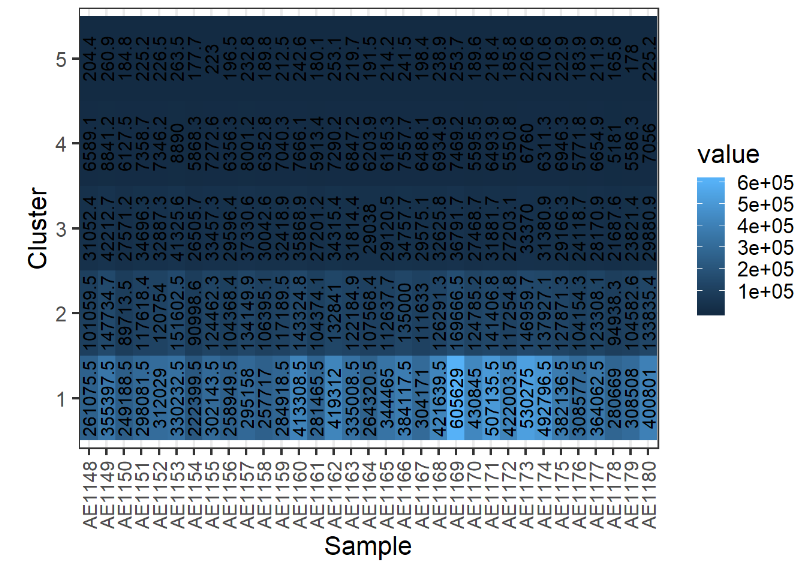
HW3 – Bioinformatics – 236523

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 1

2. c. ii.



iii.

1 2 3 4 5

2 21 169 2054 56101

iv. We see that the number of genes in each cluster varies significantly – from 2 genes in cluster number 1, to 56,101 genes in cluster number 5. This is a shortcoming since most of the genes appear in a single cluster which is not very informative for expression analysis.

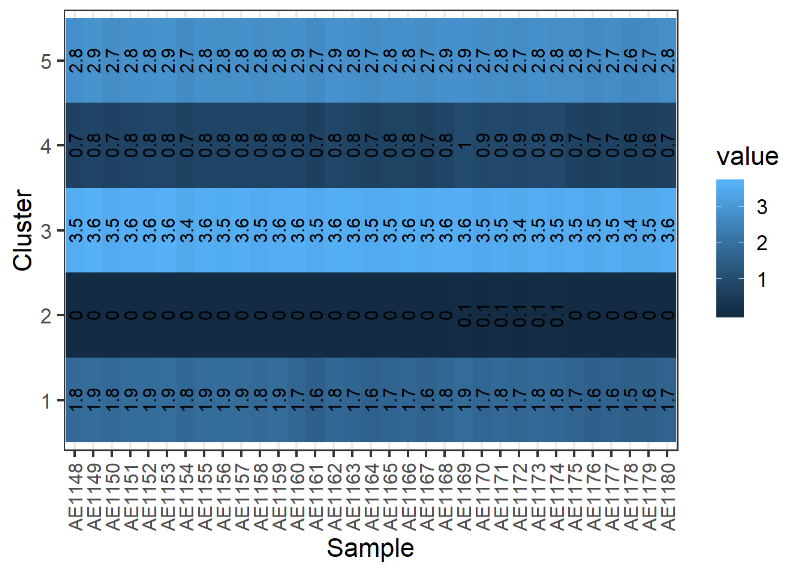
Another problem is that for clusters 1 and 2, the range of values that belong to the same cluster is big, therefore in the heatmap it looks like cluster 1 is not homogeneous. There are two possibilities: either the big difference in values (for example there is a value of 605,629 together with 222,399 in cluster 1) has biological meaning and those samples being in the same cluster is an error; or in the range of such high values, big differences are less significant biologically (a difference of ~200,000 is still in the same order of magnitude), so the clustering itself is correct but the heatmap suggests that there might be significant differences in examples in the same cluster. In the second case, a better way of representing the results could be used (as in the log transformation below).

A possible reason for these problems could be that the expression values are within a very large range, being more dense in the smaller values (we see that in cluster 5, the maximum difference is around ~100, while in cluster 1 it is around ~300,000). This in turn is caused by large differences in counts of different genes as appear in ‘rawcounts.csv’.  
This way, to make distinction between clusters the algorithm divides intro groups roughly by order of magnitude, but it might be that after a certain threshold big differences are not significant, while in the small values the algorithm should be more sensitive to small fluctuations.

v. After applying log transformation:

1 2 3 4 5

3986 37652 4922 6081 5706



We see that cluster sizes are now more balanced, and also the expression values in each cluster are closer to each other (in contrast to the previous configuration), making clusters more homogenic. This can be explained by the fact that expression values are now in smaller range – from 0 up to 4, so more of the larger values are found in the same cluster (since big differences after log transformation become much smaller and examples which previously were considered ‘too different’ are now similar).

