Sequence Statistics - Assignment 1

Franco Chiesa Docampo and Luis Felipe Aycardi

Bioinformatics - LGBIO2010

1 Introduction

A fundamental task in bioinformatics is to analyze DNA sequences. In this document real biological data were studied with the aid of several computing methods, in order to perform change point analysis, register the appearance of unusual dimers and perform ORF finding in DNA sequences.

To develop these tasks we will use R, a programming language that offers convenient graphical functions and several statistical processing packages that allow us to have a better comprehension of the genomic data.

2 Change Point Analysis

The task consisted on performing the GC-content analysis of a specific genome. The following were the computational and analysis tools implemented in the development:

- Lactococcus lactis subsp. lactis Il1403 genome sequence (Accession number NC_002662).
- Ape package for sequence recognition, in R.

The main steps went from accessing some real data, to computing some statistics and the understanding the algorithm for analyzing the GC-content of a DNA sequence.

1. The full genome of the Lactococcus lactis subsp. lactis Il1403 was downloaded. Its length was of 2365589 base pairs (bp).
2. The size of the sliding window chosen to generate the plot of its GC and AT content was of 6000 base pairs. The step of such moving window was of 2000 base pairs. These 2 parameters allowed us to represent plots that matched the ones in the assignment statement. Figures 1 and 2 showed our findings.
3. These plots exhibit anomalies in terms of GC and AT content between the vector windows positions ("positionsSeqWindow") that go from 250 to 350 and from 980 to 1180 (see Figures 3, 4, 5 and 6).
4. To understand the influence of reducing and enlarging the window size we looked at the formula used to calculate the number of window frames, located at line 19 of the script. On a constant value of step size for the window, if we decrease the value of the window we will encounter a bigger number of window frames. This provides a better data resolution when we plot the results. If we increase window size, plot resolution decrements (see Figures 7, 8). At a constant level of window size, the same phenomena happens when we modify the step size of the moving window. Bigger step size gives lower resolution and vice versa (see Figures 9, 10).

3 Unusual Dimers

The objective consisted on calculating the observed frequencies of all dimers in the previous sequence, in order to deduct which ones differed the most from their expected frequencies and if they were under-represented or over-represented in the sequence.

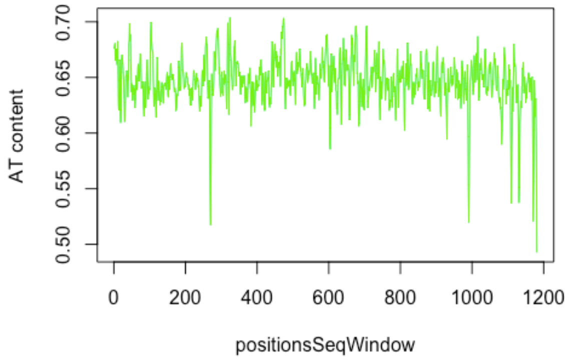


Figure 1: AT content of the DNA sequence treated.

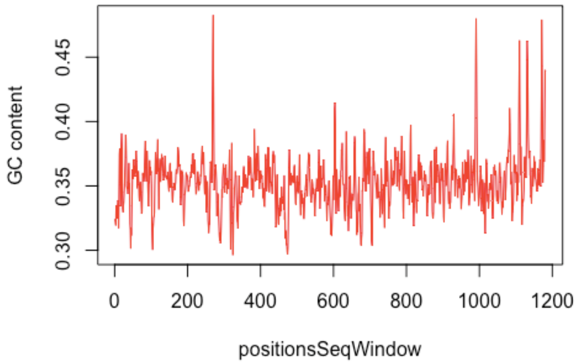


Figure 2: GC content of the DNA sequence treated.

