Homework #2: Gene Prediction

Anže Pečar (63060257)

November 15, 2012

1 Introduction

In the second homework we analyse the genomic sequence of *Paramecium tetraurelia* and *Emiliania huxleyi virus 86*.

2 Data

Data was obtained from GenBank *Nucleotide NC_006058.1* and *Nucleotide NC_007346.1*. This time we were interested in protein coding genes. These genes are under *CDS* feature type on GenBank.

3 Methods

Figuring out which part of the DNA gets transcripted into genes is not an easy endeavour. The first problem that we encounter is the correct direction of the DNA string. Should we read it left to right or vice versa? It turns out that some proteins are transcribed from left to right and some from right to left, which means that we need to check both directions. Because we don't want to write a separate function for each direction we simply reverse and complement the original string. We also need to check for overlapping proteins. Because our codon size is 3 we can have 3 different frames for our open reading frames. If we now combine this with the two possible directions we get a total of 6 frames, where the proteins can be transcripted.

Once we have detected the ORFs we need a method of determining our success rate. In this homework we used the $F1$ score, which is a combination of precision and recall, where precision is defined as

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

and recall is defined as

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

4 Results

4.1 Detecting open reading frames (ORFs)

5081 ORFs code for at least 60 amino acids.

4.2 Precision and recall

In Figure 1 we can see how precision and recall change depending on the minimum ORF length.

The values for precision and recall for 125 codons or more are 0.1026 and 0.4017.

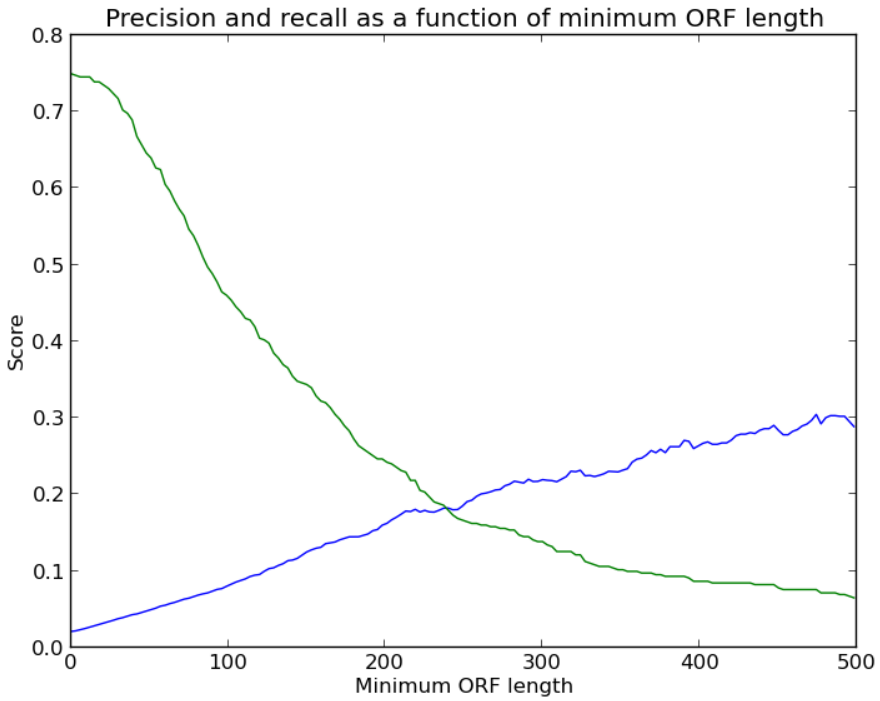

Figure 1: Precision and recall for Paramecium. Blue line = precision. Green line = recall.

4.3 Emiliania huxleyi virus 86

4.3.1 Detecting ORFs

1434 ORFs code for at least 60 amino acids.

4.3.2 Precision and recall

In Figure 2 we can see how precision and recall change depending on the minimum ORF length. The scores for precision and recall for 125 codons or more are **0.6797** and **0.6653** respectively.

It seems the *Emiliania huxleyi virus* gave a considerably better score. I am guessing the reason for this is the fact that a virus is a much simpler organism than the *Paramecium*. The

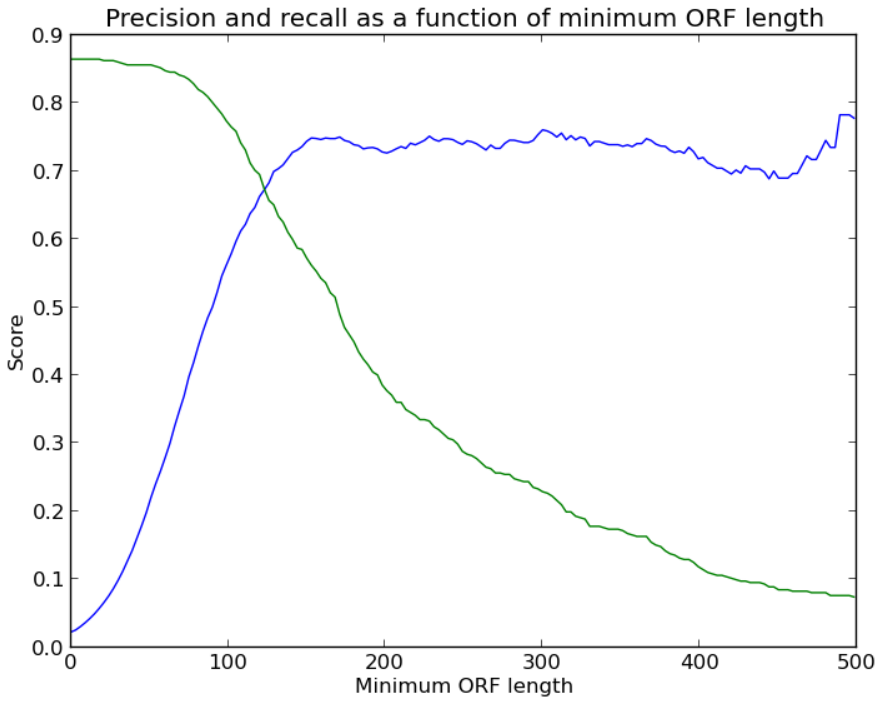

Figure 2: Precision and recall in Emiliania huxleyi virus. Blue line = precision. Green line = recall.

virus only has a few functions that need to be executed and the source code for them is not as complex and thus easier to decode.

Honor Code

My answers to homework are my own work. I did not make solutions or code available to anyone else. I did not engage in any other activities that will dishonestly improve my results or dishonestly improve/hurt the results of others.