Gene Prediction

Miha Zidar (63060317)

November 16, 2012

1 Introduction

In this homework we look at a basic prediction of open reading frames (from now ORFs) in *Paramecium tetraurelia* and *Emiliania huxleyi virus 86*. We also test how good the most simple predicting algorithm works.

2 Data

Data for this homework was retrieved from a gene bank with BioPython. We used genome sequences for *Paramecium tetraurelia (NC_006058)* and *Emiliania huxleyi virus 86 (NC_007346)*. Loaded data cotains the seqence and all known features of the genome with some usefull functions for manipulating the sequence. We used downloaded genome features to test the precision and recall of our algorithm.

3 Methods

3.1 Detecting ORFs

All genomes contains many ORFs. An ORF is determened by a start codon and a stop codon. Our algorithm will have to match start and stop codons so that the folowing will hold:

- start and stop codon have to be aligned so that there are only multiples of three codons between.
- if we have multiple possible start codons, we'll choose the one that gives us the longest ORF.
- there should not be any alinged stop codons between a start and a stop codon of an ORF.
- we have to repeat the search on the reverse complement sequence.

ORFs may overlap but only if they are not aligned, if they're on the reverse strand of the genome.

3.2 Precission and recall

To assess the ORF predictions of our algorithm, we will take a look at precission and recall, and how those change when we limit our results with a certain lenght. These two mesurements are defined as:

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

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

3.3 Area under PR-curve

If we plot precission as a function of Recall 3, we could compare the surface area unded the curve with some other algorithm. It can bs shown that this kind of comparison will usually yield better results than a simmilar AUC score.

4 Results

Table 1: Some basic data for both genomes.

	NC_006058	NC_007346
# of real ORFs	463	472
# of predicted ORFs	16646	18088
average length	548.03	261.70
average predicted length	59.75	27.09
# of predicted ORFs longer than 60	5174	1469
precision for at least 125 codons:	0.10087	0.67811
recall for at least 125 codons	0.40172	0.66949
area under PR-curve	0.10785	0.60254

The second genome, has more start and stop codons, and that meakes it easy to match a vampire from a human, since the average ORF length is half as small.

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.

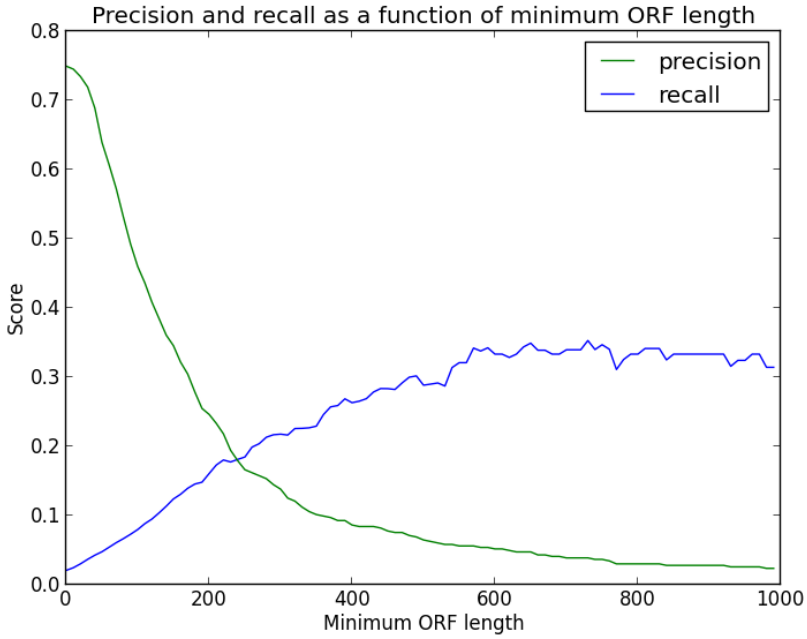

Figure 1: Precission and Recall as they depend on the minimum ORF legth for NC_006058

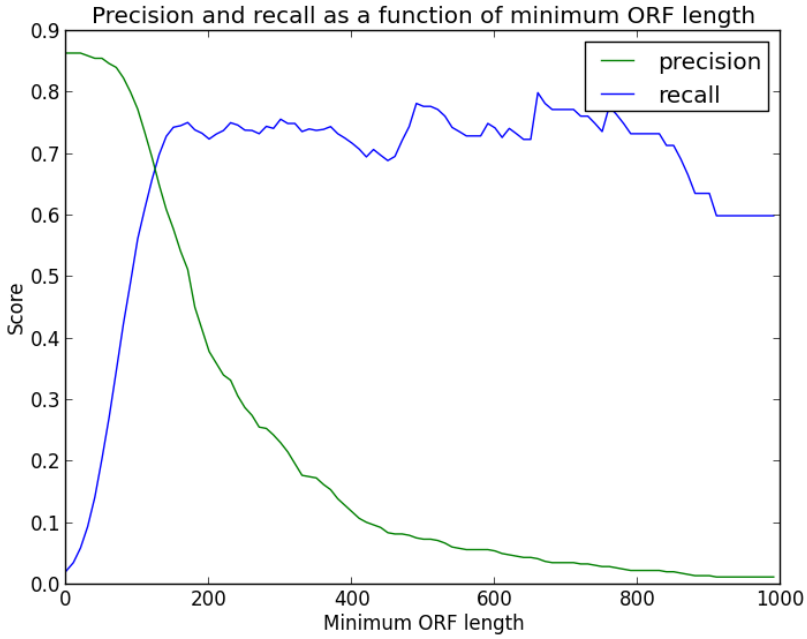

Figure 2: Precission and Recall as they depend on the minimum ORF legth for NC_007346

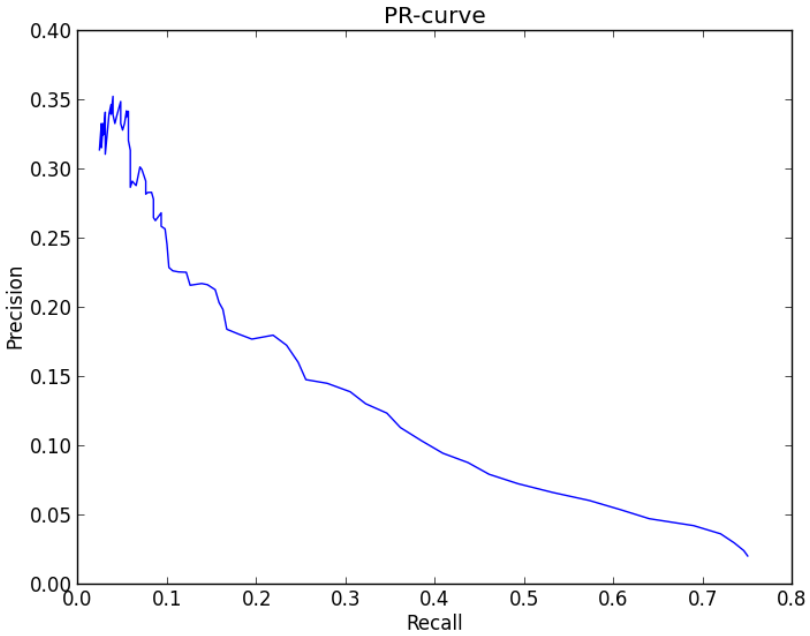

Figure 3: PR-curve from which we'll calculate the area under the curve for NC_006058