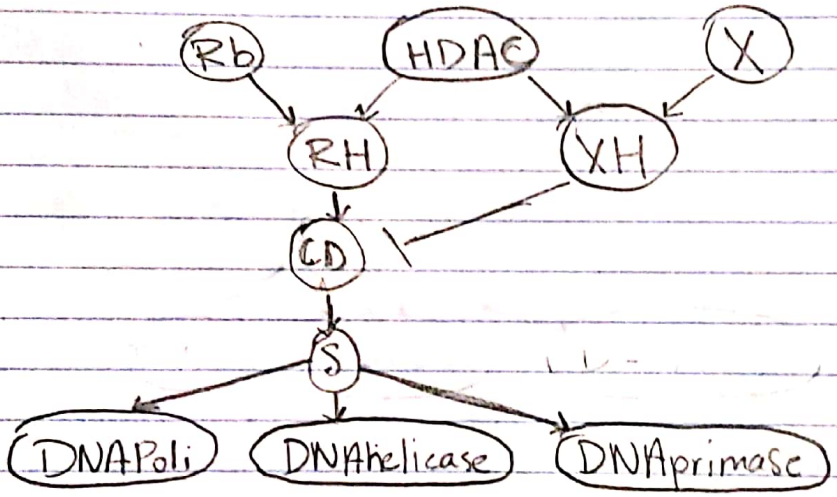
BME 230B - Homework 1

1. a) Nodes we have Rb, HDAC RH S, X, XH, Transcription - D-Pol, D-Heli, D-Prime, chromatin Dynamics CD



G

b) RH and XH likely to be hidden b/c both act upon S and depend on HDAC so we could infer which is active based on Rb or X but we can't ever know for sure.

c) The follow independence assumptions are true I(S-phase, Rb | CD - closed), I(HDAC, Rb), I(HDAC, X)

d) $P(RH+ \mid CD = open) = \frac{P(CD = open \mid RH+)\, P(RH+)}{P(CD = open)}$

neat page

$P(A,B) = P(A)P(B)$ so both independent

$$P(RH+) = P(RH \mid R+, HDAC+)\,P(R+, HDAC+) + P(RH \mid R-, HDAC+)\,P(R-, HDAC+) + P(RH \mid R+, HDAC-)\,P(R+, HDAC-) + P(RH \mid R-, HDAC-)\,P(R-, HDAC-)$$

$\therefore P(RH-) = 1 - P(RH+)$ — bc indep, $P(A,B) = P(A)P(B)$

$$P(CD=open) = P(CD=open \mid RH+, XH+)\,P(RH+)\,P(XH+) + P(CD=open \mid RH+, XH-)\,P(RH+)\,P(XH-) + P(CD=open \mid RH-, XH+)\,P(RH-)\,P(XH+) + P(CD=open \mid RH-, XH-)\,P(RH-)\,P(XH-)$$

(needed for $P(CD=open)$)

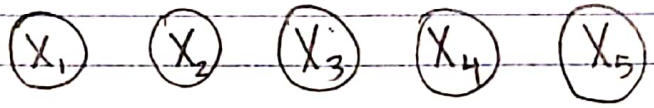
$$P(XH+) = P(XH+ \mid HDAC+, X+)\,P(X+)\,P(HDAC+) + P(XH+ \mid HDAC-, X+)\,P(X+)\,P(HDAC-) + P(XH+ \mid HDAC+, X-)\,P(X-)\,P(HDAC+) + P(XH+ \mid HDAC-, X-)\,P(X-)\,P(HDAC-)$$

$\therefore P(XH-) = 1 - P(XH+)$

$$P(CD=open \mid RH+) = P(CD=open \mid RH+, XH+)\,P(XH+)\,P(RH+) + P(CD=open \mid RH+, XH-)\,P(XH-)\,P(RH+)$$

Now all components can be plugged in to solve for bayes rule for $P(RH+ \mid CD=open)$.