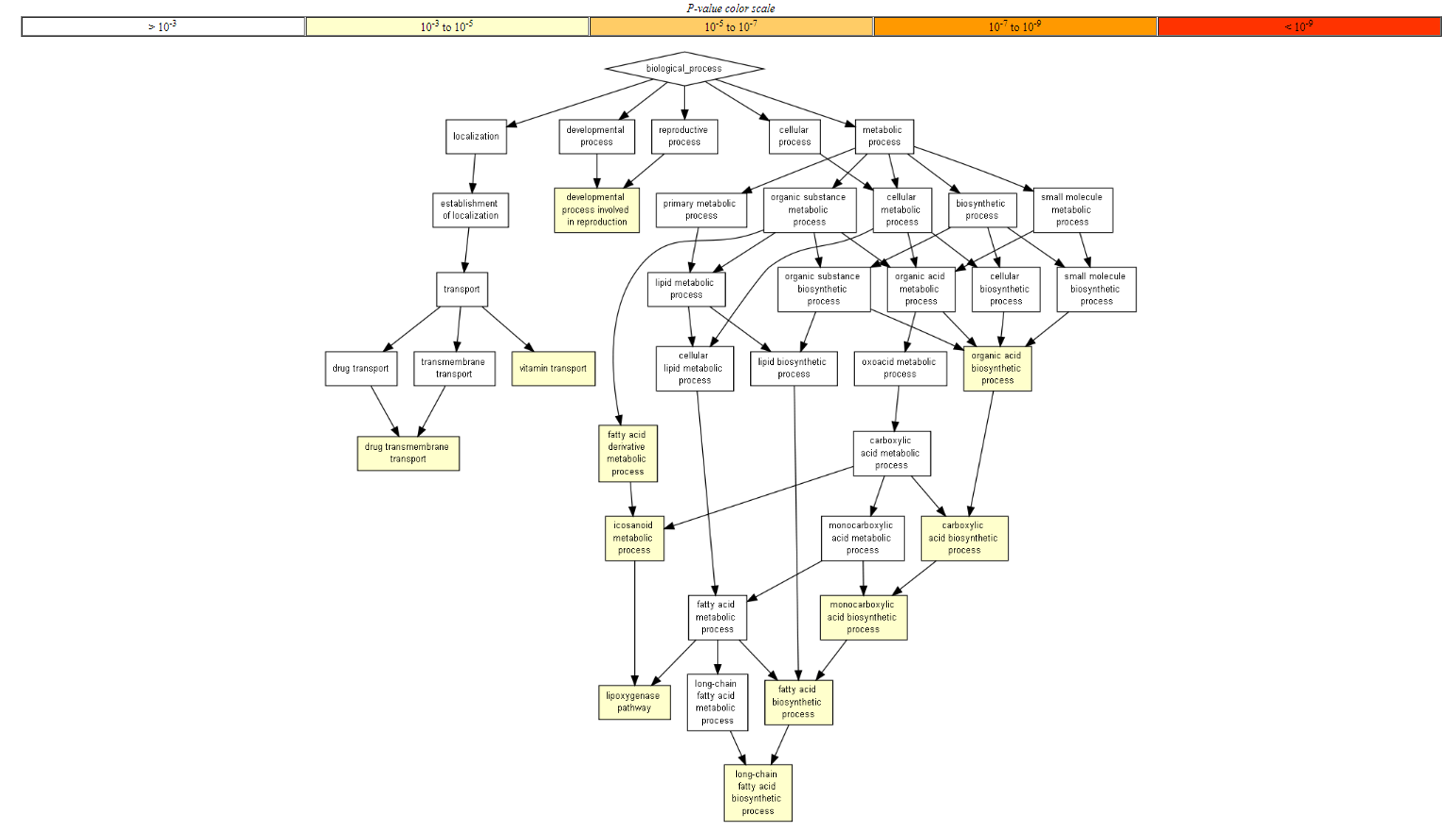
Still, this method is not ideal since many values after the log transformation become 0, which cancels the differences in values which might be important for the analysis.

vi. A way to improve?

vii.