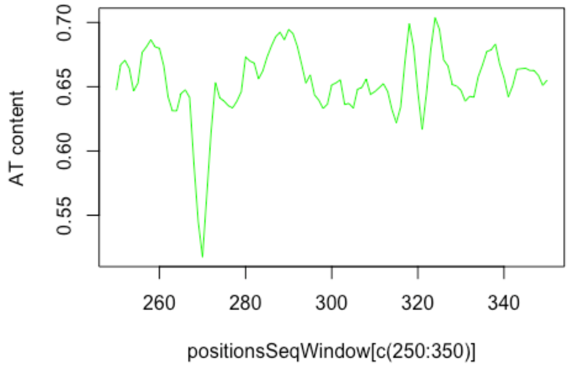


Figure 3: AT content of the DNA sequence treated from window frame 250 to 350.

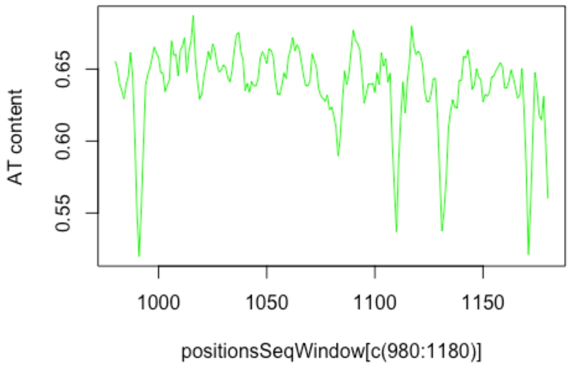


Figure 4: AT content of the DNA sequence treated from window frame 980 to 1180.

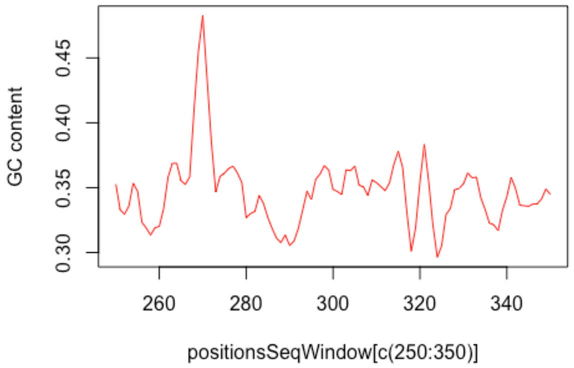


Figure 5: GC content of the DNA sequence treated from window frame 250 to 350.

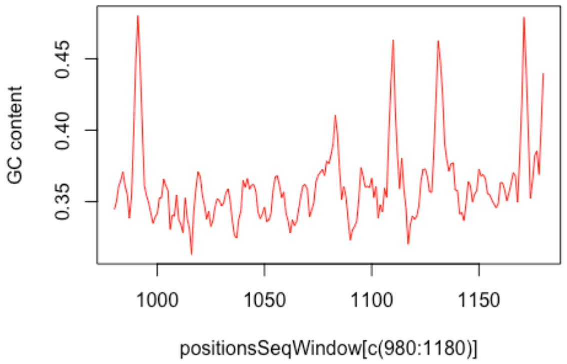


Figure 6: GC content of the DNA sequence treated from window frame 980 to 1180.

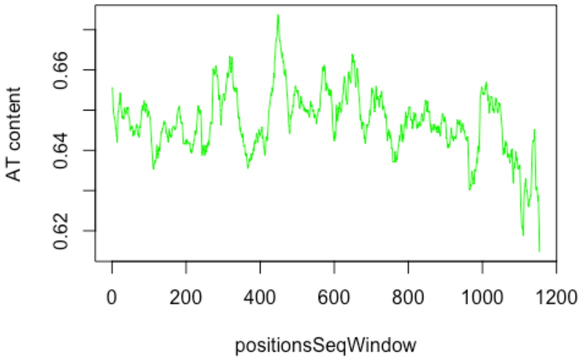


Figure 7: AT bigger window size modification (60000 bp) with a step of 2000 bp.