2.a) BN of position independent model (a zero markov chain)



b/c no dependence from one another -

$\Theta_{(X_i|Pa(i))} = \Theta_{X_i}$ for all i in vector Θ

in the PSSM. Θ_i need 4 (for i in 1 to 5)

b) BN of position dependent model

Parameters needed:

Θ_1 need 4, 1 for each nuc in "ATGC"

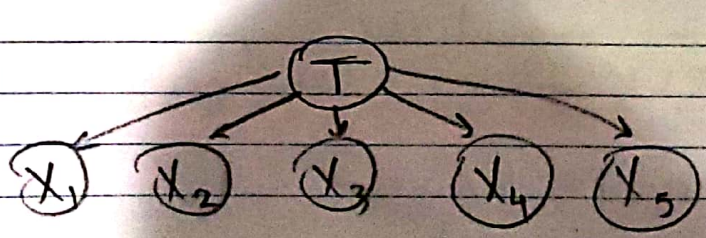
Θ_5 needs 4, 1 for each nuc in "ATGC"

Θ_2 needs 16 $\rightarrow \sum_{i \in ATGC} \sum_{j \in ATGC} P(\Theta_{2i}, \Theta_{1j})$

Θ_4 needs 16 $\rightarrow \sum_{i \in ATGC} \sum_{j \in ATGC} P(\Theta_{4i}, \Theta_{5j})$

Θ_3 needs 64 $\sum_{k \in ATGC} P(\Theta_{3i} | \Theta_{4j}, \Theta_{2k}) P(\Theta_{4j}) P(\Theta_{2k})$

c)



each Θ_i in Θ needs 8 parameters

A, T, G, or C given TF1 and A, T, G, C given TF2

d) $P(X_1 X_2 \ldots X_5 \mid \Theta\ G_{2c})$ find

$\Theta = [\Theta_1, \Theta_2, \ldots, \Theta_5]$

G_{2c} = graph made in 2c

$X = X_1 X_2 \ldots X_5$ (the sequence)

$$P(X \mid \Theta\ G_{2c}) = \prod_{i=1}^{n=5} P(X_i \mid \Theta_i)$$

where

$$P(X_i \mid \Theta_i) = P(X_i \mid \Theta_i, T = TF1)\, P(T = TF1) + P(X_i \mid \Theta_i, T = TF2)\, P(T = TF2)$$

e) I would get all the sequences and find

$P(T = TF1 \mid X_i)$

where X_i is the binding site sequence and X is the set of sequences.

for each we would find

$P(X_i \mid T = TF1)$ the likelihood

$P(X_i)$ the normalizer

$P(T = TF1)$ which is the prior

and use naive bayes.

from there I would find

$$P(T = TF1 \mid X) = \frac{1}{n} \sum_{i=1}^{n} P(T = TF1 \mid X_i)$$

next page

where n is number of sequences in X.
and since T is binary so

$$P(T = TF2 \mid X) = 1 - P(T = TF1 \mid X)$$

which ever has higher probability is most likely the transcription factor used.

4 a) $\log\left[\frac{P_z(Y \mid Model\text{-}X)}{P_z(Y \mid Model\text{-}Null)}\right] =$

$\log[P_z(Y \mid Model\text{-}X)] - \log[P_z(Y \mid Model\text{-}Null)]$

log-likelihood for Model A = 29.90

$-37.5305 - (-67.4307) = 29.9002$

log-likelihood for Model B = 10.86

$-56.5723 - (-67.4307) = 10.8584$

b) on PDF
c) on PDF
d) Model A better explains the data it has a better distribution of data with a shape like a guassian distribution.

