Jacek Radajewski
Student number: 43612772
SCIE2100 Prac 9

Exercise 1

The null hypothesis (from the question) is: X is indicative of a function known as ‘chaperone regulator activity’ known as GO:0030188 by the Gene Ontology.

Value chosen for α was 0.05 as it quite standard in statistics. P-value was calculated (please see the code and execution) to be 0.23 which was higher than than α. In conclusion the null hypothesis has not been rejected.

Code

```
'''
Created on 27/05/2014

@author: s4361277
'''
import stats
significance_level = 0.05
positives = set({"YPL106C", "YOL081W", "YOR027W", "YOR299W", "YNL006W", \
    free
              "YNL007C", "YLL039C", "YLR216C"})
has_property = set({"YER048C", "YIL016W", "YLR090W", "YOR027W", \
    "YMR161W", "YNL064C", "YNL281W", "YDR214W", "YPL106C", "YNL007C", \
    "YNL227C"})

a = len(positives.intersection(has_property))
b = 8 - a
c = 2
d = 14 - c

p_value = stats.getFETpval(a, b, c, d, left=False)

print "P value=", p_value

if p_value < significance_level:
    print "reject null hypothesis"
else:
    print "do not reject null hypothesis"
```

Execution

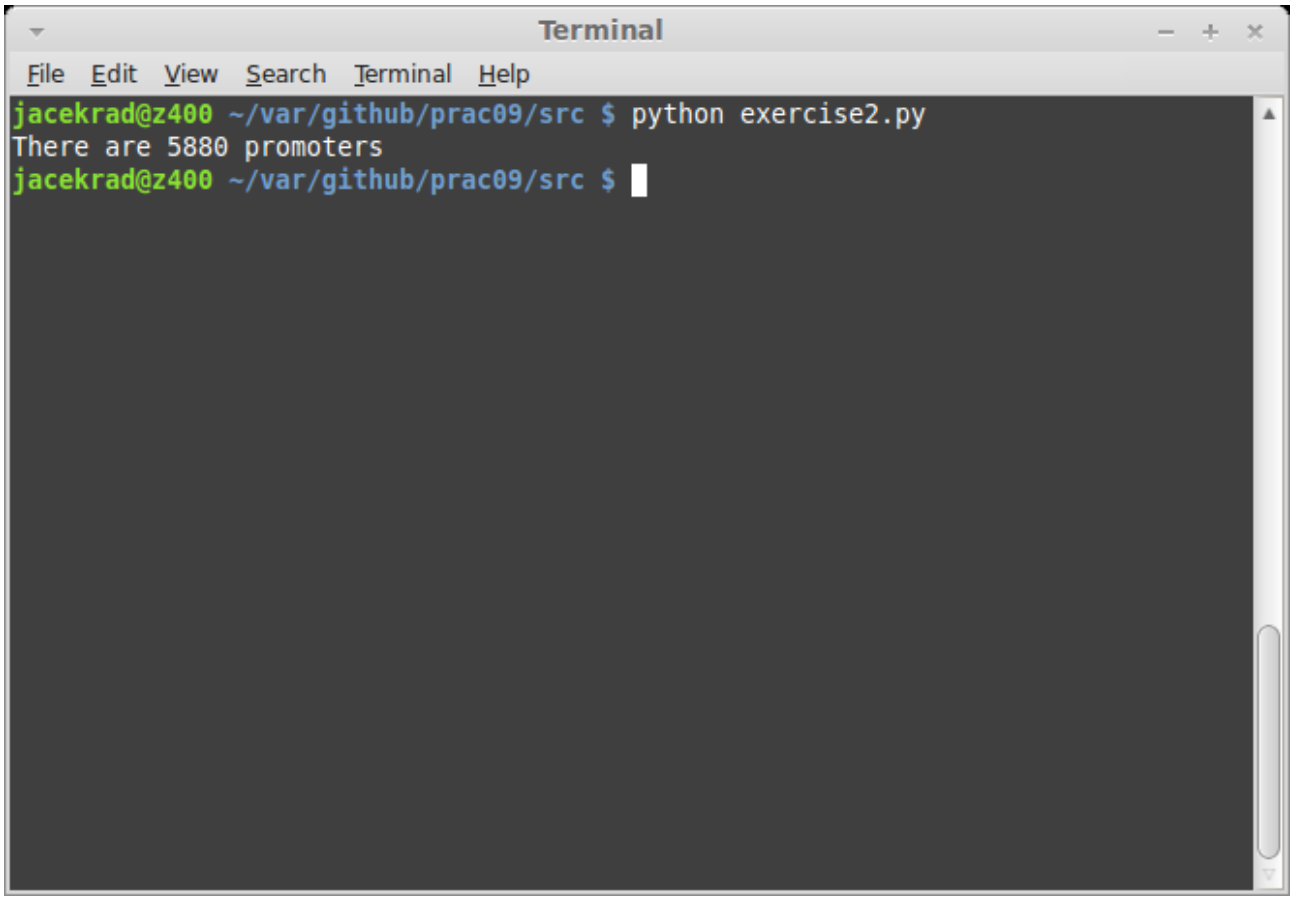
```
P value= 0.232854864434
do not reject null hypothesis
```