# Question 2

1. C. The results retrieved from Gorilla:



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **GO term** | **Description** | [**P-value**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/n5328vda/GOResults.html#p_value_info) | [**FDR q-value**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/n5328vda/GOResults.html#fdr_info) | [**Enrichment (N, B, n, b)**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/n5328vda/GOResults.html#enrich_info) | [**Genes**](http://cbl-gorilla.cs.technion.ac.il/GOrilla/n5328vda/GOResults.html#genes_info) |
| [GO:0072330](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0072330&view=details) | monocarboxylic acid biosynthetic process | 2.69E-5 | 1.21E-1 | 4.95 (395,9,71,8) | [[-] Hide genes](javascript:toggle('elements_GO:0072330'))  ASNS - asparagine synthetase (glutamine-hydrolyzing) SCD - stearoyl-coa desaturase (delta-9-desaturase) ALOX5 - arachidonate 5-lipoxygenase EPHX2 - epoxide hydrolase 2, cytoplasmic GPX4 - glutathione peroxidase 4 ABHD3 - abhydrolase domain containing 3 ALDOC - aldolase c, fructose-bisphosphate FADS3 - fatty acid desaturase 3 |
| [GO:1901568](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:1901568&view=details) | fatty acid derivative metabolic process | 1.77E-4 | 3.98E-1 | 10.39 (395,10,19,5) | [[-] Hide genes](javascript:toggle('elements_GO:1901568'))  ABCC1 - atp-binding cassette, sub-family c (cftr/mrp), member 1 ALOX5 - arachidonate 5-lipoxygenase EPHX2 - epoxide hydrolase 2, cytoplasmic GPX4 - glutathione peroxidase 4 OXCT1 - 3-oxoacid coa transferase 1 |
| [GO:0019372](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0019372&view=details) | lipoxygenase pathway | 1.93E-4 | 2.89E-1 | 65.83 (395,2,6,2) | [[-] Hide genes](javascript:toggle('elements_GO:0019372'))  ALOX5 - arachidonate 5-lipoxygenase GPX4 - glutathione peroxidase 4 |
| [GO:0006855](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0006855&view=details) | drug transmembrane transport | 2.22E-4 | 2.49E-1 | 87.78 (395,3,3,2) | [[-] Hide genes](javascript:toggle('elements_GO:0006855'))  ABCC1 - atp-binding cassette, sub-family c (cftr/mrp), member 1 STRA6 - stimulated by retinoic acid 6 |
| [GO:0051180](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0051180&view=details) | vitamin transport | 2.22E-4 | 2E-1 | 87.78 (395,3,3,2) | [[-] Hide genes](javascript:toggle('elements_GO:0051180'))  ABCC1 - atp-binding cassette, sub-family c (cftr/mrp), member 1 STRA6 - stimulated by retinoic acid 6 |
| [GO:0046394](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0046394&view=details) | carboxylic acid biosynthetic process | 3.68E-4 | 2.76E-1 | 3.85 (395,13,71,9) | [[-] Hide genes](javascript:toggle('elements_GO:0046394'))  ASNS - asparagine synthetase (glutamine-hydrolyzing) SCD - stearoyl-coa desaturase (delta-9-desaturase) ALOX5 - arachidonate 5-lipoxygenase EPHX2 - epoxide hydrolase 2, cytoplasmic GPX4 - glutathione peroxidase 4 ABHD3 - abhydrolase domain containing 3 PSAT1 - phosphoserine aminotransferase 1 ALDOC - aldolase c, fructose-bisphosphate FADS3 - fatty acid desaturase 3 |
| [GO:0042759](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0042759&view=details) | long-chain fatty acid biosynthetic process | 4.77E-4 | 3.06E-1 | 18.52 (395,4,16,3) | [[-] Hide genes](javascript:toggle('elements_GO:0042759'))  ALOX5 - arachidonate 5-lipoxygenase EPHX2 - epoxide hydrolase 2, cytoplasmic GPX4 - glutathione peroxidase 4 |
| [GO:0006690](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0006690&view=details) | icosanoid metabolic process | 5.13E-4 | 2.88E-1 | 24.69 (395,8,6,3) | [[-] Hide genes](javascript:toggle('elements_GO:0006690'))  ABCC1 - atp-binding cassette, sub-family c (cftr/mrp), member 1 ALOX5 - arachidonate 5-lipoxygenase GPX4 - glutathione peroxidase 4 |
| [GO:0003006](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0003006&view=details) | developmental process involved in reproduction | 5.49E-4 | 2.74E-1 | 3.28 (395,14,86,10) | [[-] Hide genes](javascript:toggle('elements_GO:0003006'))  STRA6 - stimulated by retinoic acid 6 RBP4 - retinol binding protein 4, plasma YBX2 - y box binding protein 2 COL9A3 - collagen, type ix, alpha 3 CENPI - centromere protein i MME - membrane metallo-endopeptidase ZMYND15 - zinc finger, mynd-type containing 15 TYRO3 - tyro3 protein tyrosine kinase HSD17B3 - hydroxysteroid (17-beta) dehydrogenase 3 RRM1 - ribonucleotide reductase m1 |
| [GO:0006633](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0006633&view=details) | fatty acid biosynthetic process | 6.48E-4 | 2.91E-1 | 4.77 (395,7,71,6) | [[-] Hide genes](javascript:toggle('elements_GO:0006633'))  SCD - stearoyl-coa desaturase (delta-9-desaturase) ALOX5 - arachidonate 5-lipoxygenase EPHX2 - epoxide hydrolase 2, cytoplasmic GPX4 - glutathione peroxidase 4 ABHD3 - abhydrolase domain containing 3 FADS3 - fatty acid desaturase 3 |
| [GO:0016053](http://www.godatabase.org/cgi-bin/amigo/go.cgi?query=GO:0016053&view=details) | organic acid biosynthetic process | 8.68E-4 | 3.54E-1 | 3.58 (395,14,71,9) | [[-] Hide genes](javascript:toggle('elements_GO:0016053'))  ASNS - asparagine synthetase (glutamine-hydrolyzing) SCD - stearoyl-coa desaturase (delta-9-desaturase) ALOX5 - arachidonate 5-lipoxygenase EPHX2 - epoxide hydrolase 2, cytoplasmic GPX4 - glutathione peroxidase 4 ABHD3 - abhydrolase domain containing 3 PSAT1 - phosphoserine aminotransferase 1 ALDOC - aldolase c, fructose-bisphosphate FADS3 - fatty acid desaturase 3 |

