Neural Networks: Gene finder
Build a neural network for finding genes in a genome sequence
Warning

- We haven’t tried this before and we don’t know how it will work.
- Consider it an experiment and try to be inventive.
- Because it is open-ended you should start early and we will spend a large part of next lecture discussing ideas and results.
The input for the project is a genome sequence.

We need to extract features (as numbers) we can feed into our neural net.

We also need to segment the sequence so we can identify the genes and the rest.
We can split the input sequence into windows (say of size 100)

\[ Z: \text{NNNCCCCCCCCNNNNNNNCCCCCCCCCCCCCCCCCCCCCCCCNNNNNNNNNNNN} \]

\[ X: \text{acgatcgctaatatgtccgatgacgtgagcataagcgacatgcag} \]
We can split the input sequence into windows (say of size 50, 100 or 200)

As features we can use nucleotide frequencies
We can split the input sequence into windows (say of size 50, 100 or 200)

As features we can use nucleotide frequencies

or we can use the frequencies of triplets
  - overlapping or non-overlapping?
The output should be a sequence indicating non-coding (N) or coding (C).

For non-overlapping windows we can assign class for the entire window
- can we handle overlapping window?