Mini Project 1: Where are the Genes

Klemen Krenker (63140132)

December 1, 2019

1 Introduction

Goal of the mini project was to implement a gene-finding algorithm that finds the genes in Mycoplasma genitalium. Genes span from start to stop codon, stop codons in this particular genome are only TAA and TAG however. A filtering system was also needed, since genes can't be too short, that value is called L, and it filters out genes shorter than L codons. As a summary of our findings across a multiple sample of L values, a precision and recall graph was needed.

2 Data

Genome analysed was Mycoplasma genitalium, accessed through NCBI database with a BioPython library. The sequence contains 580076 nucleotides or 193,358.6 codons, which are then translated into amino acids using a translation table.

3 Methods

The algorithm contains the following steps in producing a solution:

- Acquiring of the genome (and the gene solutions) from NCBI database,
- reading frames are then computed (taking into account offsets and reversing and divisibility by 3,
- the sequence is then translated into amino acids using the coding table, taking 3 nucleotides (codons) and turning them into an appropriate amino acid,
- each reading frame is then examined, storing all the start codons (M),
- for every M, a gene is found (by moving along the genome as far as possible before encountering a stop codon (*),
- found gene is valid if, and only if it is long enough (the value is compared with L which is an input parameter), if the gene is long enough, its stored as a found gene, otherwise it is rejected,
- a final analysis is made on the results and a graph is produced from the results.

4 Results

As a summary of results a few questions need to be answered..

- **What is the size of Mycoplasma genitalium genome?** Mycoplasma genitalium's genome size is 580076 bp.
- **How many genes does it include?** In total it includes 563 genes, for this homework we've taken into account only the CDS genes, 509 in total.
- **What is the length of the smallest and the longest gene (in codons)? What is the median length of the gene (in codons)?** Smallest gene contains 37 codons, the largest 1805. Median length of a gene is 287 codons.
- **What is the recall/precision of your gene finding procedure at L=50 and L=125 codons?** 50: 0.8527/0.1237, 125: 0.7525/0.2007.

My algorithm produces the following precision/recall graph, shown in Figure 1.

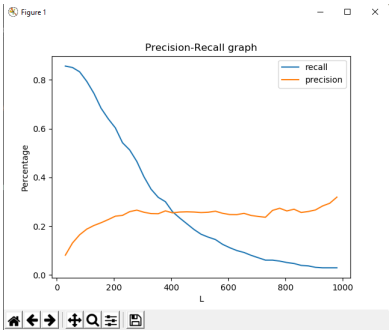

Figure 1: Precision/recall graph for the gene-finding algorithm

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.