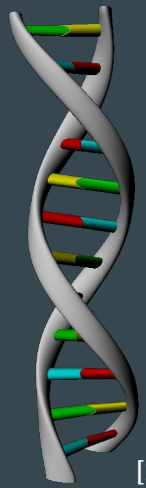
Applications of Statistics to Genomic Clinical Group Classification

by Jake Sauter

Project Goal

- The goal of this project was to train a machine learning classifier to accurately predict the true class (cancer type) of a genetic sample from a patient, given a training set to extrapolate from
    - These genetic samples were from patients with two different types of cancers that must be treated very differently, though are difficult to differentiate from more classical methods
    - A total of 72 samples were available from the data set, all pre-normalized with randomized training and testing sets
        - 47 ALL Cases and 25 AML Cases

Central Dogma of Molecular Biology



[2]

DNA

Transcription

[3]

RNA

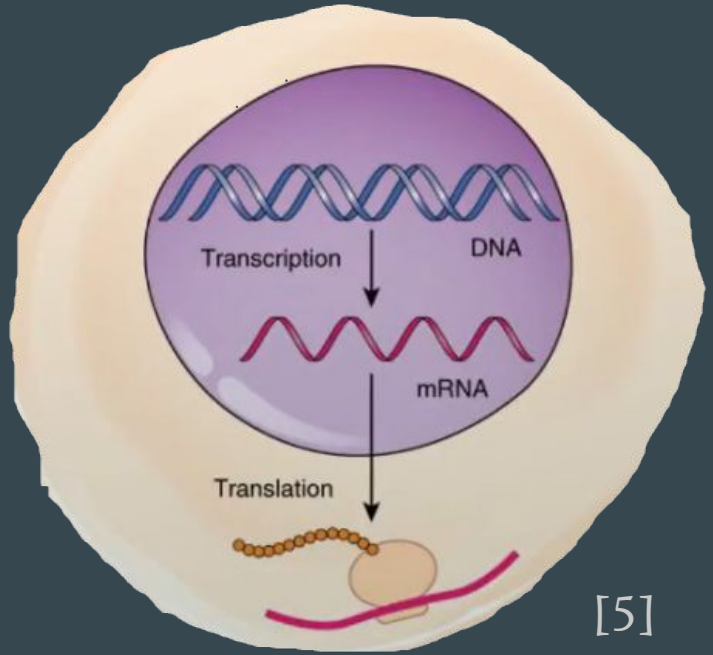
Translation

[4]

Protein

Biological Background

- The **genome** is the entire set of genes in an organism, and gives rise to the name of the study of **genomics**
- However DNA, which is contained in the **nucleus** of a cell, cannot leave this nucleus due to a non-DNA-permeable membrane
- Instead, **mRNA** (messenger RNA) copies desired DNA in the nucleus and permeates through the membrane

