ANNOTATING GENES IN NOISY READS THROUGH DEEP LEARNING

Delft University of Technology

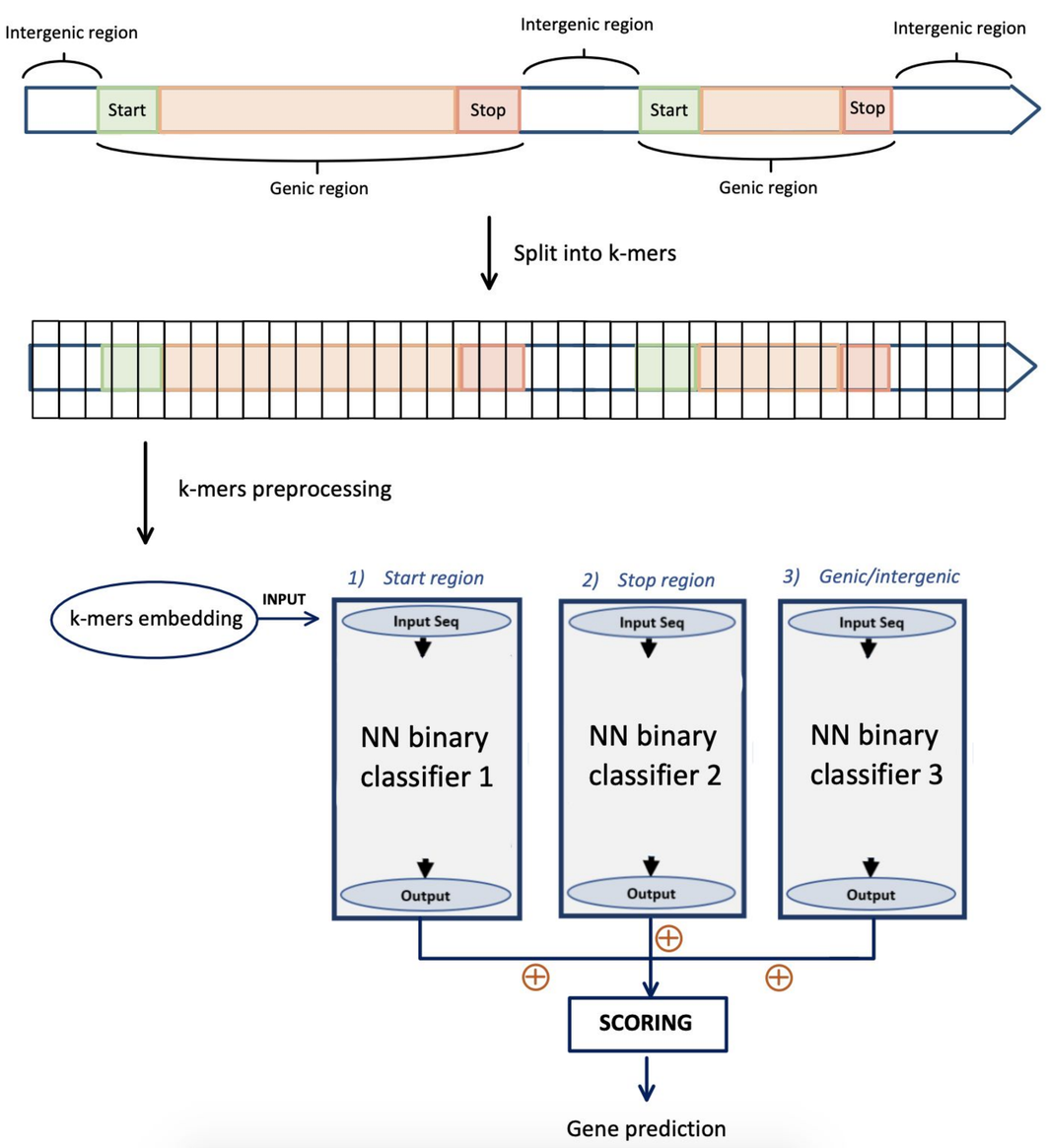
ABSTRACT

Gene annotation remains a key issue in genetic and genomic studies. Since newly sequenced genomes are usually compared with known sequences, annotation errors spread with the growth of sequence databases. In this project, we propose a new approach allowing to confront the current struggle in gene identification, with developing a deep learning based pipeline for long and noisy reads annotation. This work is a further step towards the annotation of whole new sequences, without referring to previously sequenced genomes.

WORKFLOW

What we do

- Identifying genes boundaries in long and noisy reads
- Genes are identified from scratch using Deep Learning



How

- Split long and random reads in short sequences of k nucleotides, called k-mers.

- Perform 3 different binary classifications of k-mers with LSTM neural networks:
  - k-mer ⊂ start region ⇒ label = 1, else ⇒ label = 0
  - k-mer ⊂ stop region ⇒ label = 1, else ⇒ label = 0
  - k-mer ⊂ coding region ⇒ label = 1, else ⇒ label = 0

- Candidate genes are sequences delimited by predicted start and stop distant from at least 150 bp and in frame.

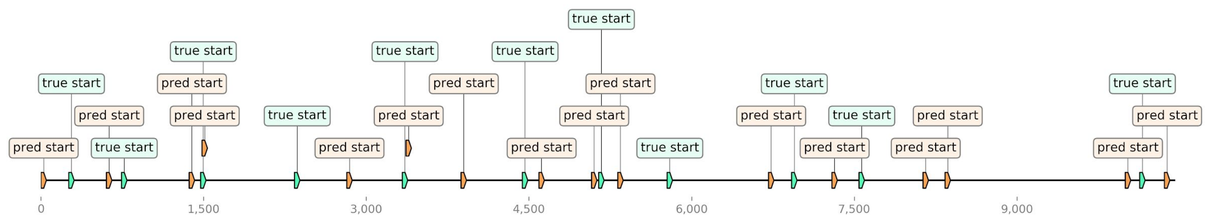
- Candidate genes are scored combining network's predictions for start, stop and genic coverage.

