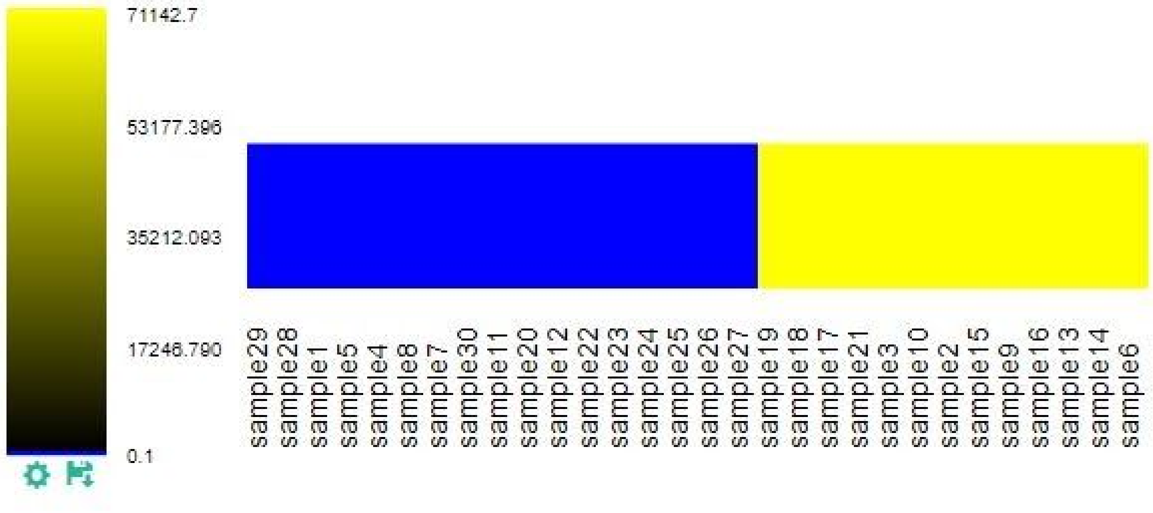
Osman Emre Bilici
2171353

Assignment #4: Analysis of Microarray Data

In this assignment, I used TM4 MEV tool to distinguish the clusters as mentioned in the homework text.

For first goal, I used k-means with 2 cluster.



The result as you can see,
sample29, sample28, sample1, sample5, sample4, sample8, sample7, sample30, sample11, sample20, sample12, sample22, sample23, sample24, sample25, sample26, sample27 for tissue1 cluster.
sample19, sample18, sample17, sample21, sample3, sample10, sample2, sample15, sample9, sample16, sample13, sample14, sample6 for tissue2 cluster.

I assumed that the data don’t need normalization. If the data needs normalization, I used deseq normalization technique, tissue2 cluster is
sample19, sample18, sample17, sample21, sample3, sample10, sample2, sample24, sample13, sample14.
Tissue1 cluster is
sample29, sample28, sample1, sample5, sample4, sample9, sample8, sample7, sample6, sample30, sample11, sample20, sample12, sample22, sample23, sample15, sample25, sample16, sample26, sample27.

But when I used the MeV tool new normalization values couldn’t be downloaded.

I used the clustering that I mentioned first.

For second goal, I wrote a script. In this script, for each gene, values of all samples for tissue1 and tissue2 is added for both seperately. Then differences are calculated with tissue1- tissue2. And max 10 values and min 10 values are detected. The script can be seen at the end of the homework.

10 genes that are highly expressed in the tissue1 and not in the tissue2:

1. '207430_s_at’,
2. '210297_s_at',
3. '205623_at',
4. '204151_x_at',
5. '214303_x_at',
6. '214385_s_at',
7. '209699_x_at',
8. '201884_at',
9. '201891_s_at',
10. '204351_at'

10 genes that are highly expressed in tissue2 and not in the tissue1:

1. '205725_at',
2. '220542_s_at',
3. '204892_x_at',
4. '203021_at',
5. '210646_x_at',
6. '206559_x_at',
7. '201257_x_at',
8. '212790_x_at',
9. '213477_x_at',
10. '215963_x_at'

I assumed tissue2 is healthy tissues and tissue1 is diseased tissues.

My script:

```
import csv
tissue1 = [29,28,1,5,4,8,7,30,11,20,12,22,23,24,25,26,27]
tissue2 = [19,18,17,21,3,10,2,15,9,16,13,14,6]
difference = 0
differences = []
with open("hw4data.txt") as tsv:
  for line in csv.reader(tsv, delimiter="\t"):
     tissue1Total = 0
     tissue2Total = 0
     if line[0] == "ID":
        continue
     for s in range(1,len(line)):
        if s in tissue1:
           tissue1Total += float(line[s])
        else:
           tissue2Total += float(line[s])
     difference = tissue1Total - tissue2Total
     differences.append({"id": line[0], "dif": difference})

topTenMax = []
topTenMin = []
for i in range(0, 10):
  topTenMax.append(max(differences, key=lambda x: x['dif']))
  topTenMin.append(min(differences, key=lambda x: x['dif']))
  differences.pop(differences.index(min(differences, key=lambda x: x['dif'])))
  differences.pop(differences.index(max(differences, key=lambda x: x['dif'])))


print(topTenMax)
print(topTenMin)
```

```
topTenMax: [{'id': '207430_s_at', 'dif': 114173.69999999997}, {'id': '210297_s_at', 'dif':
79187.40000000001}, {'id': '205623_at', 'dif': 70827.70000000001}, {'id': '204151_x_at', 'dif':
67523.09999999998}, {'id': '214303_x_at', 'dif': 66676.2}, {'id': '214385_s_at', 'dif':
60824.29999999999}, {'id': '209699_x_at', 'dif': 56351.19999999998}, {'id': '201884_at', 'dif':
48080.5}, {'id': '201891_s_at', 'dif': 47414.80000000002}, {'id': '204351_at', 'dif': 47322.8}]

topTenMin: [{'id': '205725_at', 'dif': -361614.0}, {'id': '220542_s_at', 'dif':
-237036.39999999997}, {'id': '204892_x_at', 'dif': -166822.90000000008}, {'id': '203021_at',
'dif': -96562.29999999999}, {'id': '210646_x_at', 'dif': -73386.29999999996}, {'id':
'206559_x_at', 'dif': -50763.49999999994}, {'id': '201257_x_at', 'dif': -42029.20000000001},
{'id': '212790_x_at', 'dif': -41285.30000000002}, {'id': '213477_x_at', 'dif':
-34590.00000000003}, {'id': '215963_x_at', 'dif': -31565.600000000006}]
```