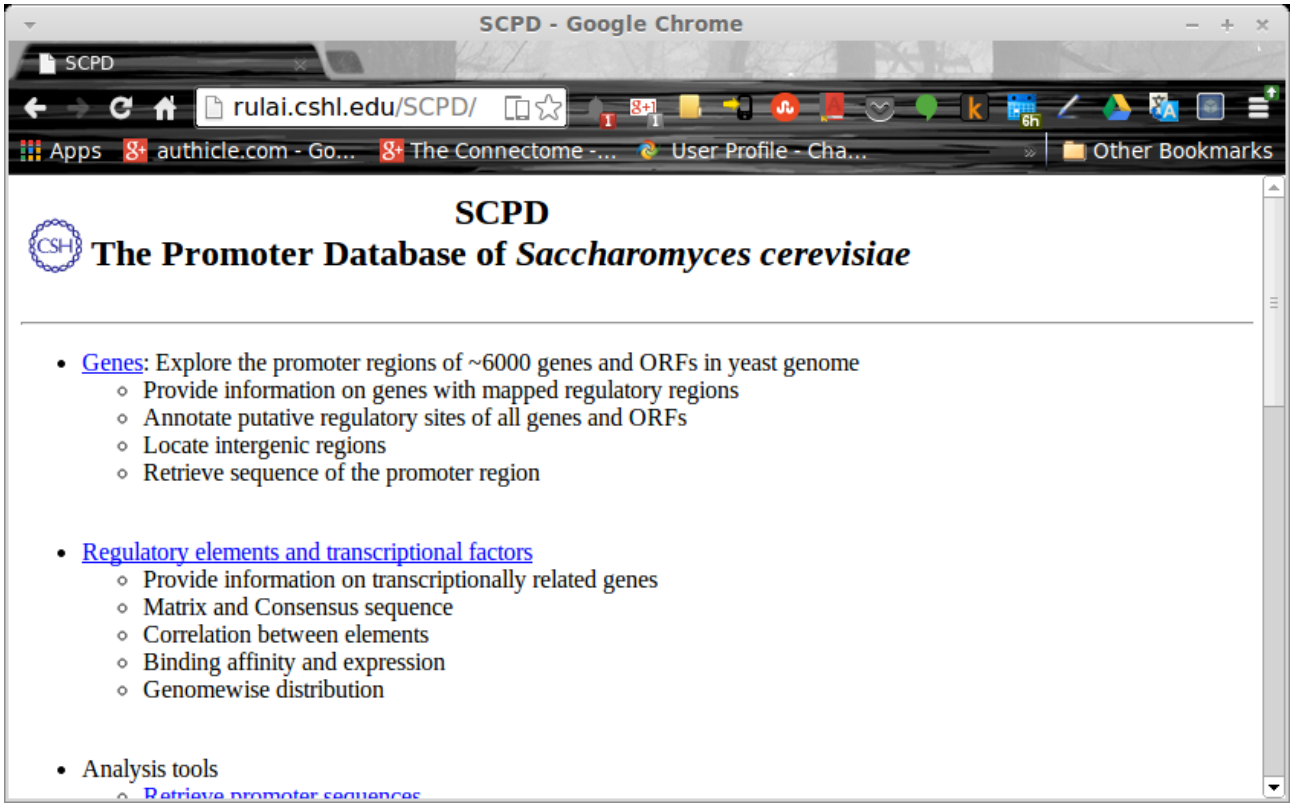
Exercise 2

In order to get the number of promoters I have simply loaded the FASTA file and counted how many were in it. The list contained 5880 promoters which agrees with the ~6000 promoter region specified at http://rulai.cshl.edu/SCPD/.

```
'''
Created on 27/05/2014

@author: s4361277
'''
from sequence import *

promoters = readFastaFile("yeast_promoters.fa", DNA_Alphabet)

print "There are", len(promoters), "promoters"
```

Running the program we get the result of 5880 promoters.




SCPD - Google Chrome
SCPD
rulai.cshl.edu/SCPD/
Apps
authicle.com - Go...
The Connectome -...
User Profile - Cha...
Other Bookmarks
SCPD
The Promoter Database of Saccharomyces cerevisiae
Genes: Explore the promoter regions of ~6000 genes and ORFs in yeast genome
Provide information on genes with mapped regulatory regions
Annotate putative regulatory sites of all genes and ORFs
Locate intergenic regions
Retrieve sequence of the promoter region
Regulatory elements and transcriptional factors
Provide information on transcriptionally related genes
Matrix and Consensus sequence
Correlation between elements
Binding affinity and expression
Genomewise distribution
Analysis tools

Exercise 3

The tenth column of the PWM is

	10th column
A	0.51
C	-25.33
G	0.84
T	-25.33

The values in a PWM are calculated by taking log ratios of frequency $q_{u,a}$ and background p_a so each field is calculated by $log(q_{ua}/p_a)$. Frequency is calculated by adding pseudo-counts of 1 to eliminate the problem of 0 counts. So for C and T we are now able to produce a highly negative value (-25.33) rather than undefined log(0).

a)

The reason we only see G and A in the tenth column is because the counts specified in the jaspr file for the other letters (C and T) are both 0.

b)

The total height represents information and is described by $I_u = \log_2 |A| - H_u$ where H_u shannon entropy and is calculated as $H_u = \sum - f_{u,a} \log_2 f_{u,a}$. From the jaspr file we can calculate frequency of A=41/98=0.42 and frequency of G=57/98=0.58.

Now, the entropy $H_u = (- 0.42 * \log_2 0.42) + (-0.58 * \log_2 0.58) = (-0.42 * - 1.25) + (-0.58 * -0.79) = 0.53 + 0.46 = 0.99$

Now, for DNA alphabets information content $I_u = \log_2 |4| - H_u$ so in our case $I_u = 2 - 0.99 =$ **1.01**

The total height is about 1 bit (half of total possible height of 2) because Shannon entropy (uncertainty) is about 1 bit (half of total possible height of 2).