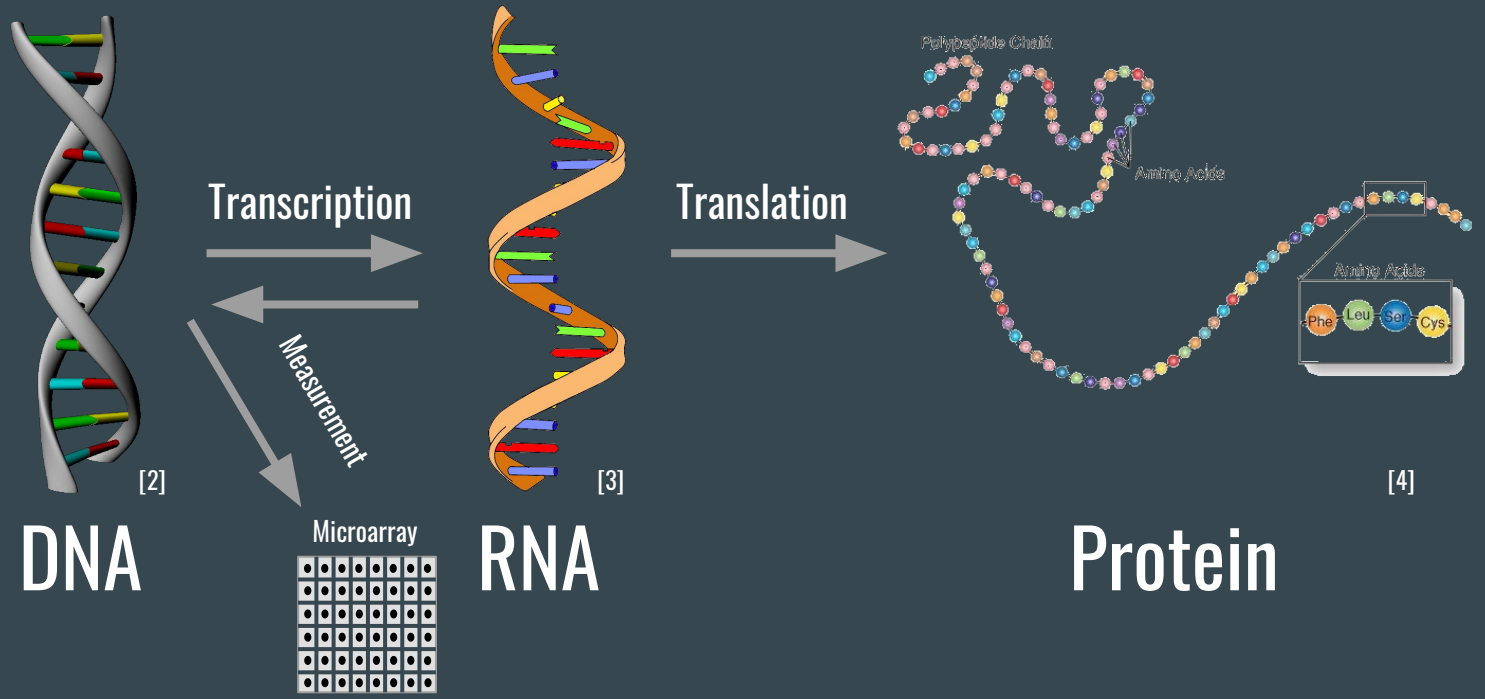


[5]

Biological Background

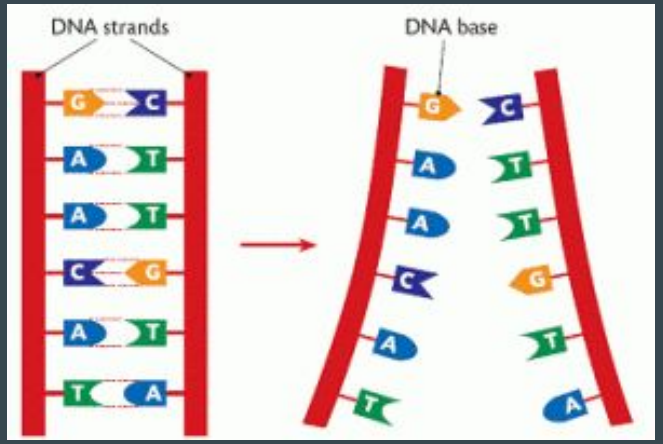
- In a last step before this sequence can be analyzed, mRNA is **reverse transcribed** into DNA
- The total **gene expression level** (how much RNA associated with a gene makes it out of the cell nucleus), which will then affect how much influence a particular gene has on the biological result
- We are interested in these **genetic expression levels** as they are measurable characteristic of different biological outcomes
    - If the gene expression levels for a gene that regulate cellular replication is modified in a particular way, that person has cancer!