3 a) $P(hap^+) = 0.5 \qquad P(X^+) = 0.5$

$P(vpsT+) = P(vpsT+ | hap+)(hap+) + P(vpsT+ | hap-)$

$P(vpsT+) = (0.1)(0.5) + (0.9)(0.5)$

$P(vpsT+) = 0.05 + 0.45 = 0.5$

$P(hapR+) = 0.5 \qquad P(hapR-) = 0.5$

$P(vpsT+) = 0.5 \qquad P(vpsT-) = 0.5$

$P(vpsR+) = P(vpsR+ | vpsT+, X+) P(vpsT+, X+) +$
$P(vpsR+ | vpsT-, X+) P(vpsT-, X+) +$
$P(vpsR+ | vpsT+, X-) P(vpsT+, X-) +$
$P(vpsR+ | vpsT-, X-) P(vpsT-, X-)$

$P(vpsR+) = (0.9)(0.5 \cdot 0.5) + (0.9)(0.5 \cdot 0.5) +$
$(0.9)(0.5 \cdot 0.5) + (0.1)(0.5 \cdot 0.5)$

$P(vpsR+) = 0.7$

$P(vpsR-) = 0.3$

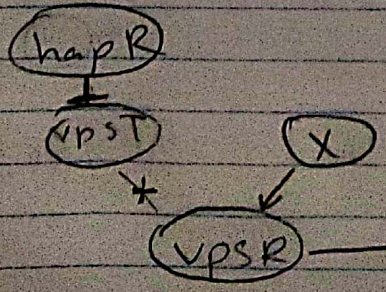
$P(Y+) = P(Y+ | vpsR+) P(vpsR+) +$
$P(Y+ | vpsR-) P(vpsR-)$

$P(Y+) = (0.9)(0.7) + (0.1)(0.3)$

$P(Y+) = (0.63) + (0.03) = 0.66$

$\boxed{P(Y+) = 0.66}$

b)



$\hat{\theta}$ includes
- vpsT- (knocked out)
- P(vpsR+), still depends on if X+ or not.
- Y+ is still wild type prob.

c) $P(vps^R+, X+ \mid Y+, G_A^*, \Theta_A^*) =$

$$\frac{P(Y+, G_A^*, \Theta_A^* \mid vps^R+, X+)\, P(vps^R+, X+)}{P(Y+, G_A^*, \Theta_A^*)}$$

$\hookrightarrow P(Y+)$

$P(vps\,R+, X+) = P(vps R+ \mid X+)\, P(X+)$

b/c vps T is knocked out

$P(vps R+ \mid X+) = P(vps\,R+ \mid X+, vps\,T-)$

$P(vps R+ \mid X+) = 0.9$

$P(X+) = 0.5$

$P(vps R+, X+) = 0.45$

given $G_A^* \Theta_A^*$

$P(Y+) = P(Y+ \mid vps\,R+)\, P(vps R+) + P(Y+ \mid vps\,R-) P(vps R-)$

$P(vps R+) = P(vps R+ \mid X+) P(X+) + P(vps R+ \mid X-) P(X-)$

$P(vps R+) = (0.9)(0.5) + (0.1)(0.5)$

$P(vps R+) = .45 + .05 = 0.5$

$\therefore P(vps R-) = 0.5$

$P(Y+ \mid vps\,R+) = 0.9$

$P(Y+ \mid vps\,R-) = 0.1$

$P(Y+) = (0.9)(0.5) + (.1)(0.5)$

$P(Y+) = 0.5$

$P(Y+ \mid vps\,R+, X+) = .9 \propto .9 * .5 = .405$

$$P(vpsR+, X+ \mid Y+, G_A, \theta_A) = \frac{.405 \cdot .45}{.5}$$

$$= \boxed{.3645}$$

3d) let D be data

find $P(D \mid G_A, \theta_A)$

$$P(D \mid G_A \theta_A) = \left(\prod_{i=1}^{2} P(Y_i+ \mid G_A, \theta_A)\right) \cdot P_{vpsT-}(Y_1- \mid G_A, \theta_A) \cdot P_{vpsT-}(Y_2+ \mid G_A, \theta_A)$$

$$P(Y_i+ \mid G_A, \theta_A) = P(Y_i+ \mid vspR+)P(vspR+) + P(Y_i+ \mid vspR-)P(vspR-)$$

$P(vspR+) = 0.7$ and $P(vspR-) = 0.3$
(both calculated in 3A)