Code

```
'''
Created on 27/05/2014

@author: s4361277
'''
from sequence import *

z = readMultiCount("abf1.jaspar")
pwm = PWM(z)
letters = ['A', 'C', 'G', 'T']
print pwm

print "PWM 10'th column"
for i in range(0, 4):
    print letters[i], pwm.m[i][9]
```

Output

```
PWM 10'th column
A 0.514898949154
C -25.3284360229
G 0.844378150284
T -25.3284360229
```

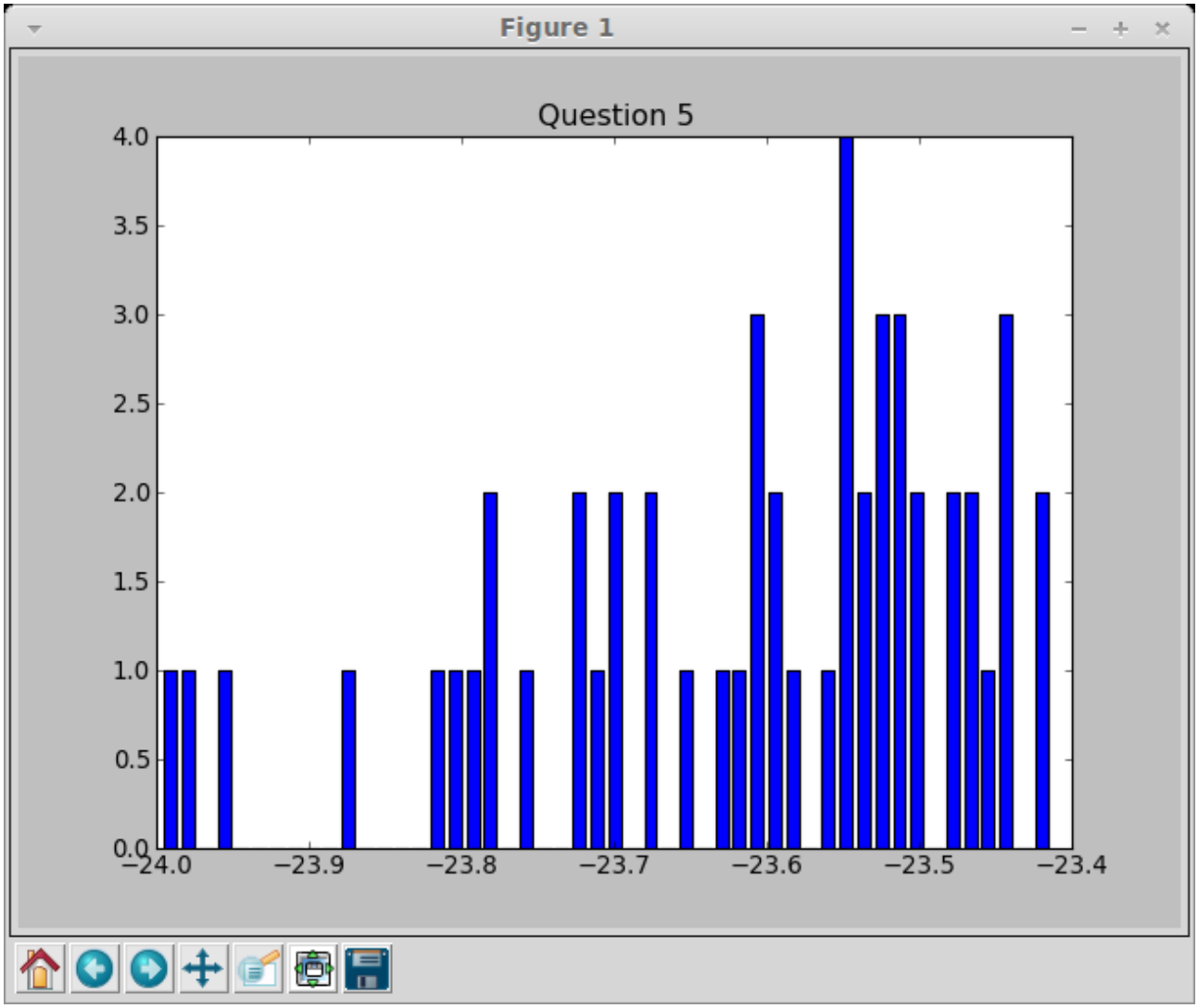
Exercise 5

Histogram and 51 genes are shown in sub sections following the code. Region from -24.0 to -23.4 was chosen as it gave us 51 genes which was close to the 50 specified by the question.

Code

```
'''
Created on 27/05/2014

@author: s4361277
'''
from sequence import *
import numpy as np
import matplotlib.pyplot as plt

yeast_prom = readFastaFile("yeast_promoters.fa", DNA_Alphabet)

z = readMultiCount("abf1.jaspar")
abf1_pwm = PWM(z)


bind_map = {}
for s in yeast_prom: #yeast_prom is an array of sequences
    if abf1_pwm.maxscore(s)[0] > -24 and abf1_pwm.maxscore(s)[0] < -23.4:
        bind_map[s.name] = abf1_pwm.maxscore(s)[0] # save score only
scores = []
for s in bind_map.keys():
    if bind_map[s] != None:
        scores.append(bind_map[s])

hist, bins = np.histogram(scores, bins=50)
width = 0.7 * (bins[1] - bins[0])
center = (bins[:-1] + bins[1:]) / 2
plt.bar(center, hist, align='center', width=width)
plt.title("Question 5")

print len(bind_map.keys()), "genes:"
for key in bind_map.keys(): print key

plt.show()
```

Histogram



Output

```
51 genes:
LOS1
CUE3
MSL1
SNC1
YBR090C
MGR2
YCR075W-A
HIR3
EUG1
YKL033W-A
YGL041W-A
SLP1
TAH18
ITT1
VPS9
ACF2
STN1
PDB1
SIR3
YML079W
MRI1
GOT1
YPI1
```

```
YFR032C-B
LAS1
CHS6
KAP114
TCO89
COA1
FAR11
THP1
TOM71
TVP38
MOB2
SCJ1
NBA1
THR1
NEJ1
TRS130
TRM112
SHU2
SLD3
YNL162W-A
MCX1
STE24
TFB3
DPB3
YDR514C
YPL264C
VPS55
CAT5
```

Exercise 6

There are 9275 probes and 8714 genes in GDS3198.soft. Please see the code below.

