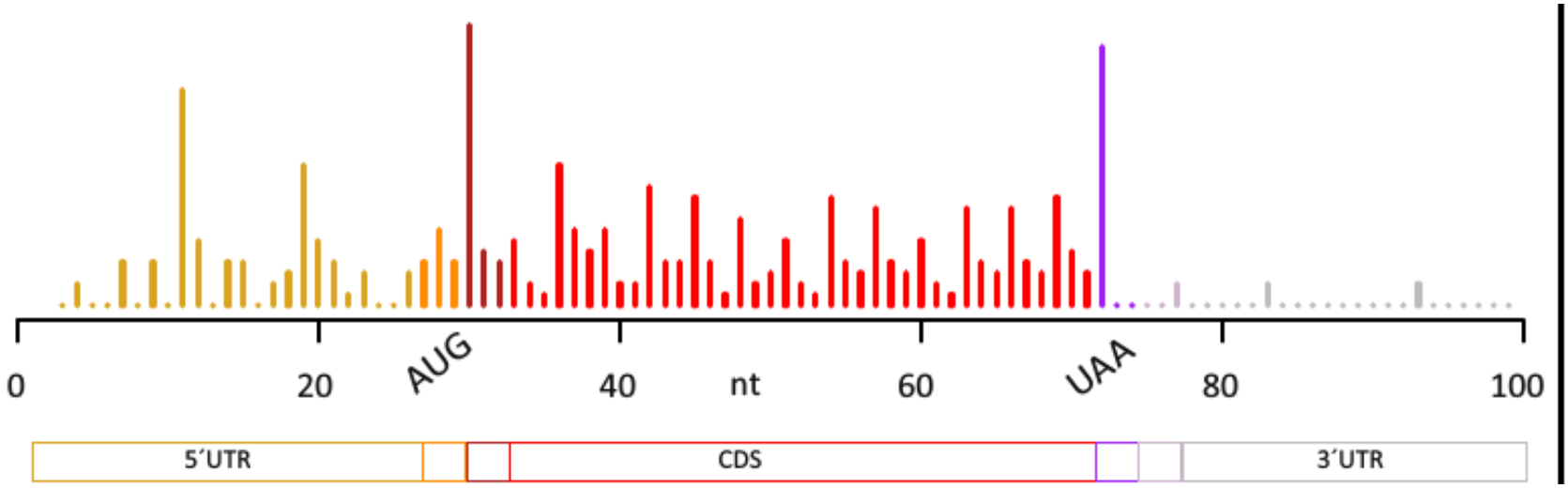


Gene clan be partitioned based on ‘coding’ potential



[?]

Gene clan be partitioned based on ‘coding’ potential



[?]

Motivation

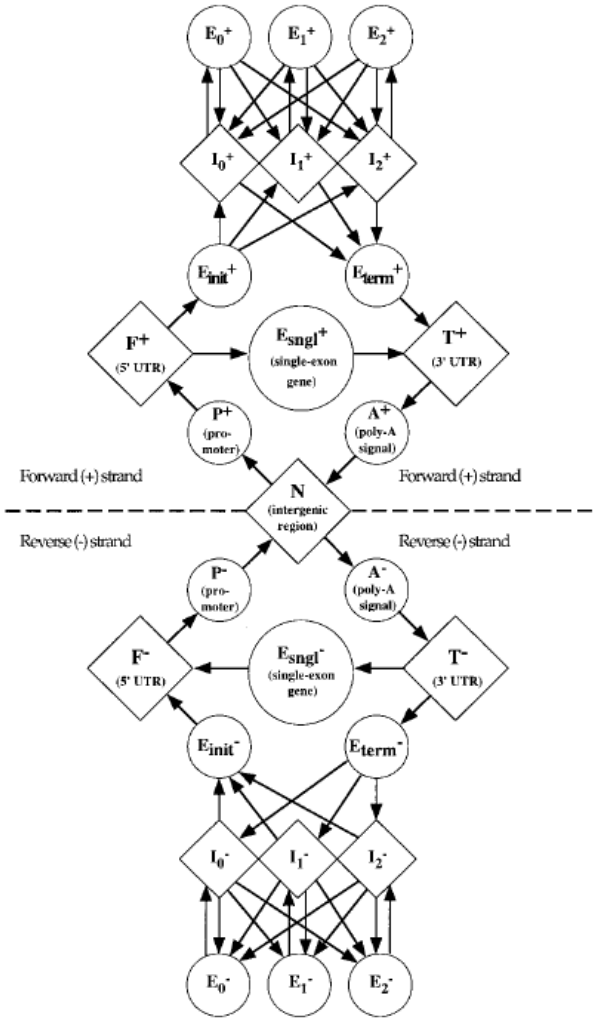
- 'Close to exact' boundaries are known for a few organisms
- Different partitions of the gene carry out different roles
- Experiments require lot of resources and time!
- Annotation is often required for any downstream application: for e.g. mutation analysis for personalized medicine

Problem Formulation/Goal

Input: A vector $\mathbf{x} \in \mathcal{V}^l$ where l is sequence length and $\mathcal{V} = \{N, A, C, T, G\}$.