$P(Y_i+ \mid vspR+) = 0.9$ $\quad P(Y_i+ \mid vspR-) = 0.1$
(from Model A)

$$P(Y_i \mid G_A, \theta_A) = (0.9)(0.7) + (0.1)\cdot(0.3)$$

$$P(Y_i \mid G_A, \theta_A) = 0.66$$

$$P_{vpsT-}(Y_1+ \mid G_A, \theta_A) = P_{vpsT-}(Y_1+ \mid vpsR+)P_{vspT-}(vpsR+) + P_{vpsT-}(Y_1+ \mid vpsR-)P_{vspT-}(vpsR-)$$

↓

b/c $P_{vpsT-}(Y_1- \mid G_A, \theta_A) = 1 - P_{vpsT-}(Y_1+ \mid G_A, \theta_A)$

$P_{vpsT-}(vspR+) = 0.50$ $\quad P_{vspT-}(vspR-) = 0.5$
(from 3c)

$P_{vspT-}(Y_{1+} | vspR+) = 0.9$

$P_{vspT-}(Y_{1+} | vspR-) = 0.1$ from model A

$P_{vspT-}(Y_{1+} | G_A, \Theta_A) = (0.9 \cdot 0.5) + (0.1 \cdot 0.5)$

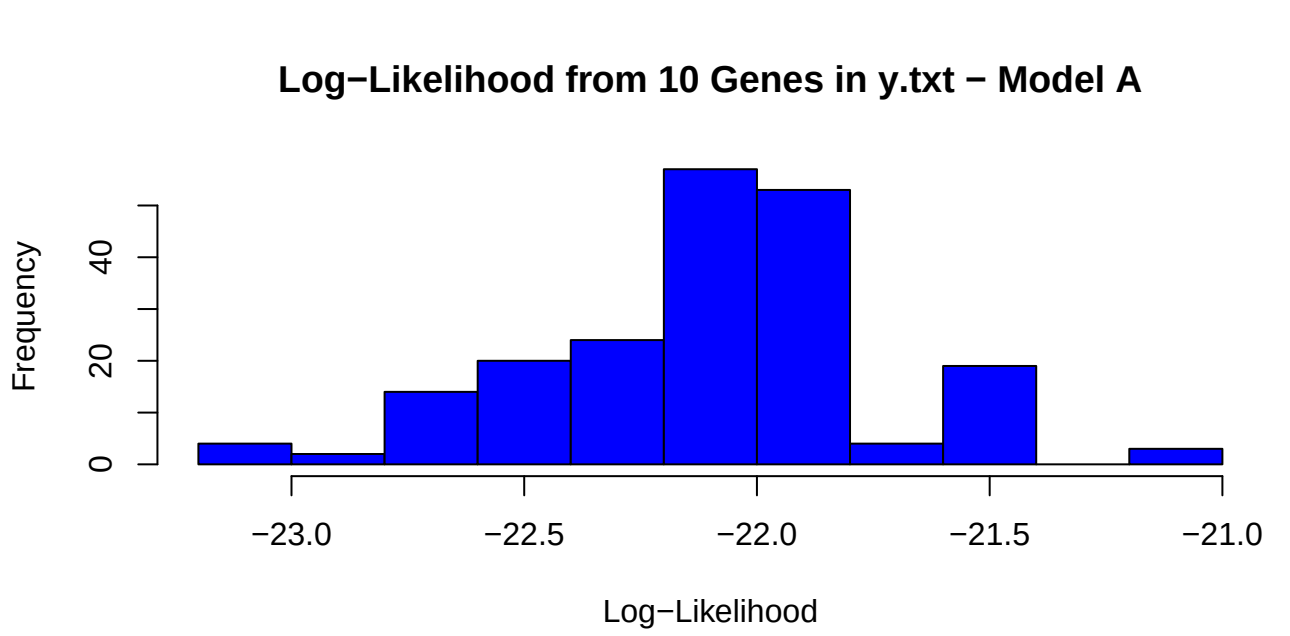
$P_{vspT-}(Y_{1+} | G_A, \Theta_A) = 0.5$