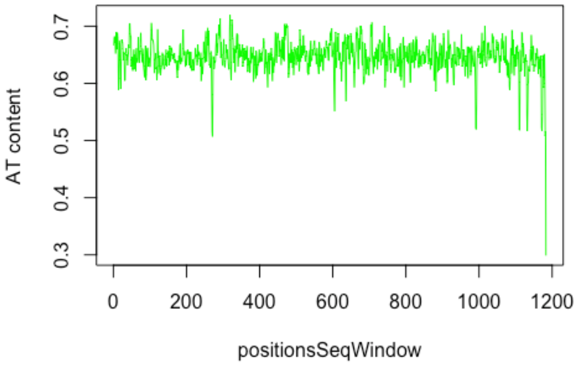


Figure 8: AT smaller window size modification (3000 bp) with a step of 2000 bp.

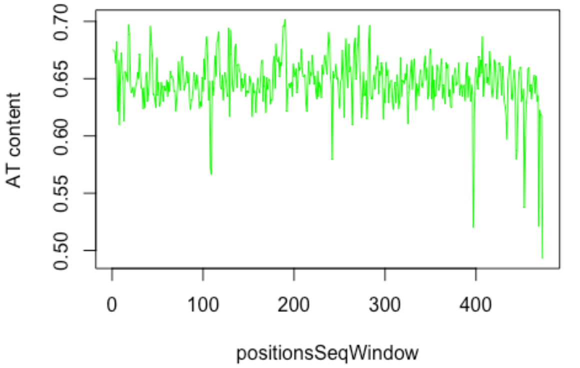


Figure 9: AT bigger step modification (5000 bp) with a window size of 6000 bp.

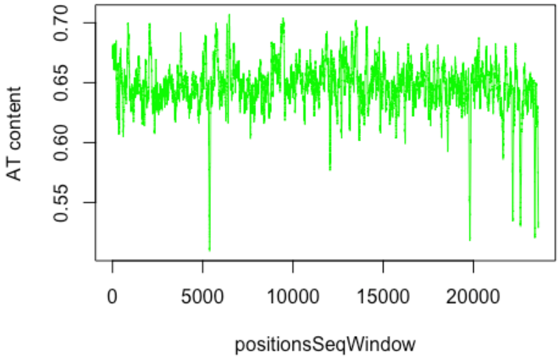


Figure 10: AT smaller step modification (100 bp) with a window size of 6000 bp.

The following were the computational and analysis tools implemented in the development:

- Lactococcus lactis subsp. lactis Il1403 genome sequence (Accession number NC_002662).
- Ape package for sequence recognition, in R.
- Seqinr package for k-mers count in a sequence.

1. The length of the "Lactococcus lactis subsp. lactis il1403" was of 2365589 base pairs.

2. The observed frequencies of all dimers in this sequence are presented in Table 1.

Table 1: Observed dimer frequencies in the *Lactococcus lactis subsp. lactis Il1403 genome.*

	*A	*C	*G	*T
A*	0.129001	0.046834	0.055454	0.092483
C*	0.063965	0.032509	0.024176	0.054891
G*	0.060319	0.037065	0.033302	0.047063
T*	0.070487	0.059133	0.064817	0.128494

3. The expected frequency of dimers in this table if all of them would be equally likely is 0.0625. The dimers departing most from this expected value are AA (over-represented), CC (rare), CG (rare), GC (rare), GG (rare) and TT (over-represented).

4. The odd ratios (under a multinomial background model) of all dimers in this sequence are presented in Table 2.

Table 2: Odd ratios of dimers in the *Lactococcus lactis subsp. lactis l1403 genome.*

	*A	*C	*G	*T
A*	1.230587	0.824029	0.963564	0.884522
C*	1.125434	1.054970	0.774820	0.968303
G*	1.048101	1.187875	1.054036	0.819896
T*	0.674151	1.043127	1.129191	1.232141

5. The most unusual dimers in this sequence are AA (unexpectedly frequent), TT (unexpectedly frequent), CG (unexpectedly rare) and TA (unexpectedly rare). According to the original analysis in question 3: AA, TT, TA are over-represented and CG is rare. Looking for example at dimer TA, results depart from initial analysis in question 3.

