Gene Expression + Class comparison

Final Exam: Dec 14th (6pm) : open book, covers all topics

Goal: Capture and compare "state" of different cells

- cells with positive outcome upon treatment
- negative outcome

→ state $\vec{P} = (p_1, p_2 \ldots, p_{20000})$

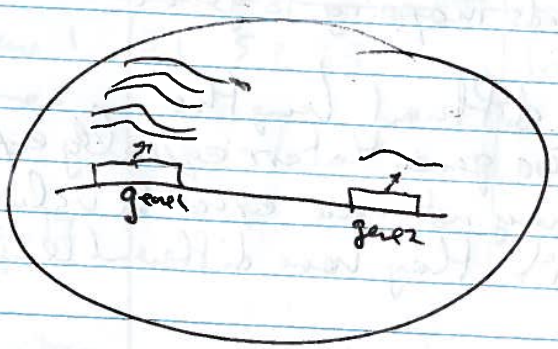
where p_i = abundance of protein p_i in cells.

Problem: Measuring protein abundance is hard (mass spectrometry)

Alternative: Measure mRNA abundance

$\vec{G} = (g_1, g_2 \ldots, g_{20000})$

where g_i = abundance of mRNA from gene i in cells



→ $g_1 = 5$
$g_2 = 1$

Note: $g_i \neq p_i$ because: mRNA degration
translation is regulated

RNA-sequencing (RNA-seq)

Goal: Measure gene expression levels $\vec{G} = (g_1 \cdots g_{20000})$

① Extract RNA from cells
② Reverse transcribed RNA to cDNA (↑ complementary)
③ Fragment cDNA in ~200bp pieces
④ Sequence each cDNA fragment

Read1 : ACTAG...
Read2 : ...
⋮
Read 10^7 : TACTA

Human Genes	R_i
gene1	1
gene2	3
gene3	1
⋮	
gene 20000	0

⑤ Map (↑ align) each read to its gene
Find the gene to which the read best aligns
⑥ Count R_i = #of reads mapping to gene i
Problem: Genes have different lengths
⇒ Two genes that are equally expressed may not have equal R values if they have different length
⑦ Normalize:

$$FPKM(g_i) = \frac{R_i}{(\text{length of } g_i \text{ (in kb)}) \cdot (\text{Total \# reads in Millions})}$$

FPKM: Fragment per kilobase per million reads

Class Comparison Problem

Given: Normalized gene expression data from two set of samples

A: (control) with $N_A = 20$ samples

$\vec{A_1} = (A_1(1), A_1(2) \ldots A_1(20000))$

$A_1(i)$ ↖ FPKM of gene i in sample 1

$\vec{A_2} = ($ $)$

$\vec{A_{N_A}} = ($ $)$

B = (treatment) with $N_B = 25$ samples

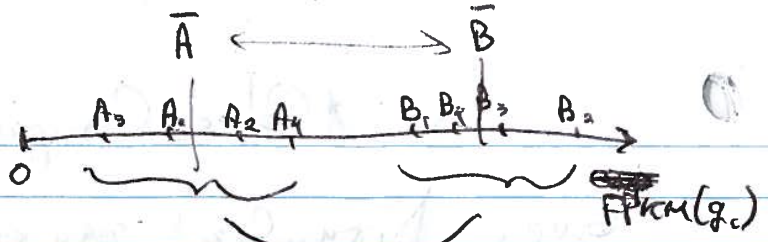
$\vec{B_1} = ($ $)$

$\vec{B_{N_B}} = ($ $)$

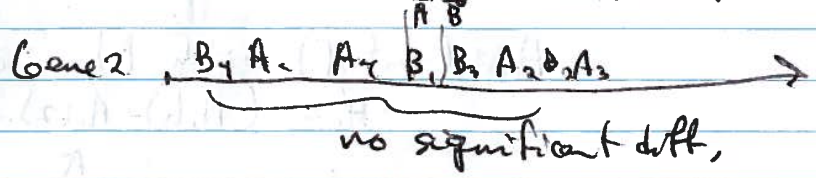
	20			25				t-stat	p-value
	A_1	A_2 ...	A_{N_A}	B_1			B_{N_B}		
Gene 1	5.1	5.7	3.7	1.7	1.3	1.9	2.1	7.1	0.001
Gene 2								1.3	0.25
⋮									⋮
Gene 17								best genes	0.0001 = 1/10^4
Gene 20000									0.35

Goal: Find genes that are "differentially expressed" b/w A, B

Example

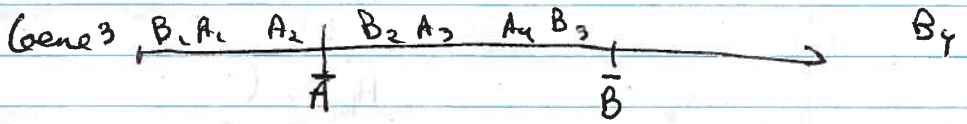


Significant diff



no significant diff.

