Exercises - Week 12

Genomics and bioinformatics

This series concerns the analysis of genomics data using multiple sequence alignments, DNA motifs and protein domains. More precisely, the goal of this exercise is to mimick the process of annotating a newly discovered gene sequence, in this case the mRNA **AY239498.1** from the Genbank database. We will search homolog candidates, reconstruct their phylogeny, and analyze the potential function of the protein by identifying conserved domains and their properties.

1 Retrieving homologs

1. Search the entry AY239498.1 in the NCBI nucleotide database [1] and save it in FASTA format as **AY239498.fasta** [click on the link *FASTA* at the top of the page and use the menu *Send*].
2. BLAST [2] the FASTA file **AY239498.fasta** against the **refseq_rna** database with parameter *Max target sequences=50* in *Algorithm parameters*.
3. Download the resulting sequences in FASTA format (in the *Descriptions* section, click on *All→Download→FASTA (Complete Sequence)*) and rename the file as **AY239498_homologs.fasta**. This will be used later.
4. Explore BLAST features:
   - Draw the corresponding tree using *Distance tree of results*. Which species are present?
   - Run BLAST against the (for example) **Nucleotide collection** database. What difference does it make?
   - Change the parameter values in *Algorithm parameters* (e.g. *Word Size*) and observe how this affects the results.

2 Multiple sequence alignment

We will next use MAFFT to perform a multiple alignment of the homologous sequences found previously. This is a required input for phylogenetic analysis. To improve visualization of results later, we choose to modify the sequence names in the FASTA file. The resulting formatted file is **AY239498_homologs_formated.fasta**.

2.1 MAFFT

1. Use MAFFT[3] with default parameters to perform a multiple sequence alignment of the file **AY239498_homologs_formated.fasta**.
2. Save the results in *CLUSTAL* format as **AY239498_homologs_formated.aln**.

[1] http://www.ncbi.nlm.nih.gov/nucleotide/
[2] http://blast.ncbi.nlm.nih.gov/Blast.cgi
[3] http://mafft.cbrc.jp/alignment/server/index.html

2.2 Jalview

1. Download and install Jalview: http://www.jalview.org/download.html
2. Launch Jalview and visualise the alignment (*File→Input Alignment→From File*).
3. Test various features of Jalview:
   - Color nucleotides (menu *Colour→Nucleotides*).
   - Display a global view (menu *View→Overview window*) and try to identify regions with the best alignment.
   - Display the logo of the consensus sequence (right-click on the consensus and select *Show Logo*).
   - Edit the alignment (remove divergent or short sequences, etc.)
4. Select the block from position **1108** to **2461**, in menu *Edit* do *Copy/Paste to new alignment*, and export the sequences in a FASTA file **conserved_block.fasta** (use *save as*).

3 Inferring the evolutionary tree

3.1 RAxML

1. Run **RAxML**[4] with the multiple alignment **AY239498_homologs_formated.aln** to infer the phylogenetic tree. *If it takes too long, go to the next section.*
2. Save the MRE consensus tree in a file named **AY239498_MRE.tree**.

3.2 Dendroscope

1. Install Dendroscope[5]
2. Visualise the consensus tree **AY239498_MRE.tree** obtained in the previous section.
3. Explore the various visualization modes, re-root the tree.

4 Identifying protein domains

The conserved block identified in the multiple sequence alignment probably contains a functionally important domain of the protein. We will search databases for proteins containing the same sequence block and see which characterized functional domains they contain.

4.1 BLASTX

1. Run BLASTX[6] with the conserved sequence block saved in section 2.2 (**conserved_block.fasta**) against the **swissprot** database with default parameters.
2. Save the top protein hits in FASTA format for the next step (**protein_sequences.fasta**).

[4] http://phylobench.vital-it.ch/raxml-bb/
[5] http://ab.inf.uni-tuebingen.de/software/dendroscope/
[6] http://blast.ncbi.nlm.nih.gov/

4.2 Prosite and InterPro searches

1. Search the BLASTX result **protein_sequences.fasta** on Prosite[7].
2. Search the first protein sequence (**P22555**) on InterProScan[8].
3. Which domain is found by these tools?
4. Are there domains described only in one source? Note the differences between the descriptors of the same domains.

5 DNA motif discovery

We have identified our sequence **AY239498.1** as coding for a **c-Myc** homolog and containing a **bHLH** DNA-binding domain. To determine its targets, we might perform a ChIP-seq experiment. Today we will use the data from a similar experiment on the Human c-Myc (from the ENCODE project http://www.genome.ucsc.edu/ENCODE/). Sequencing data have been mapped to the Human genome and, after peak calling, a list of chromosomal coordinates containing the genomic regions bound by c-Myc are obtained. We will use them to identify the motif specifically recognized by c-Myc's bHLH domain.

5.1 Retrieve sequences by their genome coordinates (with UCSC)

1. Download the BED file **posPeakMyc.bed** and go to the UCSC genome browser[9], click on *Genome* (top left) and select the Human genome assembly hg19. Click on *manage custom tracks*, then on *add custom tracks* and load the bed file.
2. Click on *go to table browser* and select *sequence* in the list of output format. Enter **posPeakMyc.fasta** in field output file. Click on *get output*.
3. In this page, select the *All upper case* option and click on *get sequence* to save the sequences into file **posPeakMyc.fasta**.
4. In the result file, you can check that the first line looks like:

```
>hg19_ct_UserTrack_3545_(null) range=chr1:8939139-8939538 5'pad=0
3'pad=0 strand=+ repeatMasking=none
```

[7] http://prosite.expasy.org/scanprosite/
[8] http://www.ebi.ac.uk/Tools/pfa/iprscan/
[9] http://genome.ucsc.edu/

5.2 Mask low complexity and repeated elements

Motif search can be hindered by repetitive sequences included in the binding regions. We will use a tool to identify and remove potential repetitive sequences.

1. Go to repeatMasker[10]. The interest of this tool is to mask regions of low complexity to facilitate the identification of relevant motifs.
2. Choose *RepeatMasking* service.
3. Select your file **posPeakMyc.fasta** and submit the job with default parameters.
4. Check which elements have been masked in *xxx.out.html* and save the masked file *xxx.masked*.
5. Which difference can you see with the input file ?

5.3 Motif discovery with MEME-ChIP

1. Go to MEME-ChIP[11], and fill the form with the masked sequences **posPeakMyc_masked.fasta** previously obtained.

   Note: In MEME options, we recommend to choose one occurrence per sequence, a maximum width of 15, and a maximum number of motif of 5.
2. Have a look at the MEME and TOMTOM html results.
3. Which transcription factor corresponds to the main motif?

[10] http://www.repeatmasker.org/
[11] http://meme.sdsc.edu/meme/cgi-bin/meme-chip.cgi