Gene Prediction

Bioinformatics algorithms | Winter Semester 2015

J. Setnička & J. Citorík

Finding genes with neural networks

1) Simplistic approach:

Input: subsequence of length w

Output: position of gene start or end (output layer of size w)

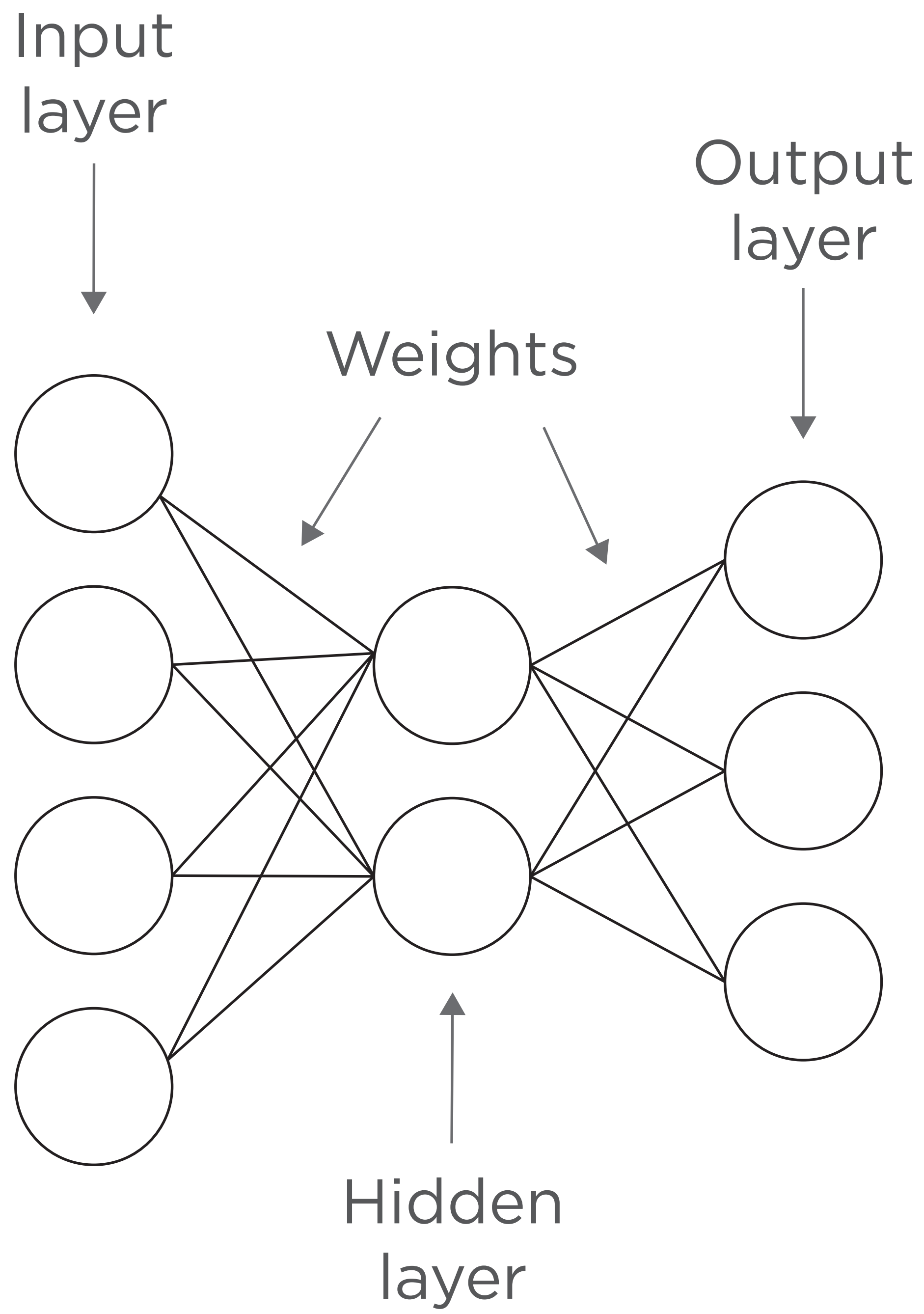
Results: terrible

2) k-mer frequencies:

Input: number of occurences for each k-mer within a window of size w

Output:
coding region: 1
non-coding region: 0

Results: it’s complicated…



Problems

What data should be used to teach?
(# of positive/negative samples, where to put the window)

Papers recommend using 6-mers
…but then we have 4096 input neurons and we have to compute frequencies of 4096 6-mers many many times

Window size

Evaluation
How to use the network once we've trained it?

Results:

Coding/Non-coding distinction

Increasing k-mer length didn’t help much
k = 3, 4, 5 didn’t differe significantly, couldn’t try k = 6 due to technical limitations

Window size that worked best: 100

Number of neurons in the hidden layer: 200

Best result for 3-mers
83.8% on training, 85% on test

Finding start/end of a gene

Move the window along the sequence, until NN says the region is a coding region
Find Start codon within that window, continue moving window while NN labels the region as coding region. In the first window which NN labels as non-coding, find stop codon.

Genes found were too long, real average was 1181 bases, found average was 1820

Approach: ORF Length (> 150 bases)

Average % genes found: 48.97%

Average % nonexistent genes found: 3294.42%

Sequences:
Bacteroides_fragilis_YCH46,
Bacteroides_ovatus_strain_ATCC8483,
Bacteroides_thetaiotaomicron_VPI5482,
Bacteroides_vulgatus_ATCC8482 ,
Bacteroides_xylanisolvens_XB1A

Approach: ORF start surroundings

Average % genes found: 38.38%

Average % nonexistent genes found: 59.83%

Trained on: Bacteroides_fragilis_YCH46, Bacteroides_ovatus_strain_ATCC8483, Bacteroides_thetaiotaomicron_VPI5482

Tested on: Bacteroides_vulgatus_ATCC8482 , Bacteroides_xylanisolvens_XB1A

Approach: CG islands

Average % genes found: 25.24%

Average % nonexistent genes found: 1147.87%

Trained on: Bacteroides_fragilis_YCH46, Bacteroides_ovatus_strain_ATCC8483, Bacteroides_thetaiotaomicron_VPI5482

Tested on: Bacteroides_vulgatus_ATCC8482 , Bacteroides_xylanisolvens_XB1A