$d(g_i)$ = diff. of means: $\bar{A}(g_i) - \bar{B}(g_i)$



Student t-test (performed separately for each gene)

H_0: Expression values from samples A and B come from the same normal distribution → Assumption

$\mu_A = \mu_B$

$\sigma_A = \sigma_B$

H_1: $\mu_A \neq \mu_B$

① Calculate $t(g_i) = \dfrac{\bar{A}(g_i) - \bar{B}(g_i)}{\sqrt{\dfrac{S_A^2}{N_A} + \dfrac{S_B^2}{N_B}}}$

S_A^2 = Variance in A

② Calculate [p-value for $t(g_i)$] = Prob that two random samples drawn from same distrib. would have t-statistics $t(g_i)$

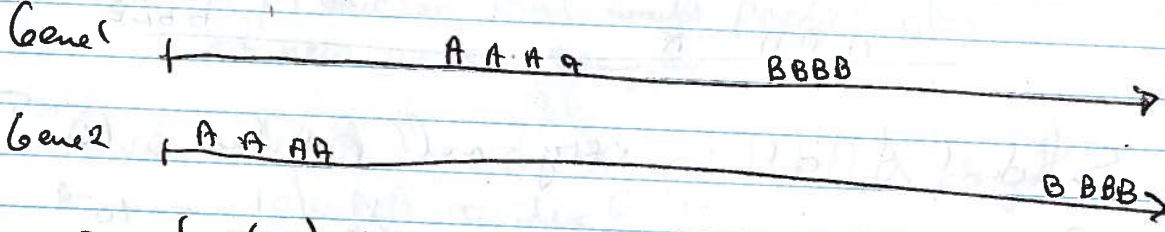
Under H_0: t follow a Student(m) ← degrees of freedom

$$m = \frac{\left(\frac{S_A^2}{N_A} + \frac{S_B^2}{N_B}\right)^2}{\frac{\left(\frac{S_A^2}{N_A}\right)^2}{N_A - 1} + \frac{\left(\frac{S_B^2}{N_B}\right)^2}{N_B - 1}}$$

If p-value(g_i) ≤ 0.05, then all g_i "diff. expressed"

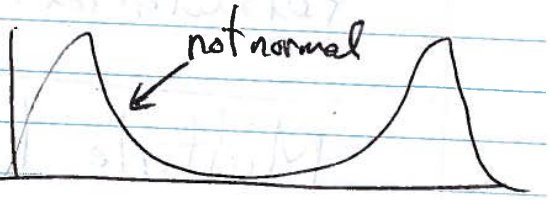
> 0.05, then g_i is not diff. exp.

Issue: Violation of assumptions: - Normality of data



p-value(g_2) < p-value(g_1)

Suppose distribution under H_0



Permutation test: Estimate p-value without assumption about underlying distrib. of data

Real $A_1\ A_2\ A_4\ A_5$

Shuffled $B_4\ A_1\ A_2\ B_3$

For each gene i

① Calculate $t(g_i)$

② Repeat $K=1000$ times

2.1 Randomly reshuffle class labels to obtain $\tilde{A}, \tilde{B}$

2.2 Calculate $\tilde{t}(g_i)$ from

2.3 If $\tilde{t}(g_i) \geq t(g_i)$ the

③ Report p-value$(g_i) = \frac{\text{success}}{K}$

A A A A B B B B

Student t-test: very small p-value: 10^{-}

Permutation: p-value: $\frac{1}{\binom{8}{4}} = 10^{-}$

Multiple hypothesis testing

We've done 20,000 tests

If all genes come from H_0, then best value we would expect to observe would be: $1/20{,}000$