Code

```
'''
Created on 27/05/2014

@author: s4361277
'''
import genome

g1 = genome.readGEOFile('GDS3198.soft', id_column = 0)
g2 = genome.readGEOFile('GDS3198.soft', id_column = 1)

print "GDS3198.soft contains", len(g1.getGenes()), "probes"
print "GDS3198.soft contains", len(g2.getGenes()), "genes"
```

Execution

```
Data set GDS3198 contains 9275 genes
Data set has 13 null-values
Data set GDS3198 contains 8714 genes
Data set has 13 null-values
GDS3198.soft contains 9275 probes
GDS3198.soft contains 8714 genes
```

Exercise 7

Looking at the distribution below we can see that it is very close to normal, i.e. z-test is applicable.

The following WT/mutant pairs were used in the experiment

	Control Identifiers	Mutant Identifiers
Pair 1	GSM140786: Abf1 wt 37 C rep1	GSM140802: Abf1 ts 37 C rep1
Pair 2	GSM140800: Abf1 wt 37 C rep2	GSM140803: Abf1 ts 37 C rep2
Pair 3	GSM140801: Abf1 wt 37C rep3	GSM140804: Abf1 ts 37 C rep3

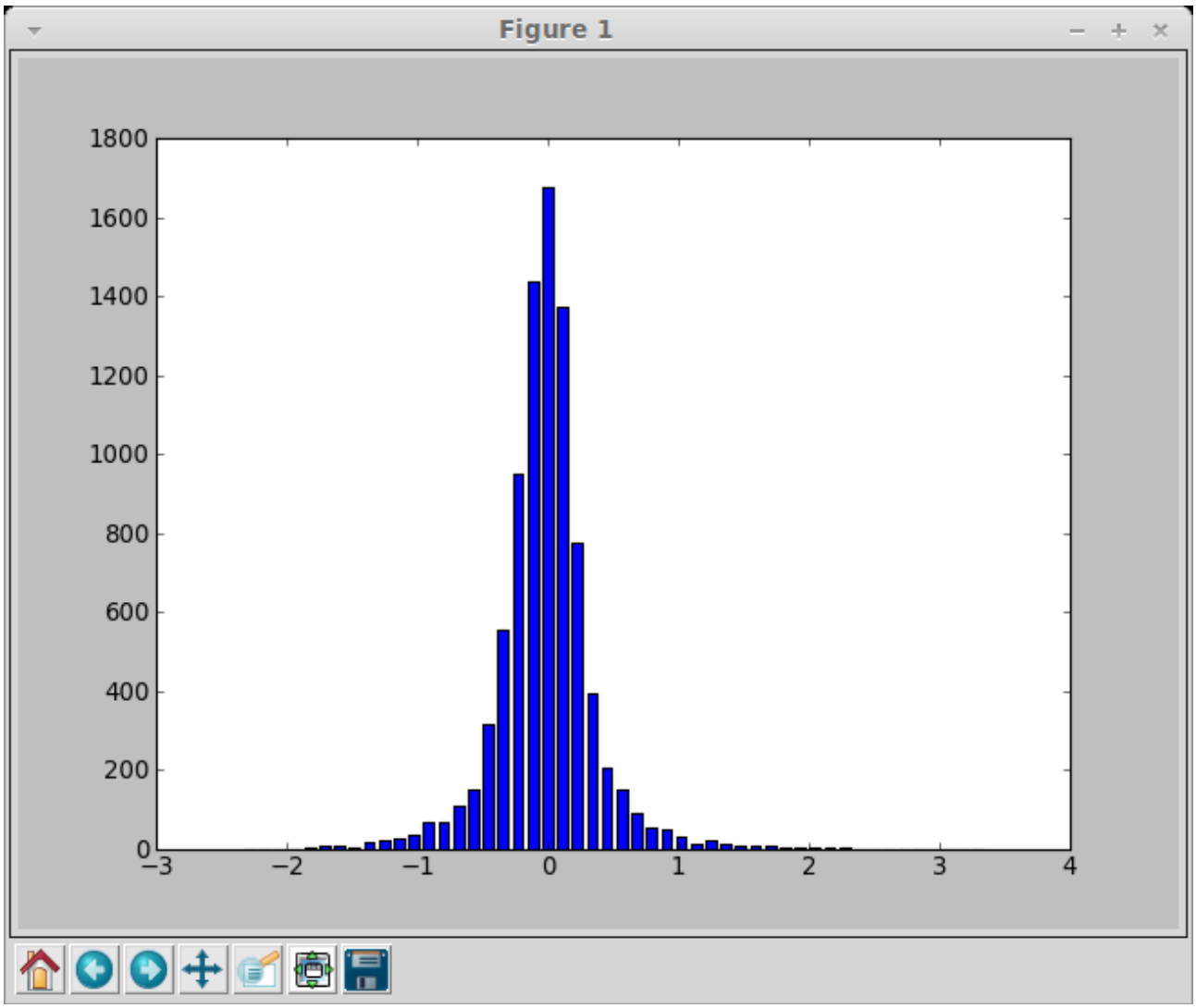
In order to obtain Gaussian distribution shown below log ratio transforms were performed in the following code:

```
    meanfold[gene] = (math.log(profile[0] / profile[3]) +
                      math.log(profile[1] / profile[4]) +
                      math.log(profile[2] / profile[5])) / 3
```

The above code is executed in a for loop iterating over the genes (see listing below the histogram for full code). For each gene we take each of the control samples and divide it by its corresponding mutant and take a log of the result. We then take the mean of the 3 results and store it in the meanfold list.