Central Dogma of Molecular Biology



Biological Background

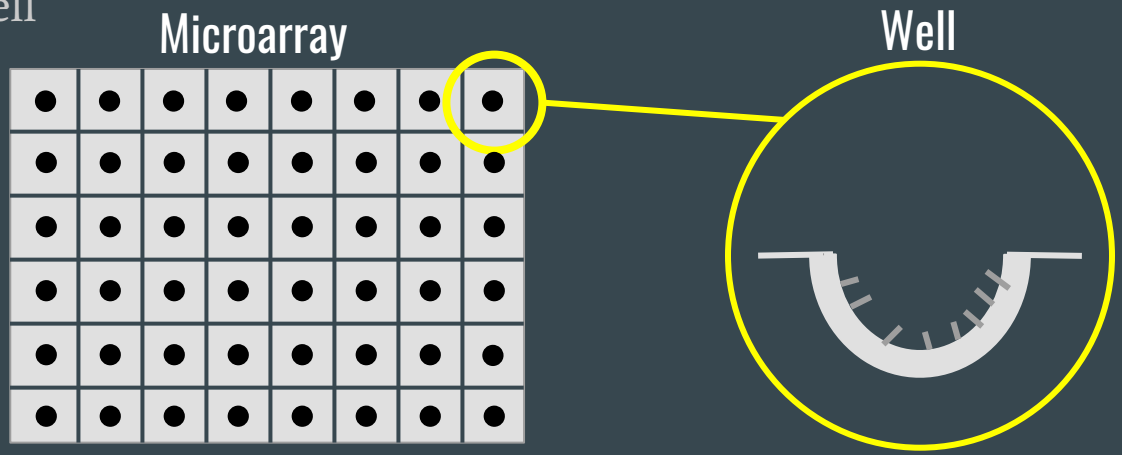
- A key idea that we will need moving forward is that of **complementary DNA**, which is when two single strand DNA (ssDNA) strands have sequences of compatible bases that allow them to **hybridize** (pair) and form a complete DNA strand



[6]

Microarrays: Introduction

- DNA Microarray - Chip that contains many cells, each with a **probe** (well) containing complementary DNA for a single gene
- Samples of genetic material are first fluorescently labelled then deposited on the chip ; afterwards the chip can be scanned to determine how much cDNA binded to each well



Microarrays: Measuring Gene Expression

- Wells contain DNA material complementary to each gene
- Sample DNA material is fluorescently labelled and will hybridize (pair to) the complementary DNA (cDNA) in the well
- The level of hybridization (fluorescence level) of each well can easily be measured in a scanner and can indicate the level of expression of a gene corresponding to the cDNA in the well
- Expression levels from DNA samples of different tissues can be compared using multiple Microarrays (Usually one Microarray per patient)

Microarrays: Challenges

- Noise
  - Introduced at each step in the complex process
- Normalization
  - Not always performed in the same manner
- Experimental Design
  - Not always thoughtfully designed
- Large number of genes
  - Sometimes finding the one influential gene in excess of 7000 is equivalent to finding a needle in a haystack
- Significance
  - Classical techniques (e.g. χ squared test) cannot be applied because the number of variables is much greater than the number of experiments

Microarrays: Challenges

- Biological factors
  - The expression level is the amount of protein produced, not the amount of mRNA (what is detected by the microarray)
  - Other tools can simply not be replaced by microarrays
  - Gene regulation to biological impact is a complicated non-linear mapping
- Array quality assessment
  - Sources of variability can include mRNA preparation, transcription and labelling processes