Bonferroni correction: corrected Pvalue(g_i) = p-value(g_i) · N=

Class Prediction Problem

	A_1 A_2 ... A_{N_A}	B_1 ... B_{N_B}	X
Gene 1			
2	RNA-seq data		
3			
Gene 20000			

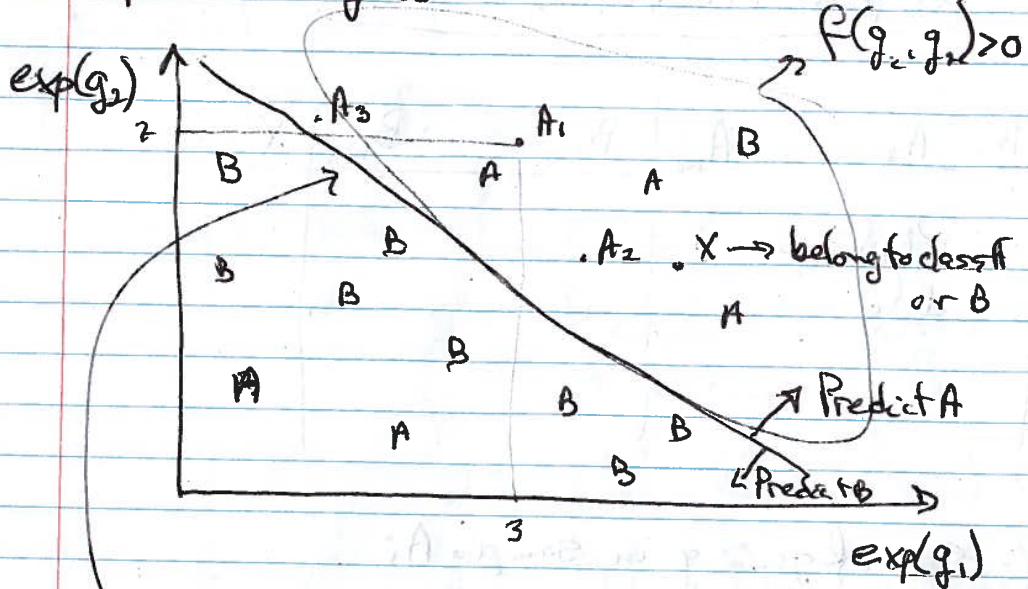
$A_i[g]$ = exp. of gene g in sample A_i

$B_j[g]$ = ...

Goal: From RNA-seq data from A, B, train a predictor that would predict class of new, unseen samples

Given: RNA-seq data X, predict if X belongs to class A or class B.

