WBO
Assignment 2
Igor Filipiuk

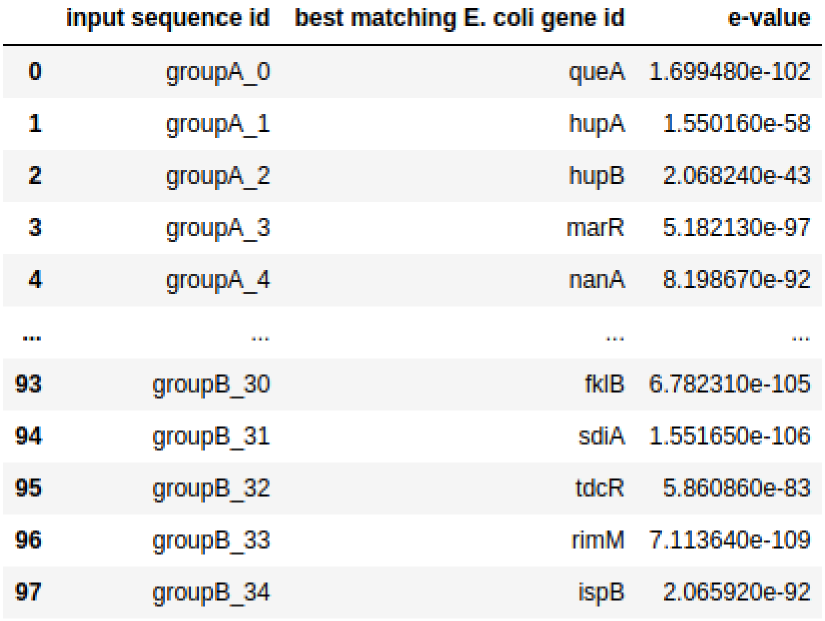
Report

Script Assignment.py performs a sequential processing and analysis of the data, provided by the lecturer.

TASK 1

Starting off with the fasta file ‘*genes_e_coli_new.fa_*’ script creates a DataBase using NcbimakeblastdbCommandline from Biopython library. As the next step running local blastn procedure with ‘*protein_fragments.fa_*’ as input file was applied (using NcbitblastnCommandline from Biopython library). This step is necessary in order to fit unknown protein-transcripts from ‘*protein_fragments.fa*’ to well-defined genes in ‘*genes_e_coli_new.fa_*’.

As the result of above procedure xml file was created from which best alignments were extracted. Best fits were identified with the lowest e-value. Result from this step was saved in csv format into ‘*outputs/output_1/bestMatches.csv*’ file containing 98 records of input sequence id, best matching E. coli gene id and the associated e-value.



	input sequence id	best matching E. coli gene id	e-value
0	groupA_0	queA	1.699480e-102
1	groupA_1	hupA	1.550160e-58
2	groupA_2	hupB	2.068240e-43
3	groupA_3	marR	5.182130e-97
4	groupA_4	nanA	8.198670e-92
...	...	...	...
93	groupB_30	fklB	6.782310e-105
94	groupB_31	sdiA	1.551650e-106
95	groupB_32	tdcR	5.860860e-83
96	groupB_33	rimM	7.113640e-109
97	groupB_34	ispB	2.065920e-92

TASK 2

Following step was to match promoters against output of the previous step, and identify which promoters belong to groupA and which to groupB (results of this intermediary step were saved in fasta format in ‘*outputs/output_2/*’ directory as ‘*promoters_groupA.fa*’ and ‘*promoters_groupB.fa*’. Further, using function *consensus(fastaFile, motifLength, returnCount)* there were overall 20 motifs with best information content obtained.

Parameters used for above function:
fastaFile = ‘*promoters_groupA.fa*’ lub ‘*promoters_groupB.fa*’

motifLength = 15
returnCount = 10

Functions were run separately on 63 sequences from groupA and 35 sequences from groupB. Result of this step are two sets of motifs of 10 sequences each (20 sequences overall). Then motifs were saved in 'pfm' format in '*outputs/output_2/motifsA_pfm'* directory for groupA and '*outputs/output_2/motifsB_pfm*' for groupB.