In order to output the results we first sorted our meanfold list using a provided function sorted(). Once the result was sorted would filter and print the top and bottom 100 based on the list index, for example, `result[0:100]` would return top 100.

```
result = sorted(meanfold.items(), key=lambda v: v[1])
print '========== Wildtype may down-regulate ==========='
for r in result[0:100]:
        print r[0], r[1]

print '========== Wildtype may up-regulate ==========='
for r in result[-1:-100:-1]:
        print r[0], r[1]
```



Code

```
'''
Created on 27/05/2014

@author: s4361277
'''
from sequence import *
import genome as ge
import matplotlib.pyplot as plt
import numpy as np

g = ge.readGEOFile('GDS3198.soft', id_column=1)
meanfold = {}
for gene in g.genes:
    profile = g.getGenes(gene)
    meanfold[gene] = (math.log(profile[0] / profile[3]) +
                      math.log(profile[1] / profile[4]) +
                      math.log(profile[2] / profile[5])) / 3

# pull out NaNs
scores = [y for y in meanfold.values() if not np.isnan(y)]
hist, bins = np.histogram(scores, bins=50)
width = 0.7 * (bins[1] - bins[0])
```

```
center = (bins[:-1] + bins[1:]) / 2
plt.bar(center, hist, align='center', width=width)
plt.title("Question 7")
plt.show()
result = sorted(meanfold.items(), key=lambda v: v[1])
print '========== Wildtype may down-regulate ==========='
for r in result[0:100]:
        print r[0], r[1]

print '========== Wildtype may up-regulate ==========='
for r in result[-1:-100:-1]:
        print r[0], r[1]
```

Exercise 8

In this question I have filtered out all names without spaces. Furthermore those without a space in the name were searched for UniProt and discarded if no match has been found. These filters produced 49 down regulated and 24 up regulated genes.

Code

```
'''
Created on 27/05/2014

@author: s4361277
'''
from sequence import *
import genome as ge
import matplotlib.pyplot as plt
import numpy as np

g = ge.readGEOFile('GDS3198.soft', id_column = 1)
meanfold = {}
for gene in g.genes:
    profile = g.getGenes(gene)
    meanfold[gene] = (math.log(profile[0] / profile[3]) +
                      math.log(profile[1] / profile[4]) +
                      math.log(profile[2] / profile[5])) / 3

result = sorted(meanfold.items(), key=lambda v: v[1])
print '========== Wildtype may down-regulate ==========='
c = 0
for r in result[0:100]:
    if not(' ' in r[0]) and len(search(r[0], dbName='uniprot',
format='list', limit=1)) == 1:
        print r[0]
        c +=1

print "total down regulated genes=", c
c = 0
print '========== Wildtype may up-regulate ==========='
for r in result[-1:-100:-1]:
    if not(' ' in r[0]) and len(search(r[0], dbName='uniprot',
format='list', limit=1)) == 1:
        print r[0]
        c += 1

print "total up regulated genes=", c
```

Output

```
Data set GDS3198 contains 8714 genes
Data set has 13 null-values
========== Wildtype may down-regulate ==========
ATG29
YLL067C
CDA1
IRC15
YAL064W-B
```

```
YHR145C
YPR078C
BSC1
RTG1
SPR3
YLR279W
YCR050C
SLZ1
YOL163W
YPL062W
YDR274C
YKL223W
YLR463C
YLR331C
YRO2
FIG1
YBL112C
YIL025C
YGL007W
BSC5
YLR428C
YDR124W
YDL152W
YNR073C
YLR169W
YGL015C
YCK3
YMR118C
YCR102C
snR31
YCR100C
AIF1
PDC6
NMD4
MPC54
YDR526C
CDA2
YBR089W
YGR045C
MET22
DMC1
TIS11
STE3
YPR123C
total down regulated genes= 49
========== Wildtype may up-regulate ==========
CEN3
POP8
CEN12
PRO3
PTH1
NIT3
MEK1
AIM27
YDR544C
YPL261C
COS12
STRP
CEN1
MRPL32
```

```
YOR008C-A
YPR153W
TPM2
SET2
SRP14
SWS2
HBS1
YKL069W
KAP120
YOR060C
total up regulated genes= 24
```

Exercise 9

E value (expectation value) is used as it is more suited than p value when multiple hypothesis testing. E value is the expectation of the number of hits (by chance) rather than probability (p value) and hence it is why it is obtained by multiplying p-value by the number of terms. In the case of `get_GO_term_overrepresentation` specifies number of GO term hits rather than the probability (p value) which is more indicative of the quality of the search.

Exercise 10

There were no GO terms that are over-represented in the Abf1 promoter set.

Code

```
'''
Created on 27/05/2014

@author: s4361277
'''
from sequence import *
from go import *
import numpy as np
import matplotlib.pyplot as plt

yeast_prom = readFastaFile("yeast_promoters.fa", DNA_Alphabet)

z = readMultiCount("abf1.jaspar")
abf1_pwm = PWM(z)


bind_map = {}
for s in yeast_prom:  # yeast_prom is an array of sequences
    if abf1_pwm.maxscore(s)[0] > -24 and abf1_pwm.maxscore(s)[0] < -23.4:
        bind_map[s.name] = abf1_pwm.maxscore(s)[0]  # save score only
scores = []
for s in bind_map.keys():
    if bind_map[s] != None:
        scores.append(bind_map[s])

godb = GODB("yeast_go")
r = godb.get_GO_term_overrepresentation(bind_map.keys(), evalThreshold=1.0)
print "r=", r
```

Output

```
Loaded map with 4674 gene symbols
r= {}
```

Exercise 11

Please see the output section for the list of GO terms that are over-represented in the Abf1 differential set.

