Babraham
Bioinformatics

Exercises:
Motif Searching

Version 2017-07

Licence

Promoter Analysis

In this exercise you are given a list of genes coming out of a previous experiment and you are going to analyse their promoter sequences to see if they contain any enrichment motifs. You will also analyse them for occurrences of known motifs.

Exercise 1a: Retrieving sequences

You should find a file called “promoter_gene_list.txt” in your data folder, this contains a list of genes of interest. To retrieve the promoter sequences for these genes we’re going to use the Ensembl biomart system.

- Open a web browser and navigate to ensembl.org/biomart/
- Choose the “Ensembl Genes” database to query and the “Mouse genes” dataset.

- Click on the “Filters” on the left of the screen, and then expand the “GENE” section on the right
- Copy and paste your gene list into the “Input external references ID list” box and tick the checkbox next to it.
- Change the ID type in the dropdown to “Gene Name(s)”. You can press the “Count” button at the top to check your genes are being found. You should see around 100 genes in your dataset

- Click on the “Attributes” link on the left and change the type of output from “Features” to “Sequences”
- Expand the “SEQUENCES” section and select “Flank (Gene)” as the sequence type.
- In the flank options tick the “Upstream flank” box and set the value to 500

- Expand the “Header Information” section
- Under “Gene Information” deselect “Ensembl Gene ID” and select “Associated Gene Name”
- Under “Transcript Information” deselect “Ensembl Transcript ID”

- Click on the “Results” button at the top of the page. You should see something similar to the text below. If yours looks different check your attributes preferences carefully.

```
>Cnn1
GTGACATTTCCTCCCTGCCCTTTACTGCACCCTTGTGAGGCAAGCACAGTTGTTAGCCCC
TCTAGAGATTTGGCAATAGGGTCCCATAGAGGGGAAGGCTCTGTGTAGAGGGTTGGTGAA
TGGAGGTCTGTCAGATCATCTTGCTATGCTAGGGGCTTGGGTGGGGGTGCAGCGGTTCTG
TCTTGGGACTGAAGAAGGAGATCAATGACCTCTGAGATAAGGAGTCTCAGAGACGGGACA
GTTGGCATGGGGAGAAGGGTGGAAAAGGGGTGGGCTGTATGAGAACCCCTCTCCCAGAAT
AAGGCATTCAGCCCCCTAGGTGGAAACAATGACACAGTCAGCTCCCAATACCAAGGCTCT
GACATCAGGAGGTGGGGGTGGCCAGAGTATGTGTGGGGTGCCACGCCTCTTGGCAGCCCC
CGTGGCCAATGGGACAGGCTTGGAAGAGCCCAGGGCTGGCTTCTGGAGCTTCAAAAACGT
GTGAGGAGGGAAGAGGGTGC
>Hpn
TCAGGGCCGAGCCCCCATCCCCCTATCAGGACTTTCCGGGCTCCAGAGGTCACTGGGCAG
GCAGCACAGCGGCTGCCAAAACAAACCACCCCCGTGACTTGTAGGAAAGTGGCTGGGGGA
AGGGCTGGGCCCCGCTAGGGACAGGAAGGGTTAGAGGCCCATAGCTGTATGAACTGGATC
AGTGGAGGGCAGGGGACTGGCATGGAAGAAAAATCGAATGAAGGACCACATTGCTACAAG
CATGGTGGACCGGGGACAGTCAGAACTTTCCACCTCATTCCTCCAGTTTGGAATCTTAGC
CTTAGGGCCTCTGAGCGCCCCCAAGTTTGTCACAGACACACCTGAGTCCCAGTGGACGTG
GGAGGCCTGGCTGGTAACTGAGCCTGCCTGGTCCAACCCCTGCGGGCCCCGCCCCATGAG
CCAGGTGGCCTCTGTGGCAGCCTGTCCTGGGGTCCGCCCTTCCCCGCCCCGTCCCCGCCT
CCCACCCAGCCCCCTCCTCC
```

- Export your file by using the option at the top of the Results page to “Export all results to File FASTA” (press the Go button to start the export). Save the results to a file called promoter_sequences.txt.

Exercise 1b: Repeatmasking

Since repetitive sequences could cause false positive discoveries within our data we are going to use the repeatmasker service to remove them from our promoter data.

- Navigate to www.repeatmasker.org
- Click on the “Repeat Masking” link in the services section on the left hand side.
- Select the file you downloaded in Exercise 1a
- Change the sensitivity to “quick”
- Change the DNA source to “Mouse”
- Press “Submit sequence” to start the search
- Once the search has completed scroll to the bottom of the results page and find the link to the masked file.
- View the masked file in a browser, then save it to a file called promoter_sequences_masked.txt

Exercise 2: Novel motif discovery

We can now use the repeat-masked sequences from exercise 1 to search for novel motifs.

- Navigate to meme-suite.org and select MEME from the Motif Discovery section on the left
- Upload your masked sequence file
- Set the number of motifs to be found to 5
- Run the search
- View the HTML output

Note that due to the time it takes for MEME to process this type of search you should not wait for it to complete but continue with exercise 4. If it still hasn’t finished when you’ve completed that then you can look at the example result we’ve provided you with instead of waiting for it to complete.

Exercise 3: Enrichment of known motifs

Rather than searching for a novel motif we can use the same repeat-masked sequences and search these against the known motif database to look for enrichment of known transcription factor binding sites.

- Navigate to meme-suite.org and select AME from the Motif Enrichment section on the left
- Upload your masked sequence file
- Set the motif database to be “JASPAR DNA” and “JASPAR CORE Vertebrates”
- Run the search
- Compare the results with what you saw in the MEME search results. Could you have got the same answer from AME without having to run MEME?

Exercise 4: Differential Motif Searching

For this exercise you will do a comparison of two sets of CpG island sequences. We have already prepared these for you and there are two files called “positive_cgi_set.txt” and “negative_cgi_set.txt”.

To start with we want to check that there is no gross compositional difference between the two sets of sequences since this would override any differences in specific motifs. To do this we will use the Compter kmer analysis tool to visualise their composition

- Navigate to https://www.bioinformatics.babraham.ac.uk/compter/

- Select the two CGI sequence files as the first and second search sequence sets.

- Press "Run compter" to start the analysis. Look at the results and see if you can see any obvious differences in composition between the two sequence sets.

We will then use DREME to compare these and compare the top hit to known motifs to see if we can find a match.

- Navigate to meme-suite.org and select DREME from the Motif Discovery section on the left
- Change the control sequence option to "User provided sequences"
- Select your positive and negative sequence sets and run the search
- Do you see a convincing hit returned? How good is the detected motif at distinguishing the two groups?
- From the top hit press the right pointing arrow under the "Submit" column and submit the motif to TomTom.
- Look at the TomTom results and see if the motif identified looks like any existing known motifs.