   This mismatch allowed us to think there is no statistical connection between the observed dimer frequencies being higher or lower than 0.0625 (thus being “over-represented” or “rare”) and the observed frequencies being higher or lower than the expected one (which is defined by the product of the relative base frequencies involved in the dimer) because when the observed frequencies are compared to 0.0625 the relative base frequencies are not being taken into account. We believe the odd ratio offers a more precise way of determining if a dimer is over-represented or rare because it takes into account the expected frequency (formed by the product of the relative base frequencies).

4 ORF Finding

Reading frames are understood as the three possible sequences of codons by which a genetic translation may occur from one nucleotide sequence [1]. There are 6 possible reading frames for a DNA sequence: 3 on the original sequence and 3 on its reverse complement. In each one of them, Open Reading Frames (ORF) can be found. An ORF is a sequence of successive nucleotide triplets that are read as codons specifying amino acids and begin with an initiator codon and end with a stop codon [1].

In this work the bacterial DNA of the Bacillus cereus B4264 was analyzed; for which the start codons are TTG, CTG, ATA, ATT, ATC, ATG and GTG and the stop codons are TGA, TAA and TAG.

The following were the computational and analysis tools implemented in the development:

- Bacillus cereus B4264 genome sequence (Accession number NC_011725 from GenBank).
- Biosting library from the Bioconductor package for R programming.
- Ape package for sequence recognition in R.

The main steps went from accessing the data, to computing some statistics and algorithms for finding ORF in a DNA sequence.

1. The full genome of Bacillus cereus B4264 was downloaded. Its length was of 5419036 base pairs (bp).

2. In the sequence (considering all reading frames), a total of 418913 ORFs were found. Of them, a classification regarding the minimum number of nucleotides long of each ORF was performed. The results are presented in Table 3. The length of each sequence was measured from the first nucleotide in the start codon, to the first nucleotide of the end codon.

Table 3: ORFs quantity with at least k nucleotides minimum length value long.

k	Number of ORF found
10	331989
50	133014
100	48480
300	6547
500	3996

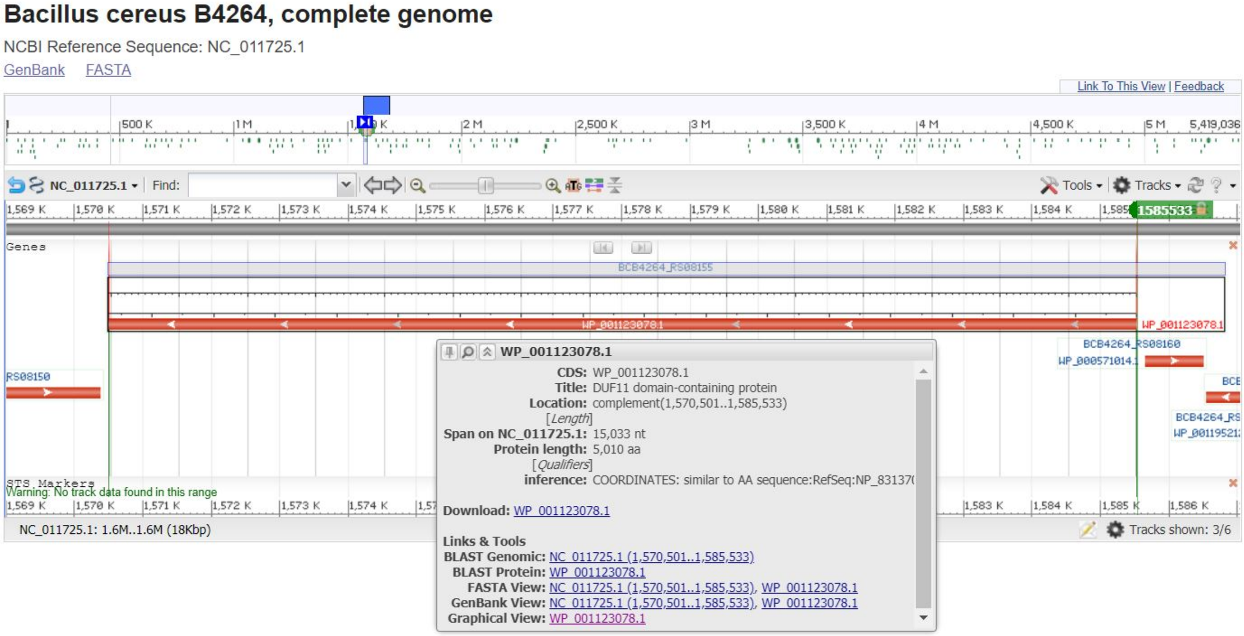
3. Then, to asses the statistical significance of the found ORFs, the distribution of ORF length in a NULL model estimated from a random permutation of the DNA was computed.
   The maximal length of an ORF found according to this random model was of 570 nucleotides. Which meant that 3535 ORFs in the actual DNA were strictly longer than the maximum obtained by the permutation.
4. Considering a p-value of 1% as significance threshold, 13801 candidate genes were found in this DNA.
5. According to the information found, the longest gene candidate had the following characteristics:
   - Exact position: 3833504 bp
   - Length: 15030 bp
   - Reading frame: 3
   - DNA sequence: Reverse complement DNA