Code

```
'''
Created on 27/05/2014

@author: s4361277
'''
from sequence import *
import genome as ge
from go import *

g = ge.readGEOFile('GDS3198.soft', id_column=1)
meanfold = {}
for gene in g.genes:
    profile = g.getGenes(gene)
    meanfold[gene] = (math.log(profile[0] / profile[3]) +
                      math.log(profile[1] / profile[4]) +
                      math.log(profile[2] / profile[5])) / 3

result = sorted(meanfold.items(), key=lambda v: v[1])
genes = []

for r in result[0:100]:
    if not(' ' in r[0]) and len(search(r[0], dbName='uniprot',
format='list', limit=1)) == 1:
        genes.append(r[0])

for r in result[-1:-100:-1]:
    if not(' ' in r[0]) and len(search(r[0], dbName='uniprot',
format='list', limit=1)) == 1:
        genes.append(r[0])

godb = GODB("yeast_go")
r = godb.get_GO_term_overrepresentation(genes, evalThreshold=1.0)
print "GO terms:"
for term in r: print term
```

Output

```
Data set GDS3198 contains 8714 genes
Data set has 13 null-values
Loaded map with 4674 gene symbols
GO terms:
GO:0030435
GO:0048610
GO:0030476
GO:0032505
GO:0048869
GO:0030154
GO:0042244
GO:0022414
GO:0022413
```

Exercise 12

Looking at the summary paragraph of the ABF1 (http://www.yeastgenome.org/cgi-bin/locus.fpl?locus=abf1#summaryParagraph) it looks that most of the GO terms retrieved fit with yeast except the cell differentiation which would most likely belong to multi cellular organisms. Note that these are GO terms from question 11 as question 10 produced none.

Output (GO descriptions)

```
Data set GDS3198 contains 8714 genes
Data set has 13 null-values
Loaded map with 4674 gene symbols
GO terms:
biological_process: sporulation
biological_process: reproductive cellular process
biological_process: spore wall assembly (sensu Fungi)
biological_process: reproduction of a single-celled organism
biological_process: cellular developmental process
biological_process: cell differentiation
biological_process: spore wall assembly
biological_process: reproductive process
biological_process: reproductive process in single-celled organism
```

Code

```
from sequence import *
import genome as ge
from go import *

g = ge.readGEOFile('GDS3198.soft', id_column=1)
meanfold = {}
for gene in g.genes:
    profile = g.getGenes(gene)
    meanfold[gene] = (math.log(profile[0] / profile[3]) +
                      math.log(profile[1] / profile[4]) +
                      math.log(profile[2] / profile[5])) / 3

result = sorted(meanfold.items(), key=lambda v: v[1])
genes = []

for r in result[0:100]:
    if not(' ' in r[0]) and len(search(r[0], dbName='uniprot',
format='list', limit=1)) == 1:
        genes.append(r[0])

for r in result[-1:-100:-1]:
    if not(' ' in r[0]) and len(search(r[0], dbName='uniprot',
format='list', limit=1)) == 1:
        genes.append(r[0])

godb = GODB("yeast_go")
r = godb.get_GO_term_overrepresentation(genes, evalThreshold=1.0)
print "GO terms:"

for term in r:
    print godb.get_GO_description(term)
```

