# 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*. This time we were interested in protein coding genes. In the GenBank database this relates to CDS feature type.

## 3 Methods

## 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.

### 4.3 Precision and recall for 125 codons or more

0.1026 precision, 0.4017 recall

### Emiliania huxleyi virus 86

#### 4.3.1 Detecting ORFs

2264 ORFs code for at least 60 amino acids.

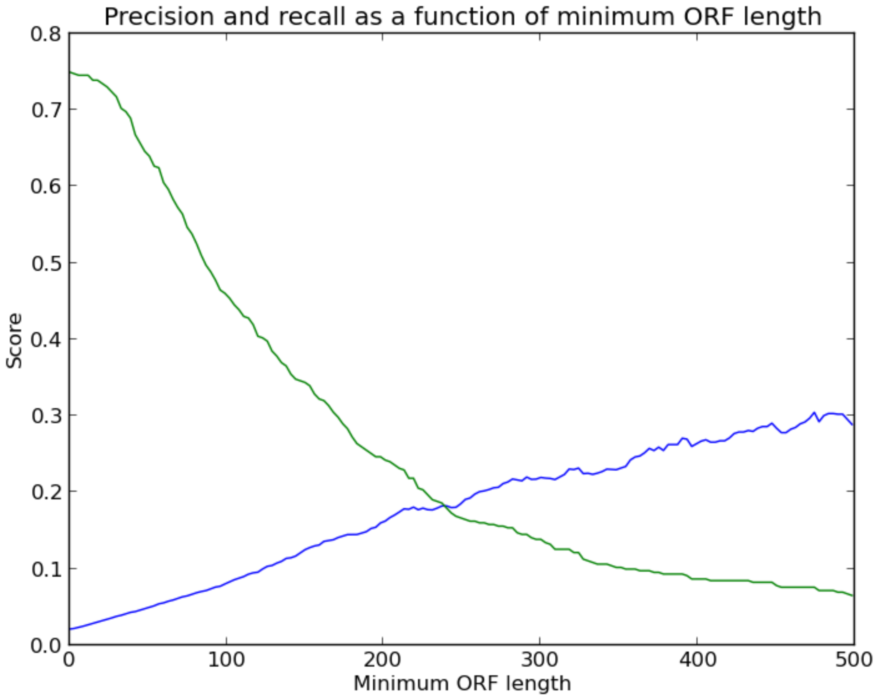

Figure 1: Blue line = precision. Green line = recall.

### 4.3.2 Precision and recall

In Figure 2 we can see how precision and recall change depending on the minimum ORF length. 0.1254 precision, 0.2691 recall

## 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.

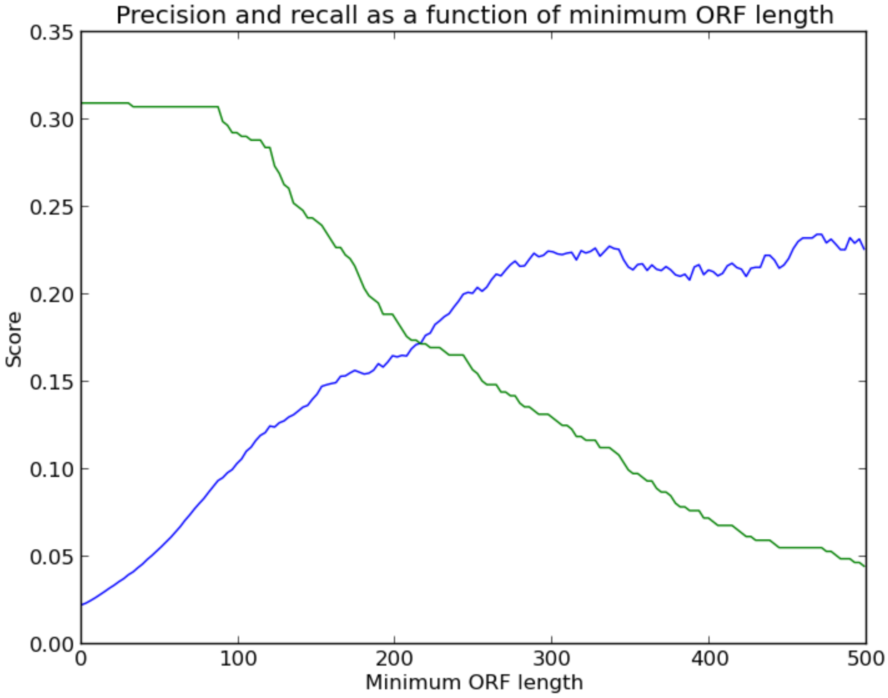

Figure 2: Blue line = precision. Green line = recall.