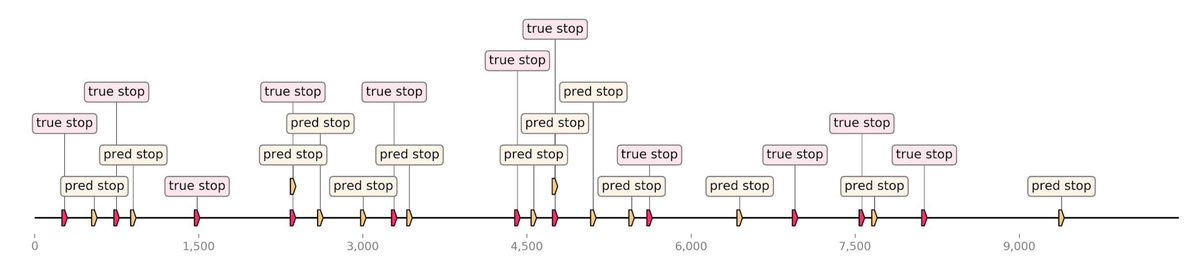
RESULTS

	WG TRAINING *Precision*	WG TRAINING *Recall*	WG VALIDATION *Precision*	WG VALIDATION *Recall*	NR TRAINING *Precision*	NR TRAINING *Recall*	NR VALIDATION *Precision*	NR VALIDATION *Recall*
Starts in WG	0.74	0.80	X	X	X	X	X	X
Stops in WG	0.63	0.63	X	X	X	X	X	X
Genic in WG	0.60	0.60	X	X	X	X	X	X
Starts in NR	X	X	0.71	0.71	0.72	0.72	0.64	0.62
Stops in NR	X	X	0.63	0.63	0.62	0.61	0.64	0.62
Genic in NR	X	X	0.59	0.57	0.57	0.62	0.58	0.62

Table summarizing the performance of the three networks.

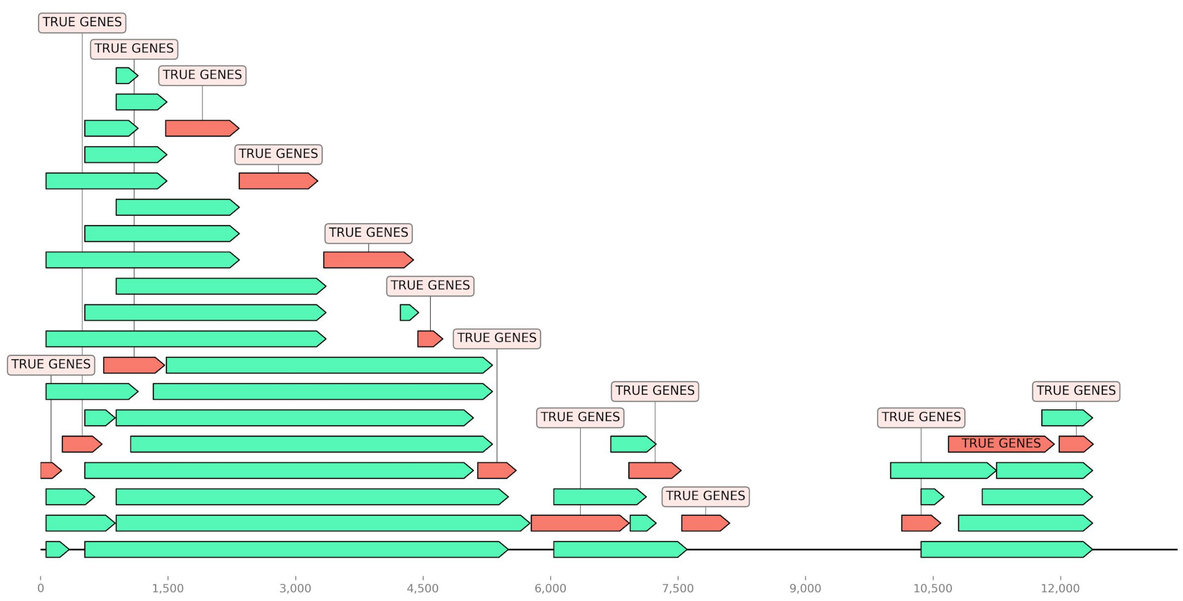
- Columns : Training metric (model trained on whole genomes, WG, or on noisy reads, NR)
- Rows: classification (start, stop, genic/intergenic) and the type of dataset on which the model is evaluated (WG or NR)





Boundary prediction in a random read from the E. coli genome. The scale shows relative location in the read.

- In green: true start codons.
- In red : true stop codons.
- In orange : predicted starts (above), or predicted stops (below)



Final candidate genes, having the highest predicted genic coverage by the network. These candidates are delimited by predicted starts and stops.

CONCLUSION

➢ ***Good points***

- Basic LSTMs are able to predict boundary DNA structures with a good precision
- These basic LSTMs can be trained on whole genomes
- Models are robust to noise

➢ ***Problem points***

- Model fails to classify more complex pattern as for coding/non-coding classification
- Do not reach statistical approach performance yet
- Still early to solve complex and multifactorial problems with Deep Learning.