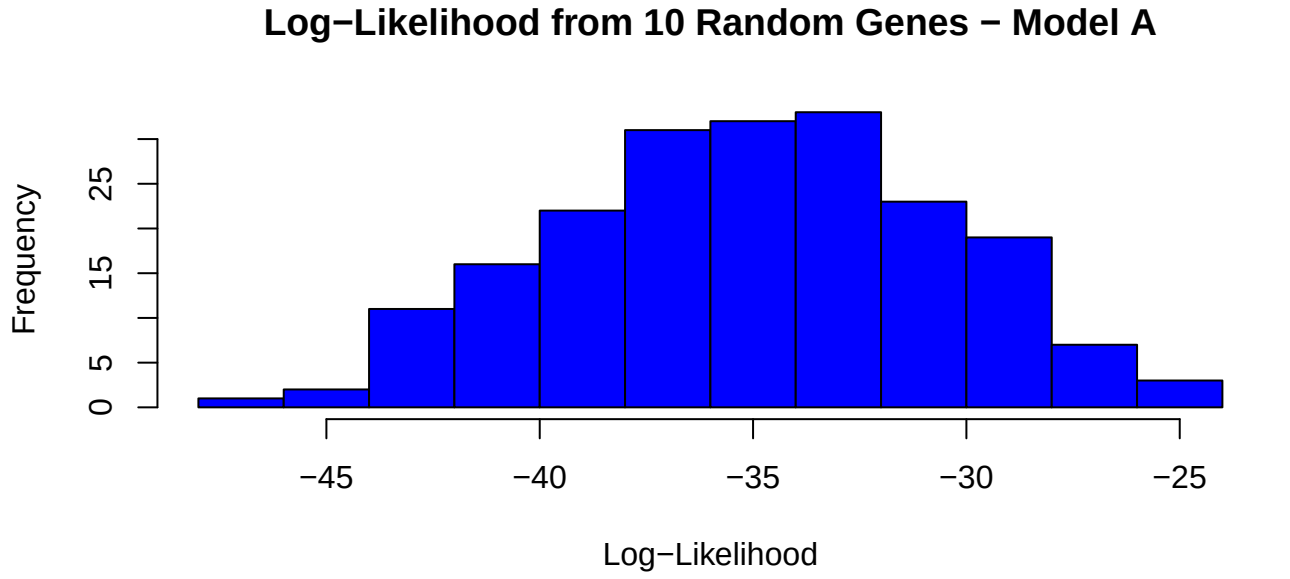
$P_{vspT-}(Y_{1-} | G_A, \Theta_A) = 0.5$

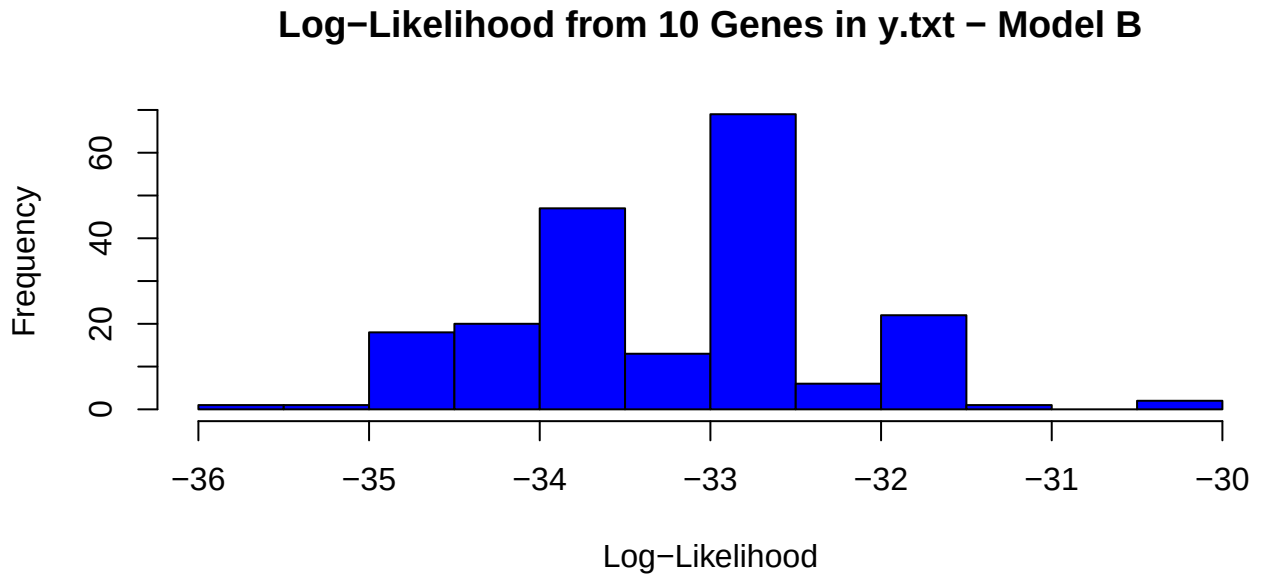
b/c same model $P_{vspT-}(Y_{2+} | G_A, \Theta_A) = 0.5$