Appendix A – Question 7 full output

```
Data set GDS3198 contains 8714 genes
Data set has 13 null-values
========== Wildtype may down-regulate ==========
ATG29 -2.18070388835
YCLWOMEGA2 -2.01198659566
non-annotated SAGE orf Found forward in NC_001143 between 173981 and 174175
with 100% identity. -1.83565207722
YLL067C -1.76054129809
non-annotated SAGE orf Found forward in NC_001145 between 481528 and 481713
with 100% identity. -1.70198182463
CDA1 -1.67897716821
IRC15 -1.67618097337
Found forward in NC_001146 between 130021 and 131020 with 100% identity.
-1.66006848578
Found forward in NC_001140 between 299647 and 300646 with 100% identity.
-1.51101177265
YAL064W-B -1.42163792379
YHR145C -1.35384433684
Found forward in NC_001137 between 535026 and 536025 with 100% identity.
-1.3415498064
YPR078C -1.33600224127
non-annotated SAGE orf Found forward in NC_001145 between 271996 and 272136
with 100% identity. -1.32083328432
YDR112W -1.30337730421
Found forward in NC_001142 between 368944 and 369943 with 100% identity.
-1.3017380545
Found forward in NC_001148 between 387268 and 388267 with 100% identity.
-1.26886360087
BSC1 -1.25950906434
Found forward in NC_001141 between 204053 and 205052 with 100% identity.
-1.23323012341
Found forward in NC_001133 between 88357 and 89356 with 100% identity.
-1.17025564047
non-annotated SAGE orf Found reverse in NC_001139 between 1037741 and
1037887 with 100% identity. -1.12646580225
RTG1 -1.12474596396
YOLCDELTA2 -1.08417523985
non-annotated SAGE orf Found forward in NC_001134 between 181316 and 181477
with 100% identity. -1.04365302025
non-annotated SAGE orf Found forward in NC_001134 between 9384 and 9605
with 100% identity. -1.03173671189
Found forward in NC_001133 between 216649 and 217143 with 100% identity.
-1.01669743793
non-annotated SAGE orf Found forward in NC_001136 between 979658 and 979807
with 100% identity. -1.01433136487
SPR3 -1.00494678089
Found forward in NC_001139 between 924127 and 925126 with 100% identity.
-1.00215863918
non-annotated SAGE orf Found reverse in NC_001136 between 1149727 and
1149861 with 100% identity. -0.977140699145
Found forward in NC_001139 between 334616 and 335615 with 100% identity.
-0.976485598348
YLR279W -0.968323595511
YCR050C -0.956059092271
SLZ1 -0.954652026641
YOL163W -0.948184287541
YPL062W -0.914903251787
```

```
YDR274C -0.906722856219
non-annotated SAGE orf Found reverse in NC_001145 between 234512 and 234685
with 100% identity. -0.897340703468
non-annotated SAGE orf Found reverse in NC_001147 between 974085 and 974252
with 100% identity. -0.897126558349
YKL223W -0.890999521014
non-annotated SAGE orf Found reverse in NC_001147 between 271475 and 271732
with 100% identity. -0.886852018222
YLR463C -0.863683751206
YLR331C -0.847726347571
YRO2 -0.81960665351
non-annotated SAGE orf Found forward in NC_001146 between 140489 and 140683
with 100% identity. -0.804691718167
FIG1 -0.798885833407
YBL112C -0.792777057205
YIL025C -0.788409292989
YGL007W -0.787153587495
non-annotated SAGE orf Found forward in NC_001139 between 788087 and 788224
with 100% identity. -0.785958587854
Found forward in NC_001144 between 1043294 and 1044293 with 100% identity.
-0.783906622993
Found forward in NC_001148 between 448337 and 449336 with 100% identity.
-0.780854875827
YERWDELTA11 -0.777280466865
BSC5 -0.775164457964
YLR428C -0.769387842883
Found forward in NC_001134 between 754813 and 755812 with 100% identity.
-0.763599045747
non-annotated SAGE orf Found reverse in NC_001135 between 8959 and 9150
with 100% identity. -0.754499411231
YDR124W -0.738802141034
YFLWTAU1 -0.734666568676
Found forward in NC_001134 between 76564 and 77563 with 100% identity.
-0.724856235081
Found forward in NC_001144 between 310356 and 311355 with 100% identity.
-0.721447568523
YDL152W -0.715508899423
non-annotated SAGE orf Found reverse in NC_001137 between 251194 and 251418
with 100% identity. -0.708712262618
YDRWSIGMA4 -0.704618706798
YNR073C -0.690429272851
YLR169W -0.689408132064
YGL015C -0.682041207299
non-annotated SAGE orf Found reverse in NC_001142 between 312518 and 312670
with 100% identity. -0.678566742115
Found forward in NC_001133 between 208649 and 209648 with 100% identity.
-0.661261198567
YCK3 -0.65629720026
non-annotated SAGE orf Found reverse in NC_001143 between 108918 and 109193
with 100% identity. -0.649539081097
YMR118C -0.649064365635
YCR102C -0.646648830069
non-annotated SAGE orf Found reverse in NC_001148 between 408869 and 409009
with 100% identity. -0.646426134689
snR31 -0.646198041102
YHRCTAU4 -0.645979296508
YCR100C -0.643323210404
YGRCDELTA25 -0.638103462763
Found forward in NC_001144 between 634184 and 635183 with 100% identity.
-0.627123875973
```

```
Found forward in NC_001144 between 309356 and 310355 with 100% identity.
-0.624617404822
AIF1 -0.622612762368
Found forward in NC_001144 between 629184 and 630183 with 100% identity.
-0.617927890466
PDC6 -0.606010683159
NMD4 -0.600597019536
MPC54 -0.594373634442
YDR526C -0.591195658992
CDA2 -0.59064988505
YIR020c-a -0.584726322877
YBR089W -0.581631726409
YGR045C -0.576838079382
YBRWDELTA16 -0.575249182566
MET22 -0.574645215501
non-annotated SAGE orf Found reverse in NC_001147 between 241012 and 241308
with 100% identity. -0.56325921133
DMC1 -0.563141405096
Found forward in NC_001143 between 412757 and 413756 with 100% identity.
-0.559969664554
TIS11 -0.549879557275
Found forward in NC_001144 between 1049294 and 1050293 with 100% identity.
-0.542159792763
STE3 -0.541608586463
non-annotated SAGE orf Found forward in NC_001143 between 146588 and 146755
with 100% identity. -0.53744734736
YPR123C -0.532080392713
========== Wildtype may up-regulate ==========
Found forward in NC_001146 between 659170 and 660169 with 100% identity.
2.97594184414
YJLWDELTA9 2.51300901861
Found forward in NC_001141 between 9696 and 10695 with 100% identity.
2.44769886642
Found forward in NC_001144 between 353507 and 354506 with 100% identity.
2.31705582148
CEN3 2.24327469016
YJLWDELTA2 2.22647722254
non-annotated SAGE orf Found reverse in NC_001142 between 227571 and 227705
with 100% identity. 2.14066738999
Found forward in NC_001141 between 108607 and 109606 with 100% identity.
2.13374898256