Output: Labels $\mathbf{y} = \{5'\text{UTR}, \text{CDS}, 3'\text{UTR}\}^l$

Related Work: HMM based



[?]

Data Availability

National Center of Biotechnology Information - Sequence Read Archive(NCBI SRA) hosts genomic data from multiple organisms hosted publicly.

`https://www.ncbi.nlm.nih.gov/sra`

- Each organism has multiple genes. For example humans 25000 genes with close to gold standard annotation
- Around 7 more organisms with golden standard annotation

Milestones

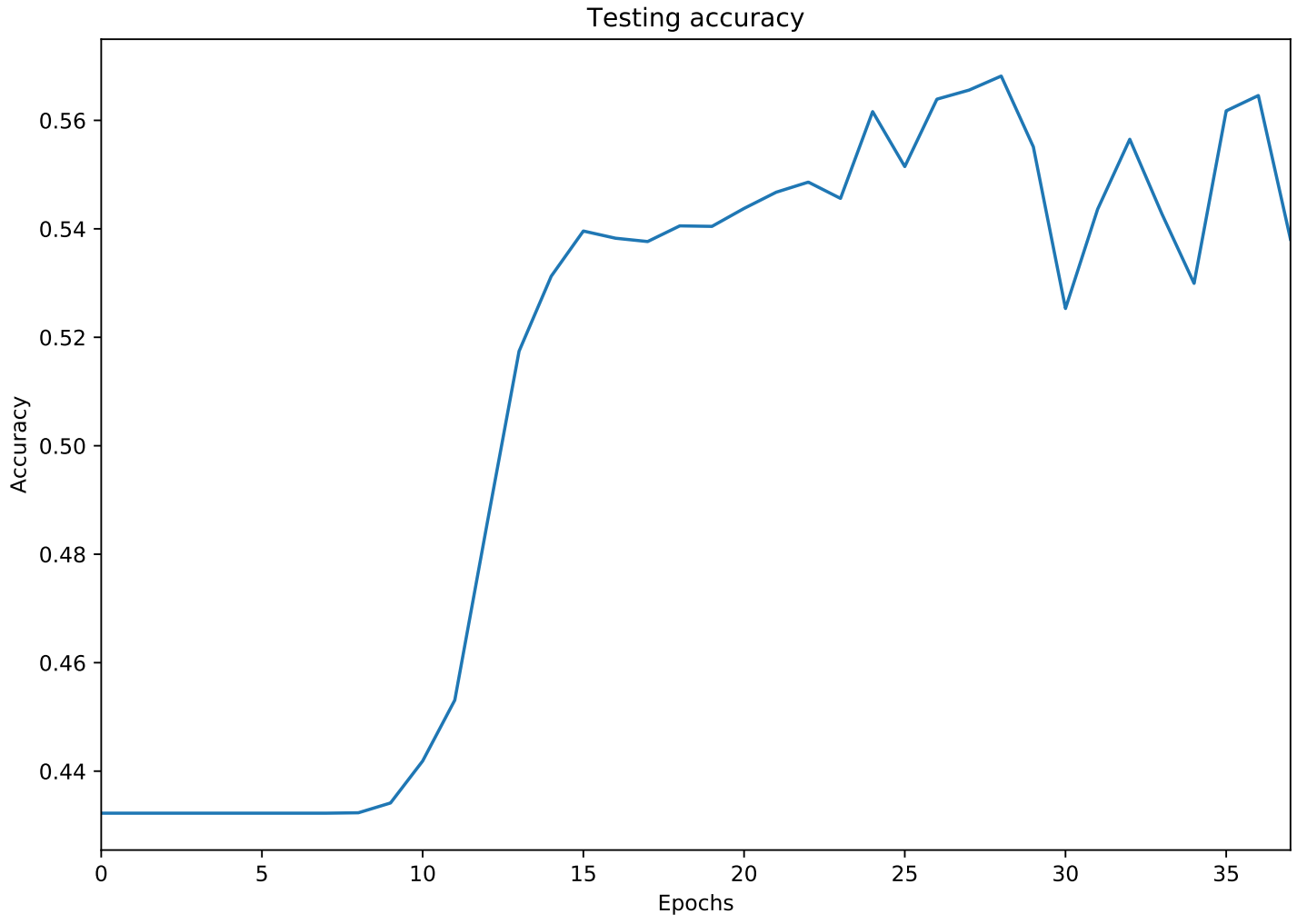
1. Preprocessed datasets : Raw data $\Rightarrow$ Encoded
2. Minimal implementation : RNN
3. Within organism prediction : Well annotated genes in humans and mouse