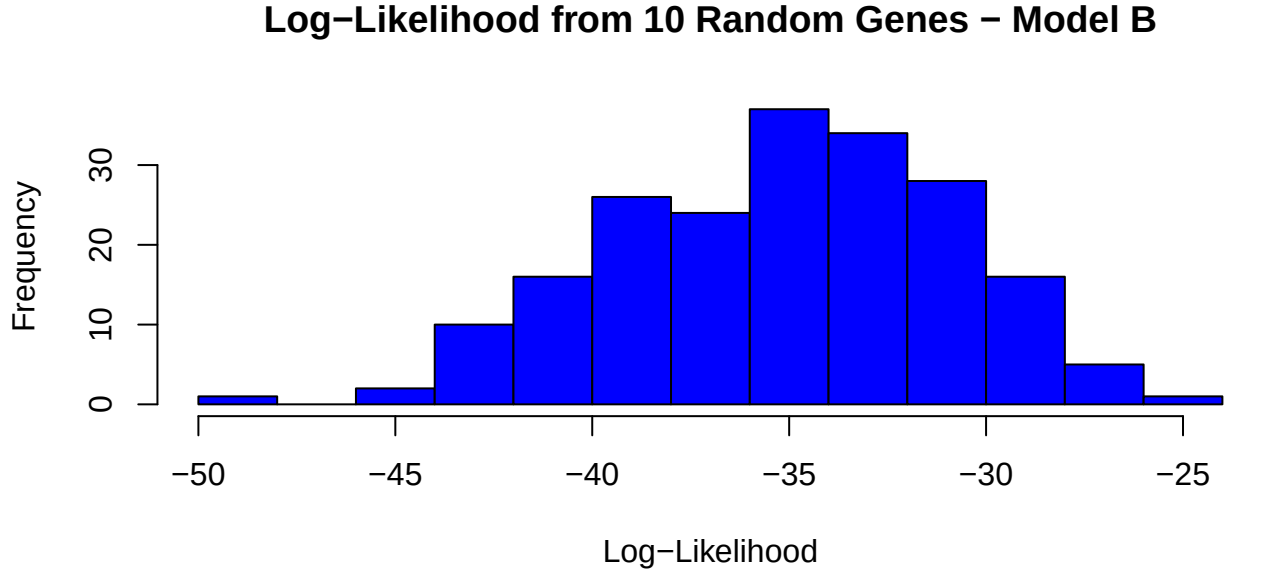
$$P(D | G_A, \Theta_A) = \left[\prod_{i=1}^{2} (0.666)\right] \cdot 0.5 \cdot 0.5$$

$$\boxed{P(D | G_A, \Theta_A) = 0.1089}$$


Log−Likelihood from 10 Genes in y.txt − Model A
Frequency
0
20
40
−23.0
−22.5
−22.0
−21.5
−21.0
Log−Likelihood