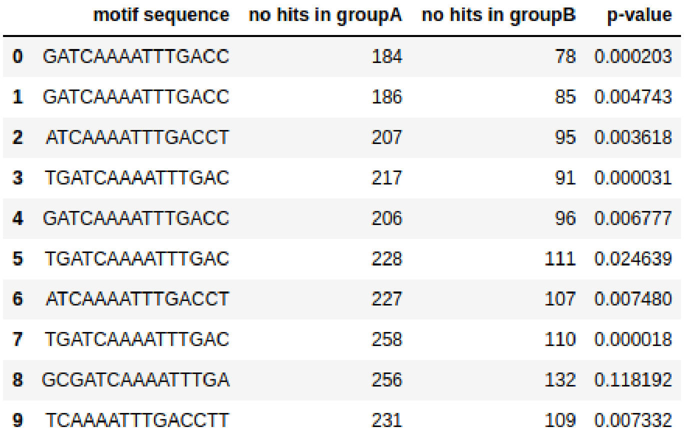
As an example, first motifs (with the highest information content) was written thusly:

```
-3.0660891904577725 1.393342428179525 -1.0660891904577725 -1.2587342684001683
1.6618312641054267 1.4257639058719025 1.63435052768332 1.5188733102633836 -
3.0660891904577725 -0.89616418901546 -1.4811266897366162 -1.4811266897366162
1.6618312641054267 -2.0660891904577725 -1.74416109557041
-1.4811266897366162 -1.2587342684001683 -3.0660891904577725 1.4574727655992403 -
2.4811266897366164 -1.74416109557041 -1.4811266897366162 -1.74416109557041 -
0.7441610955704102 -1.2587342684001683 -1.0660891904577725 -1.74416109557041 -
3.0660891904577725 1.63435052768332 1.488499661219865
1.6887983117056962 -1.4811266897366162 -1.2587342684001683 -1.2587342684001683 -
1.74416109557041 -1.74416109557041 -2.0660891904577725 -1.74416109557041 -
0.89616418901546 -1.2587342684001683 -1.74416109557041 1.6618312641054267 -
1.0660891904577725 -1.4811266897366162 -1.0660891904577725
-1.74416109557041 -0.7441610955704102 1.5777669993169523 -1.2587342684001683 -
1.4811266897366162 -0.4811266897366163 -1.74416109557041 -0.89616418901546
1.4574727655992403 1.393342428179525 1.5188733102633836 -2.4811266897366164 -
2.0660891904577725 -1.74416109557041 -1.2587342684001683
```

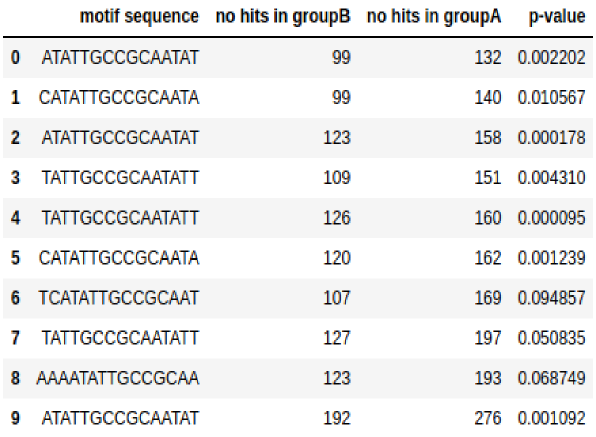
TASK 3

Last step of the analysis was identify motifs obtained from previous step (TASK 2), and test if motifs that are from a given group are significantly enriched against the other group. Function *testBinomial(motif, group)* performs binomial tests and returns p-value for a given motif of a given group. During the run function needs to compute three parameters:
x – get number of positions in a given set of promoter sequences that have a log-odds score higher than 0 for a given motif
n – number of possible position that motif frame will slide over sequences in a given group (i.e. 63 x (100 - 15) for the groupA and 35 x (100 – 15) for the groupB)
p – a threshold against which we check if motif is significantly enriched. In case of motif from groupA we would comput p as (no of log-odds score > 0 for all position for all sequences from groupB)/(35 sequences in group B x (length(sequences from group B) – length(motif)).

For the above described procedure results were obtained (motifsA to the left, motifsB to the right):



	motif sequence	no hits in groupA	no hits in groupB	p-value
0	GATCAAAATTTGACC	184	78	0.000203
1	GATCAAAATTTGACC	186	85	0.004743
2	ATCAAAATTTGACCT	207	95	0.003618
3	TGATCAAAATTTGAC	217	91	0.000031
4	GATCAAAATTTGACC	206	96	0.006777
5	TGATCAAAATTTGAC	228	111	0.024639
6	ATCAAAATTTGACCT	227	107	0.007480
7	TGATCAAAATTTGAC	258	110	0.000018
8	GCGATCAAAATTTGA	256	132	0.118192
9	TCAAAATTTGACCTT	231	109	0.007332



	motif sequence	no hits in groupB	no hits in groupA	p-value
0	ATATTGCCGCAATAT	99	132	0.002202
1	CATATTGCCGCAATA	99	140	0.010567
2	ATATTGCCGCAATAT	123	158	0.000178
3	TATTGCCGCAATATT	109	151	0.004310
4	TATTGCCGCAATATT	126	160	0.000095
5	CATATTGCCGCAATA	120	162	0.001239
6	TCATATTGCCGCAAT	107	169	0.094857
7	TATTGCCGCAATATT	127	197	0.050835
8	AAAATATTGCCGCAA	123	193	0.068749
9	ATATTGCCGCAATAT	192	276	0.001092

Based on the p-values, assuming alpha = 0,05, we can conclude that for motifs of groupA all sequences are significantly enriched except 8-th sequence (GCGATCAAAATTTGA). For motifs from groupB there are three sequences that we cannot reject hypothesis 0 that motifs are not significantly enriched (TCATATTGCCGCAAT, TATTGCCGCAATATT and AAAATATTGCCGCAA).