**Data School Synthesis Project**

**Project aim:**

build a command-line python program count the number of transcription factor binding sites (TFBs) from differentially expressed genes.

**Background:**

A typical transcription factor binds to DNA at a certain target sequence (or motif). Once it's bound, the transcription factor makes it either harder or easier for RNA polymerase to bind to the promoter of the gene, and consequently regulates the amount of messenger RNA (mRNA) produced by the gene. Some transcription factors **activate transcription**, while others **repress transcription**.

Transcription factor binding sites(TFBS) are often located in the 5’-upstream region of target genes to modulate the rate of gene transcription. DNA binding sites can be thus defined as short DNA sequences (typically 4 to 30 base pairs long) that are specifically bound by one or more [DNA-binding proteins](https://en.wikipedia.org/wiki/DNA-binding_protein) or protein complexes.

**To find TFBS my function should complete the following tasks:**

1. **Find the location (start and end coordinates) of differentially expressed genes (DE genes):**

a. **Input:** a list of differentially expressed genes (gene symbol and Entrez ID) and an annotation file (gff file)

b.  **Output:** gene scaffold, and gene coordinates (start and end location of gene CDS region)

*Signature:* **get\_gene\_coordinates**: string, string à string, integer, integer

*Purpose:* give the name of the scaffold and the coordinates of a differentially expressed gene

*Stub:* def **get\_gene\_coordinates**(DEgenes, annotation): name, start coordinates, end coordinates.

*Example:*

get\_gene\_coordinates((‘gene1, ‘gene2’, ‘gene3’), salmon.gff3): chromosome: chr1, start: 120, end: 150.

1. **Extract the upstream sequences from target genes (DE genes):**

a. Input: Gene location (name of scaffold, start and end coordinates)

b. Output: Sequences of a certain number of base pairs (e.g. 5000) located on the 5’-upstream region of target genes.

*Signature:* **extract\_sequences**: string, integer, integer, string à string

*Purpose:* Extract a given number of base pairs upstream from a target gene

*Stub:* def **extract\_sequences**(gene\_scaffold\_name, start\_coordinate, number\_basepairs, fasta\_file): gene\_names, gene\_sequences (in fasta format)

Example:

extract\_sequences(NC\_027306.1, 39830756, 5000, salmon.fasta):

>gene2

CGCACAAACAGAGATTTGATCTGTGAGTTCGGTGCTAGTCAGTTTGCACTAAGCCACCTACTGCAAACAAGGCTGCGTCTAAACATATACCTCTGCAACTTGTTGAATCGGCTTAATAATATAGAAACTTTGTCAGGGAAAAGTTGAGGAACTTTTGAGTGGGACTGGATACGACCTTGGACTTATTTATGACAACATTCTAG ….

1. **Find TFBSs or motifs in the extracted sequences:** this is when it gets a bit hairy…

a. **Input:** A list of upstream sequences from target genes, and motifs (sequences and matrices) for the most common transcription factors (I could also find these in my list of DE genes, but for now I will only use a few well-known ones present in my DE gene list)

b. **Output:** Number of times these motifs appear in the genome before each DE gene.

*Signature:* **count\_tfbs**: string, string, integer à integer

*Purpose:* count the number of times a certain motif is present in a gene sequence

*Stub:* def **count\_tfbs**(gene\_sequences, motifs\_for\_tfbs, motif\_matrices): counts

*Example:*

**count\_tfbs(**CGCACAAACAGAGATTTGATCTGTGAGTTCGGTGCTAGTCAGTTTGCACTAAGCCACCTACTGCAAACAAGGCTGCGTCTAAACATATACCTCTGCAACTTGTTGAATCGGCTTAATAATATAGAAACTTTGTCAGGGAAAAGTTGAGGAACTTTTGAGTGGGACTGGATACGACCTTGGACTTATTTATGACAACATTCTAGt, TCTAA, ATGD\_matrix): gene2 : 1