   Ones this results were found, a research of the DNA sequence was conducted at the National Center for Biotechnology Information (NCBI) page, more precisely the primary nucleic acid sequence databank *GeneBank*. Following the information found in the previous procedure, a gene DUF11 domain-containing protein was matched, as presented in the Figures 11 and 12.

This protein (of NCBI Reference Sequence WP_001123078.1) DUF11 stands for a domain of unknown function [2]. In the Figure 13 a summary of the information found about this protein is presented.

5 Conclusion

The previous work constitutes the first approach to the analysis of real biological data through several computing methods. With the use R, a free software environment for statistical computing and graphics, and several packages there available, procedures including change point analysis, the registration of



Bacillus cereus B4264, complete genome

NCBI Reference Sequence: NC_011725.1

GenBank FASTA

WP_001123078.1

CDS: WP_001123078.1
Title: DUF11 domain-containing protein
Location: complement(1,570,501..1,585,533)
[*Length*]
Span on NC_011725.1: 15,033 nt
Protein length: 5,010 aa
[*Qualifiers*]
inference: COORDINATES: similar to AA sequence:RefSeq:NP_83137(

Download: WP_001123078.1

Links & Tools
BLAST Genomic: NC_011725.1 (1,570,501..1,585,533)
BLAST Protein: WP_001123078.1
FASTA View: NC_011725.1 (1,570,501..1,585,533), WP_001123078.1
GenBank View: NC_011725.1 (1,570,501..1,585,533), WP_001123078.1
Graphical View: WP_001123078.1

Warning: No track data found in this range

NC_011725.1: 1.6M..1.6M (18Kbp)

Figure 11: Gene match with the longest gene in the Bacillus cereus B4264 complete genoma.

