Scanning non-coding sequences with a TFBM

DUBii 2019

Jacques van Helden

2019-02-06

Preparation of the exercise

Collective table for the 2019 practical

Students will store their results in a shared spreadsheet, which will be used to compare their results and get a broader landscape from the comparison of the results obtained with different transcription factors.

- Folder: https://tinyurl.com/lcg-beii-19
- Motif scanning exercise:

In your computer, create a folder to store the results of this practical, for example : `$HOME/LCG_BEII_practicals/` (you can change the path and name according to your own organisation of folders).

Choosing a TF on RegulonDB

- Open a connection to RegulonDB http://regulondb.ccg.unam.mx/
- Click on the link regulon list. This opens a table with all the regulons.
- Choose a TF of interest and open its record
- Fill up the details of the collective exploration table (https://tinyurl.com/lcg-beii-19).
- Save a fasta file with the sequences of the known binding sites for your TF (tip: click on the bug "+" button in the header of the binding site section)
- Save in a text file the matrix associated to your factor.

Computing the degenerate consensus from the reference matrix

- Connect RSAT server: http://rsat.eu/
- Choose the bacterial server
- Use **convert-matrix** to compute frequencies, weights, parameters and display a logo of your matrix.
- In the result, get the degenerated consensus and save it to a separate text file.

Getting all upstream ("promoter") sequences o E.coli

- Open the tool **retrieve-seq**
- Select organism *Escherichia coli K12* (top : type simply K12 in the organism query box)
- Set all parameters to get the non-coding sequences located upstream of all genes with a maximal distance of 400 bp from the gene start
- Copy the URL of the result file and save it in a text file (we will use it several times below)

Coverage of the annotated binding sites by the reference motif

- Use **dna-pattern** to scan the annotated binding sites (extracted from RegulonDB) with the degenerate consensus.
- Use **matrix-scan** to scan the same sites with the RegulonDB matrix
- Compare the coverage rate of the two motifs

Binding site prediction in all promoters

- Use the same tools (dna-pattern and matrix-scan) to predict binding sites in all the promoters of E.coli.
- For **matrix-scan**, run the analysis with a threshold of p-value of either 0.001 or 0.0001.
- Compare the number of matches obtained in these respective searches.
- With the respective p-values used for the scanning, how many matches would you expect by chance ?

Negative control 1: scan artificial sequences with your motif

- RSAT random sequences

Negative control 2: permute the columns of the matrix

- Use the tool **permute-matrix** in order to generate 10 randomized copies of the motif
- Send these randomiazed matrices to **convert-matrix** and check their logo.
- Run the same analyses as above with the randomized matrix
- Compare the number of sites obtained between the RegulonDB matrix and the randomized matrix derived from it.