Log−Likelihood from 10 Random Genes − Model A
Frequency
0
5
15
25
−45
−40
−35
−30
−25
Log−Likelihood


Log−Likelihood from 10 Genes in y.txt − Model B
Frequency
0
20
40
60
−36
−35
−34
−33
−32
−31
−30
Log−Likelihood


Log−Likelihood from 10 Random Genes − Model B
Frequency
0
10
20
30
−50
−45
−40
−35
−30
−25
Log−Likelihood

```
# ---------------------------------------------------------
# HW1 Problem 4b
#
# Script by: Mohammad Abdulqader
#
# this shell script is used to solve problem 4b in
# the HW1 assignment for BME 230B
# ---------------------------------------------------------

let a=0

# create these files
echo log_likelihood > experiment_1.txt
echo log_likelihood > experiment_2.txt
tail -3850 data.txt | awk -F '\t' '{print $1}' > y2.txt

while [ $a -lt 200 ]
do

    # shuf to get 10 random Y values.
    shuf -n 10 y.txt > sample_Y.txt
    shuf -n 10 y2.txt > sample_Y2.txt

    # now run pathy.jar for first experiment
    java -jar pathy.jar sample_Y.txt data.txt modela.bif >> experiment_1.txt

    # now run pathy.jar for second experiment
    java -jar pathy.jar sample_Y2.txt data.txt modela.bif >> experiment_2.txt

    # increment a (number of iterations)
    let a=a+1

done

# remove excess files
rm sample_Y.txt ReadMe.txt YgenesIndScore.txt y2.txt sample_Y2.txt

# done
```

```
# ---------------------------------------------------------
# HW1 Problem 4c
#
# Script by: Mohammad Abdulqader
#
# this shell script is used to solve problem 4c in
# the HW1 assignment for BME 230B
# ---------------------------------------------------------

let a=0

# create these files
echo log_likelihood > experiment_3.txt
echo log_likelihood > experiment_4.txt
tail -3850 data.txt | awk -F '\t' '{print $1}' > y2.txt

while [ $a -lt 200 ]
do

    # shuf to get 10 random Y values.
    shuf -n 10 y.txt > sample_Y.txt
    shuf -n 10 y2.txt > sample_Y2.txt

    # now run pathy.jar for first experiment
    java -jar pathy.jar sample_Y.txt data.txt modelb.bif >> experiment_3.txt

    # now run pathy.jar for second experiment
    java -jar pathy.jar sample_Y2.txt data.txt modelb.bif >> experiment_4.txt

    # increment a (number of iterations)
    let a=a+1

done

# remove excess files
rm sample_Y.txt ReadMe.txt YgenesIndScore.txt y2.txt sample_Y2.txt


# done
```

```
# ---------------------------------------------------------
# HW1 Problem 4
#
# Script by: Mohammad Abdulqader
#
# this R script is used to solve problem 4b and 4c in
# the HW1 assignment for BME 230B
# ---------------------------------------------------------

# read data
data1 <-read.csv('experiment_1.txt')
data2 <-read.csv('experiment_2.txt')
data3 <-read.csv('experiment_3.txt')
data4 <-read.csv('experiment_4.txt')


# Model A
pdf('Problem_4b.pdf')
par(mfrow = c(2, 1))

hist(data1$log_likelihood, main='Log-Likelihood from 10 Genes in y.txt - Model A',
    col = 'blue', xlab = 'Log-Likelihood')
hist(data2$log_likelihood, main='Log-Likelihood from 10 Random Genes - Model A',
    col = 'blue', xlab = 'Log-Likelihood')

dev.off()

# Model B
pdf('Problem_4c.pdf')
par(mfrow = c(2, 1))

hist(data3$log_likelihood, main='Log-Likelihood from 10 Genes in y.txt - Model B',
    col = 'blue', xlab = 'Log-Likelihood')
hist(data4$log_likelihood, main='Log-Likelihood from 10 Random Genes - Model B',
    col = 'blue', xlab = 'Log-Likelihood')

dev.off()

par(mfrow = c(1, 1))
```