D. We see that there are enriched GO terms (the ones in the list above). The biological insight we gain is that differentially the expressed genes found in HW3 are part of certain biological processes as listed above. In particular, the lowest P-value (most significant) belongs to the GO term GO:0072330: monocarboxylic acid biosynthetic process. From looking at the diagram, we see that many of the differentially expressed terms are related to carboxylic acid biosynthetic process, being in it’s subtree.

2. a. We used the ‘rawcounts.csv’ file to obtain the background list, and “DE\_results.csv” to obtain the differentially expressed genes (similar to previous section, without sorting by P-value).

c. The results of running Gorilla:

# 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 |

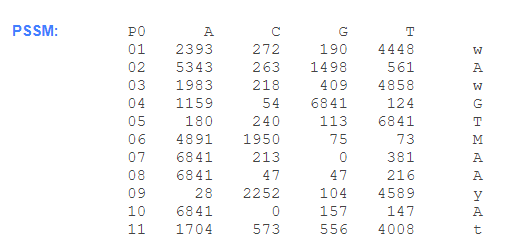
What can you say about the motif?

The motif is FOXBD1 taken from the Homo Sapiens Database.

The length of the motif is 11.



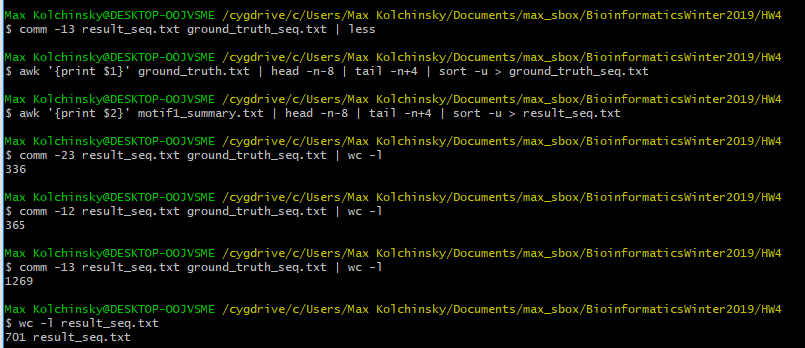
The PSSM of the motif is:

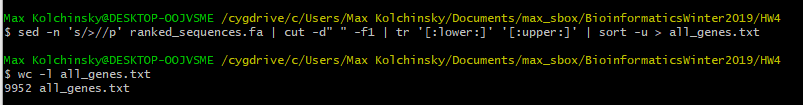


What possible experiment generated the ranked list of sequences?

The list of sequences is a fasta file. An example for a possible experiment that made this file could be PCA on a patient being treated for cancer.

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.