Model/Results

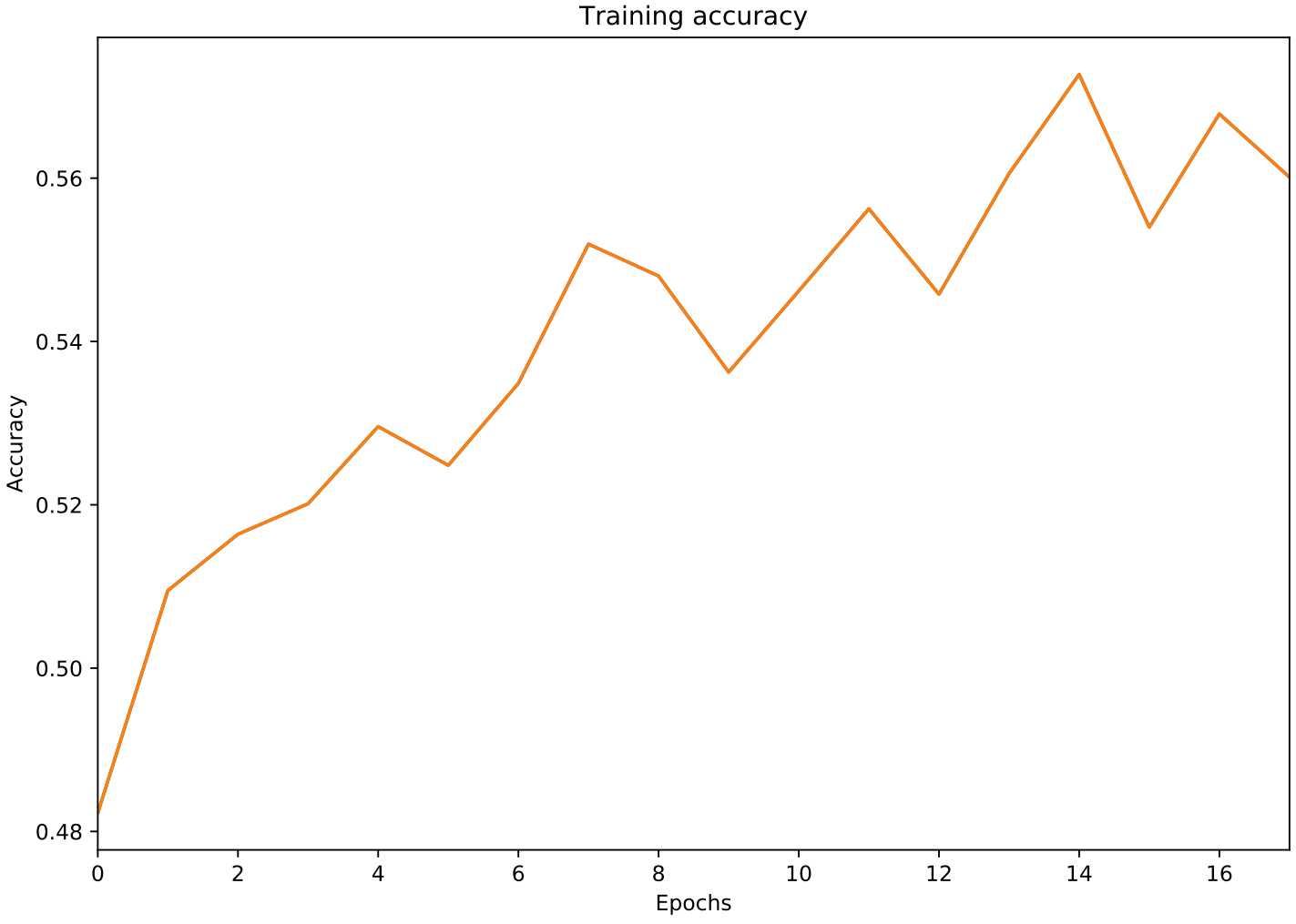
1. LSTM with sigmoid activation, 0.25 dropout
2. 20000 genes with length 200 - 10,000
3. Downsampled genes to 1000 first maintaining the length distribution
4. train:test = 70:30
5. 20 epochs so far

