HW3 – Bioinformatics – 236523

Anna Romanov 321340580 [annarom@campus.technion.ac.il](mailto:annarom@campus.technion.ac.il)

Maxim Kolchinsky 320983216 [kolchinsky@campus.technion.ac.il](mailto:kolchinsky@campus.technion.ac.il)

# Question 3

1. Since a sequence may contain more than one occurrence we will have to uniq filter our results.

I took the motif occurrences file motif1\_summary.txt and ran the following linux command to get the amount of different sequences:



So we have 702 sequences that contain a k-mer associated with the motif.

1. PSSM:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PSSM: |  |  |  |  |  |  |  |  |
| ----- |  |  |  |  |  |  |  |  |
| A | 0 | 0 | 0 | 1 | 0.52 | 1 | 0 | 0.66 |
| T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| U | 1 | 0 | 1 | 0 | 0.48 | 0 | 1 | 0.34 |

What can you say about the motif?

The motif has 20 matches over 3 target databases

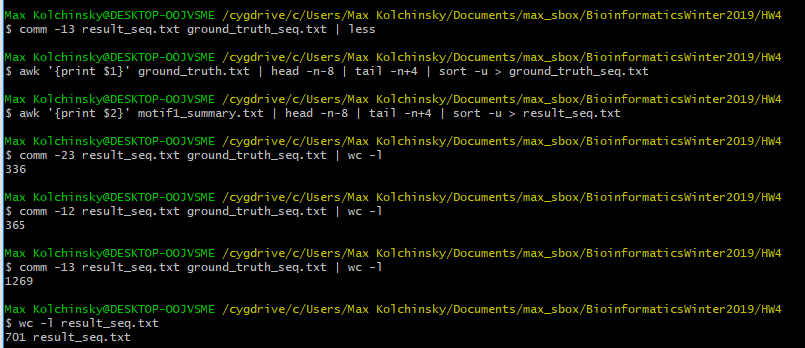
[FOXB1\_DBD\_3](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_1),  [MA0845.1 (FOXB1)](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_11),  [FOXB1\_DBD\_2](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_0),  [UP00094\_2 (Zfp128\_secondary)](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_14),  [FOXL1\_full\_2](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_9),  [FOXB1\_full](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_2),  [FOXC1\_DBD\_3](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_4),  [MA0032.2 (FOXC1)](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_10),  [UP00250\_1 (Irx5\_2385.1)](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_19),  [UP00223\_2 (Irx3\_2226.1)](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_17),  [FOXC1\_DBD\_2](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_3),  [FOXC2\_DBD\_1](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_5),  [UP00236\_1 (Irx2\_0900.3)](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_18),  [UP00150\_1 (Irx6\_2623.2)](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_15),  [FOXD3\_DBD\_1](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_7),  [UP00194\_1 (Irx4\_2242.3)](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_16),  [MA1103.1 (FOXK2)](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_12),  [FOXD3\_DBD\_2](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_8),  [UP00008\_2 (Six6\_secondary)](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_13),  [FOXD2\_DBD\_1](http://meme-suite.org/opal-jobs/appTOMTOM_5.0.3_1547199588627-577132523/tomtom.html#match_0_6)

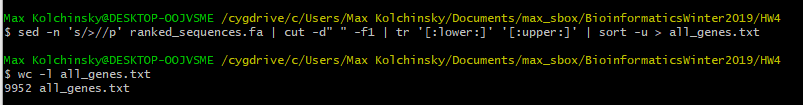
The Alt. ID is TGTAWATW

What possible experiment generated the ranked list of sequences?

I have no idea

1. The following commands gave me the results:





So:

Left\_only = 336

Right\_only = 1269

Both = 365

TotalGenesLeft = 701

TP = Both = 365

FP = TotalGenesLeft – Both = 701 – 365 = 336

FN = Right\_only = 1269

TN = TotalGenes – FP = 9952 – 1269 = 8636

From which I can conclude that my confusion matrix will be like so:

|  |  |  |
| --- | --- | --- |
|  | Predicted:  No | Predicted:  Yes |
| Actual:  No | 8636 | 336 |
| Actual:  Yes | 1269 | 365 |

1. Calculation:
2. If the tool is very complex it is also very slow. Since we already have 96.2% specifity if we value specifity a lot more than sensitivity and we also value fast runtime we will prefer to run the motif search to get almost 100% results a lot faster.