Assume # genes = 2



Linear classifier: linear function of $exp(g_1)$, $exp(g_2)$

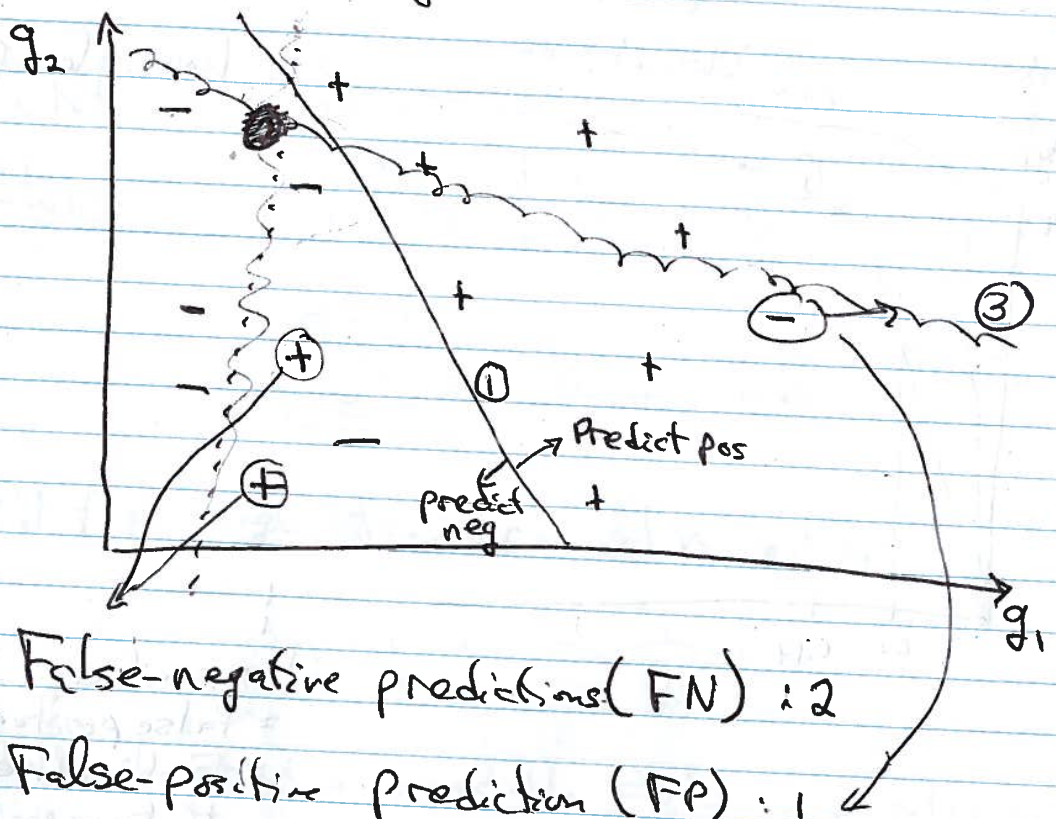
$g_2 = 10 - 2g_1$

$f(g_1, g_2) > 0 \rightarrow$ predict A

$< 0 \rightarrow$ predict B

$f(g_1, g_2) = 2g_1 + g_2 - 10$

Assessing a classifier's accuracy



False-negative predictions (FN) : 2

False-positive prediction (FP) : 1

True-positive (TP) : 7

True-negative (TN) : 5

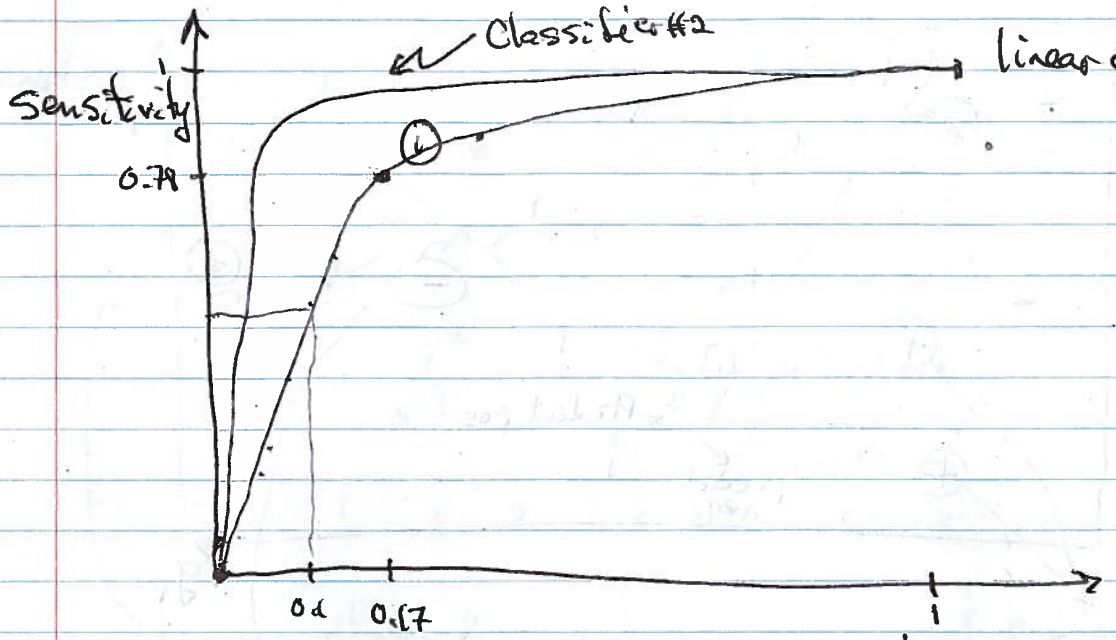
① Sensitivity : $\frac{TP}{TP+FN} = \frac{7}{7+2} = \frac{7}{9} \approx 78\%$

Specificity : $\frac{TN}{TN+FP} = \frac{5}{5+1} = \frac{5}{6} = 83\%$

② Sensitivity : 100%
Specificity : 50%

③ Sensitivity = 4/9 ≈ 55%
Specificity : 100%

Receiving-Operating Curve



1-specificity
= false positive rate
= Fraction neg. example that are predicted pos.