Preliminary results : Human (training)

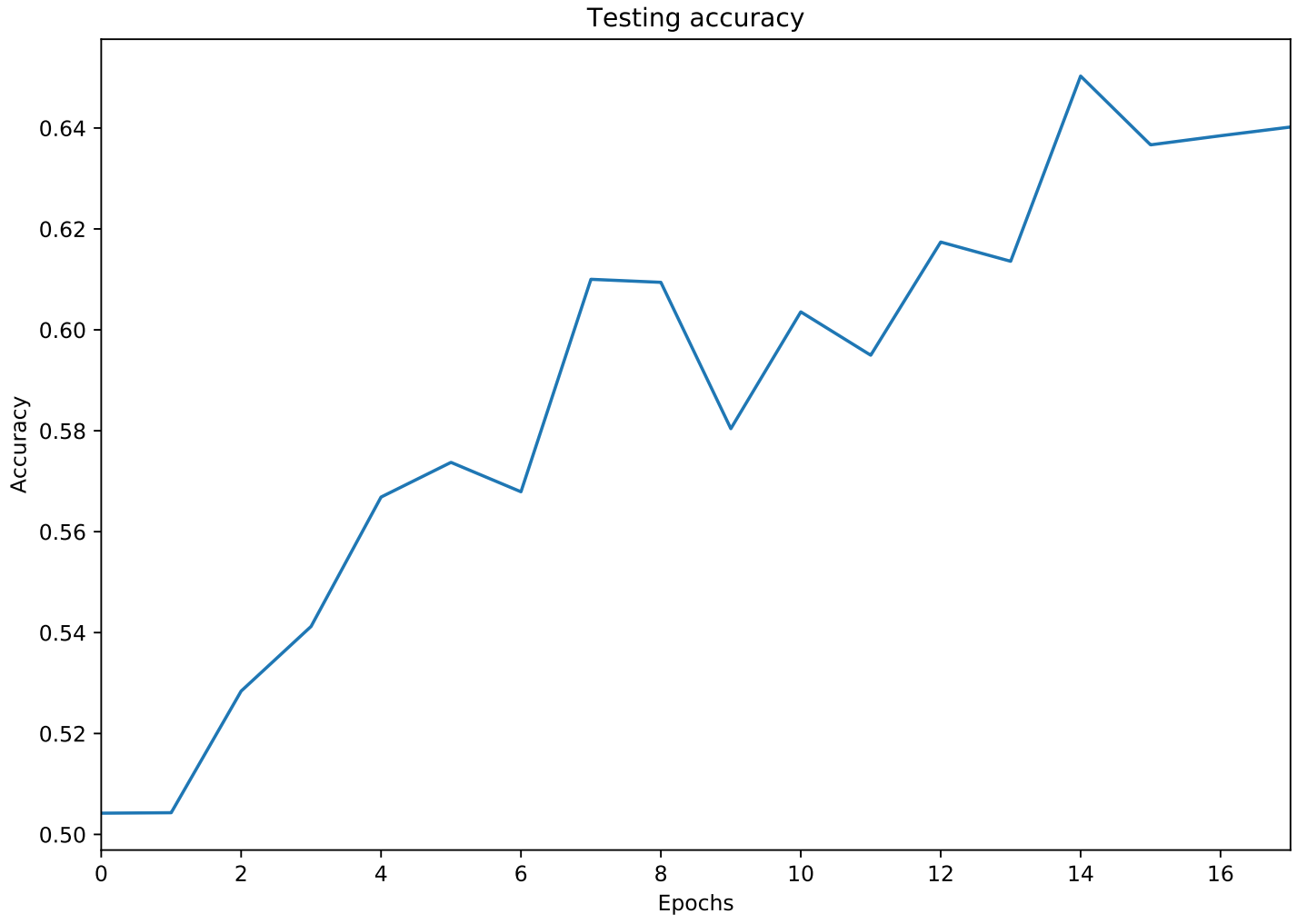
Preliminary results : Human (testing)



Preliminary results : Mouse (training)



Preliminary results : Mouse (testing)



Plan

1. Try with more epochs
2. Try increasing sample size

References I