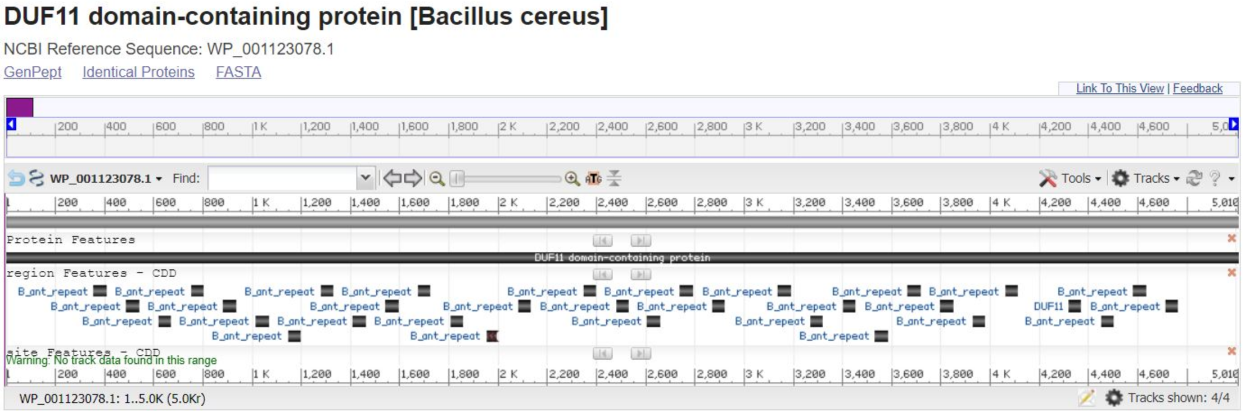
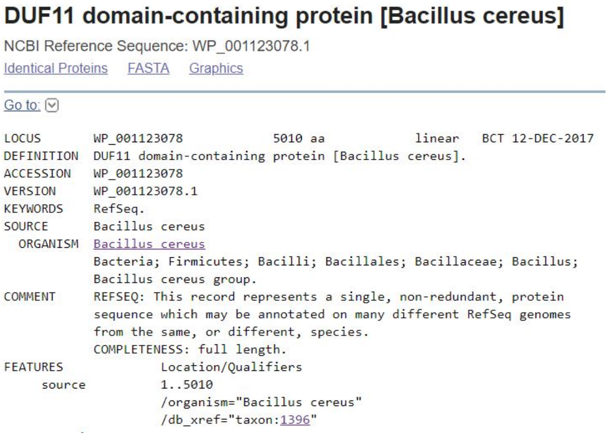

DUF11 domain-containing protein [Bacillus cereus]
NCBI Reference Sequence: WP_001123078.1
GenPept
Identical Proteins
FASTA
Link To This View | Feedback
WP_001123078.1
Find:
Tools
Tracks
Protein Features
DUF11 domain-containing protein
region Features - CDD
B_ant_repeat
DUF11
site Features - CDD
Warning: No track data found in this range
WP_001123078.1: 1..5.0K (5.0Kr)
Tracks shown: 4/4

Figure 12: Grapich representation of the DUF 11 domain-containing protein.


DUF11 domain-containing protein [Bacillus cereus]

NCBI Reference Sequence: WP_001123078.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       WP_001123078            5010 aa            linear   BCT 12-DEC-2017
DEFINITION  DUF11 domain-containing protein [Bacillus cereus].
ACCESSION   WP_001123078
VERSION     WP_001123078.1
KEYWORDS    RefSeq.
SOURCE      Bacillus cereus
  ORGANISM  Bacillus cereus
            Bacteria; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus;
            Bacillus cereus group.
COMMENT     REFSEQ: This record represents a single, non-redundant, protein
            sequence which may be annotated on many different RefSeq genomes
            from the same, or different, species.
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..5010
                     /organism="Bacillus cereus"
                     /db_xref="taxon:1396"
```

Figure 13: General attributes of the found gene.

appearance of unusual dimers and the ORF finding in DNA sequences were successfully performed. As secondary exploration methods, DNA sequence databases were consulted.

References

[1] Open reading frames - mesh result - ncbi. https://www.ncbi.nlm.nih.gov/mesh?Db=mesh &term=Open+Reading+Frames. Accessed: 2019-03-11.

[2] Lillian Reiter, Nicolas J. Tourasse, Agnès Fouet, Raphaël Loll, Sophie Davison, Ole Andreas Økstad, Armin P. Piehler, and Anne-Brit Kolstø. Evolutionary history and functional characterization of three large genes involved in sporulation in bacillus cereus group bacteria. *Journal of Bacteriology*, 193(19):5420–5430, 2011.