Found forward in NC_001144 between 355507 and 356506 with 100% identity.
2.11006179385
Found forward in NC_001147 between 177973 and 178972 with 100% identity.
2.0840928313
Found forward in NC_001137 between 442412 and 443411 with 100% identity.
2.0778163782
Found forward in NC_001146 between 175087 and 176086 with 100% identity.
2.07610734481
snR57 2.07503347341
Found forward in NC_001143 between 541784 and 542783 with 100% identity.
2.01619459546
non-annotated SAGE orf Found reverse in NC_001137 between 212169 and 212351
with 100% identity. 2.00448115128
POP8 1.95014917009
Found forward in NC_001146 between 164130 and 165129 with 100% identity.
1.94647656097
Found forward in NC_001144 between 633184 and 634183 with 100% identity.
1.88943743174
CEN12 1.85636919044
```

```
Found forward in NC_001143 between 665918 and 666445 with 100% identity.
1.84984678039
PRO3 1.8431041243
Found forward in NC_001140 between 522872 and 523871 with 100% identity.
1.78457446804
Found forward in NC_001136 between 80795 and 81794 with 100% identity.
1.74322152443
PTH1 1.66049189774
Found forward in NC_001138 between 188470 and 189469 with 100% identity.
1.65280372771
non-annotated SAGE orf Found reverse in NC_001147 between 978298 and 978459
with 100% identity. 1.64930181574
Found forward in NC_001140 between 40533 and 41532 with 100% identity.
1.61844847274
non-annotated SAGE orf Found forward in NC_001143 between 298846 and 299052
with 100% identity. 1.5951861565
non-annotated SAGE orf Found reverse in NC_001141 between 139370 and 139600
with 100% identity. 1.58652629901
Found forward in NC_001136 between 345736 and 346735 with 100% identity.
1.58501266598
Found forward in NC_001145 between 882563 and 883562 with 100% identity.
1.57367000049
Found forward in NC_001144 between 936909 and 937908 with 100% identity.
1.57161822438
non-annotated SAGE orf Found reverse in NC_001133 between 199737 and 199886
with 100% identity. 1.54490061303
Found forward in NC_001147 between 174973 and 175972 with 100% identity.
1.52075081097
Found forward in NC_001143 between 620375 and 621374 with 100% identity.
1.47896401901
Found forward in NC_001148 between 765478 and 766477 with 100% identity.
1.45870499568
Found forward in NC_001148 between 871140 and 872139 with 100% identity.
1.45373355441
non-annotated SAGE orf Found reverse in NC_001140 between 5778 and 5924
with 100% identity. 1.40218141944
NIT3 1.39889558633
Found forward in NC_001140 between 525872 and 526871 with 100% identity.
1.39440102221
MEK1 1.37536442018
Found forward in NC_001144 between 1047294 and 1048293 with 100% identity.
1.37450961755
Found forward in NC_001148 between 111147 and 112146 with 100% identity.
1.37441863524
Found forward in NC_001143 between 653494 and 654493 with 100% identity.
1.3721885919
AIM27 1.36025528501
YDR544C 1.35650382111
YOLCDELTA1 1.32050256894
Found forward in NC_001144 between 997726 and 998725 with 100% identity.
1.32011985066
YPLWDELTA6 1.30196508799
YPL261C 1.29595923186
Found forward in NC_001145 between 884563 and 885562 with 100% identity.
1.29247718689
YERCDELTA26 1.28548854215
Found forward in NC_001144 between 333178 and 334177 with 100% identity.
1.2785507066
Found forward in NC_001143 between 59435 and 60434 with 100% identity.
1.27222146232
```

```
Found forward in NC_001142 between 372944 and 373943 with 100% identity.
1.26569767958
Found forward in NC_001140 between 41533 and 42532 with 100% identity.
1.26008927073
Found forward in NC_001148 between 386268 and 387267 with 100% identity.
1.24983469198
Found forward in NC_001147 between 948724 and 949723 with 100% identity.
1.24063011194
COS12 1.22709860863
YORWDELTA19 1.22463433331
non-annotated SAGE orf Found forward in NC_001142 between 445314 and 445592
with 100% identity. 1.20055839816
Found forward in NC_001141 between 109607 and 110606 with 100% identity.
1.19619953589
Found forward in NC_001148 between 389268 and 390267 with 100% identity.
1.19606711914
STRP 1.19005438242
Found forward in NC_001147 between 349694 and 350693 with 100% identity.
1.18106216548
non-annotated SAGE orf Found forward in NC_001145 between 115459 and 115659
with 100% identity. 1.13854667564
CEN1 1.13787330055
Found forward in NC_001134 between 638162 and 639161 with 100% identity.
1.13713225824
Found forward in NC_001148 between 761478 and 762477 with 100% identity.
1.12282382375
Found forward in NC_001146 between 761618 and 762617 with 100% identity.
1.10975476024
Found forward in NC_001142 between 13138 and 14137 with 100% identity.
1.09676016127
MRPL32 1.0871275504
YOR008C-A 1.08344449185
YPR153W 1.0743413276
YLRCTAU1 1.07141997494
non-annotated SAGE orf Found forward in NC_001141 between 197558 and 197818
with 100% identity. 1.07111832079
Found forward in NC_001144 between 101543 and 102542 with 100% identity.
1.06927636543
Found forward in NC_001145 between 154719 and 155718 with 100% identity.
1.06475560112
Found forward in NC_001148 between 627964 and 628963 with 100% identity.
1.05282515148
TPM2 1.04643986703
non-annotated SAGE orf Found forward in NC_001136 between 1385623 and
1385760 with 100% identity. 1.04205954141
Found forward in NC_001147 between 961693 and 962692 with 100% identity.
1.04039249971
non-annotated SAGE orf Found forward in NC_001138 between 234229 and 234471
with 100% identity. 1.0166256304
non-annotated SAGE orf Found forward in NC_001142 between 447920 and 448102
with 100% identity. 1.01037806687
YJLWDELTA1 1.00890417302
non-annotated SAGE orf Found forward in NC_001137 between 434581 and 434727
with 100% identity. 1.00558654692
SET2 1.00212427325
Found forward in NC_001137 between 531026 and 532025 with 100% identity.
0.983530670847
SRP14 0.980429628833
mRNA maturase bI3 Found forward in NC_001224 between 39141 and 40265 with
98.577778% identity. 0.977064584139
```

```
SWS2 0.976914871776
HBS1 0.975200438019
CEN8 0.974907079304
YKL069W 0.971805536014
Found forward in NC_001146 between 527084 and 528083 with 100% identity.
0.968934140856
KAP120 0.952338049781
Found forward in NC_001147 between 730506 and 731505 with 100% identity.
0.952141718314
YOR060C 0.951326856559
Found forward in NC_001135 between 276986 and 277985 with 100% identity.
0.947321068121
```