Seminario III: R/Bioconductor

José Reyes
jreyes@lcg.unam.mx

29 de Septiembre del 2009

Professor: Leonardo Collado (lcollado@lcg.unam.mx)

Assistants: Alejandro Reyes (areyes@lcg.unam.mx), José Reyes (jreyes@lcg.unam.mx), Víctor Moreno (jmoreno@lcg.unam.mx)

Abstract

These is the exercise for the Biostrings package session. The only mandatory exercise is the first one. The other ones are interesting examples for those who want to learn a little more about this package. Even though these are optional, if you do them they will be considered for your evaluation as extra points.

1 GC skew in bacterial genomes

This is the only mandatory exercise. You will visualize the GC skew of a bacterial genome.

1. Load the E. coli genome from the BSgenome package. You will be working with *Escherichia coli str. K12 substr. MG1655* (NC_000913)

2. Create a set of 500bp long overlapping windows, sliding 100bp and covering as much of the chromosome as possible (The first window will go from 1-500, the next one will go from 101-600 and so on)

3. For each window, estimate the (G-C)/(G+C) index

4. Obtain the start position and strand of all the protein coding genes from this bacterium, using the "bacterial_mart_54" mart and the "esc_18_gene" dataset

5. Using the function *density*, estimate the density of protein coding genes in the plus and minus strands. The density function will return an object with a x field and a y field. The x corresponds to the chromosome location and the y corresponds to the estimated density.

6. Use the *makeGene* function to plot the location of DnaA (emsembl_gene_id = "EBESCG00000003429").

7. Make a genome graph with five tracks:

   - Gene density in + strand.

- Chromosome axis.
- Gene density in - strand.
- DnaA gene.
- The index you calculated

8. Write a brief conclusion about the results.

2 Optional exercises

1. Obtain a nucleotide sequence from EBESCG00000001846 (571591, 571689) and use the function *replaceLetterAt* to introduce every posible point mutation to the sequence. What is the proportion of:
   - Synonymous substitutions?
   - Non-synonymous substitutions?
   - Non-sense mutations?
2. Use the Biostrings package as an alignment editor. Try to add numbers to the names of a FASTA alingment and to eliminate columns containing more than 50% gaps.
3. Extract the (-100,+10) region for each transcript in the E. coli chromosome using a combination of BSgenome and biomaRt packages. For a sample of these sequences, create overlapping windows of lenght 20bp, sliding 10bp each step, and calculate the GC content. Plot the resulting profile of the sampled sequences (use lines to overlap the profiles). Do you notice an interesting pattern?
4. Make a function to generate random BStrings given a background model (for example, nucleotide or dinucleotide frequencies). Extract the upstream sequence for each transcript in E. coli. Calculate the nucleotide and dinucleotide fequencies to generate random sequences of about 5000bp. The sequences you generated may serve as random controls for a transcription factor binding site motif search.