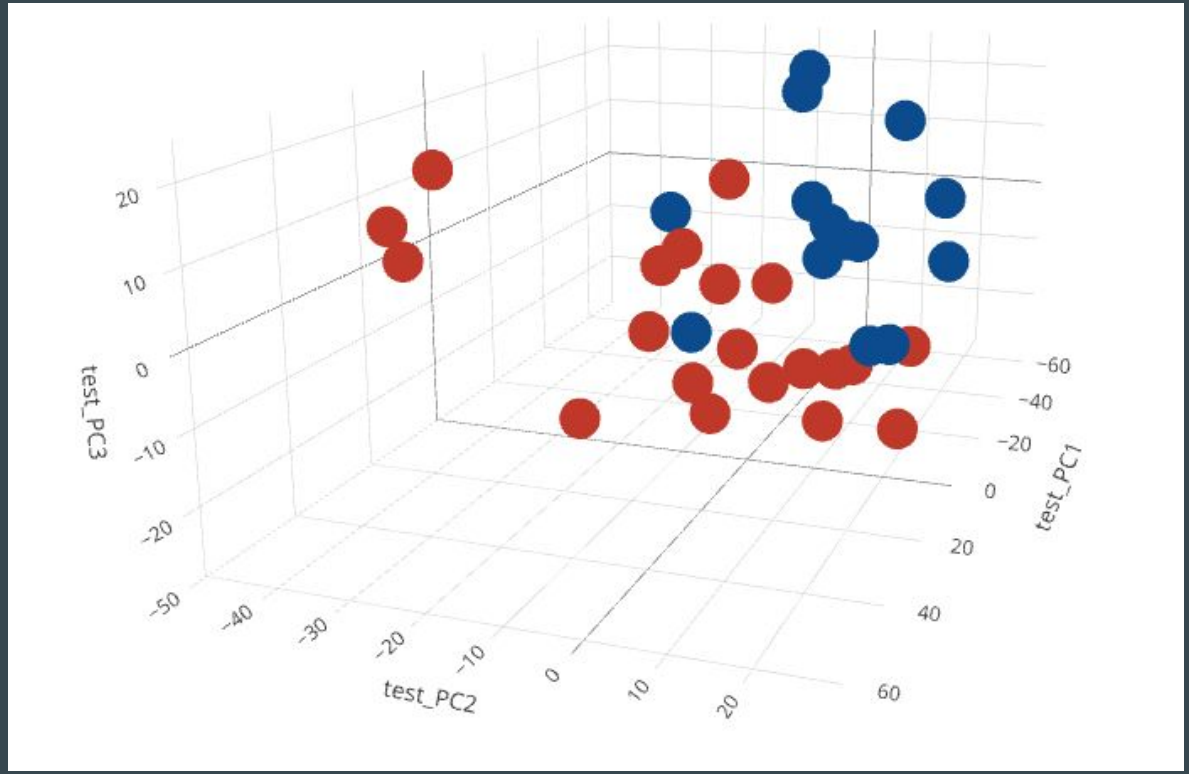
Microarrays: Reliability of Results

- Results are still marred by several technical issues that are often neglected
- Each manufacturing technique has its own weaknesses
- Different arrays have different results concerning accuracy, sensitivity, specificity and robustness
    - Claims have been made of reduced reproducibility of results from one platform to another
- These issues became much more important when microarrays were proposed as a diagnostic tool in molecular disease classification
    - Regulatory agencies such as the FDA require solid, empirically supported data about the accuracy, sensitivity, specificity, reproducibility and reliability of microarrays for clinical use

Microarray Quality Control Project (MAQC)

- FDA spawned project initiated September 2006 with the goals of
  - Provide quality control, tools to the microarray community to avoid procedural failures
  - Develop guidelines for microarray data analysis by providing the public with larger reference data sets along with readily accessible reference RNA samples
  - Establish QC metrics and thresholds for objectively assessing the performance achievable by various microarray platforms
  - Evaluate the advantages and disadvantages of various data analysis methods
- Goals of this project update periodically once the direction wished has been sufficiently explored

PCA: Principle Components Graphed



Principal Components Analysis: Motivation

- A form of dimensionality reduction would be very helpful, in which the dimensions of the data set are reduced, **while the majority of variance and the relation of patients are preserved**

$$< \text{expression level}_{gene_1}, \text{expression level}_{gene_2}, \cdots, \text{expression level}_{gene_{7000}} >$$

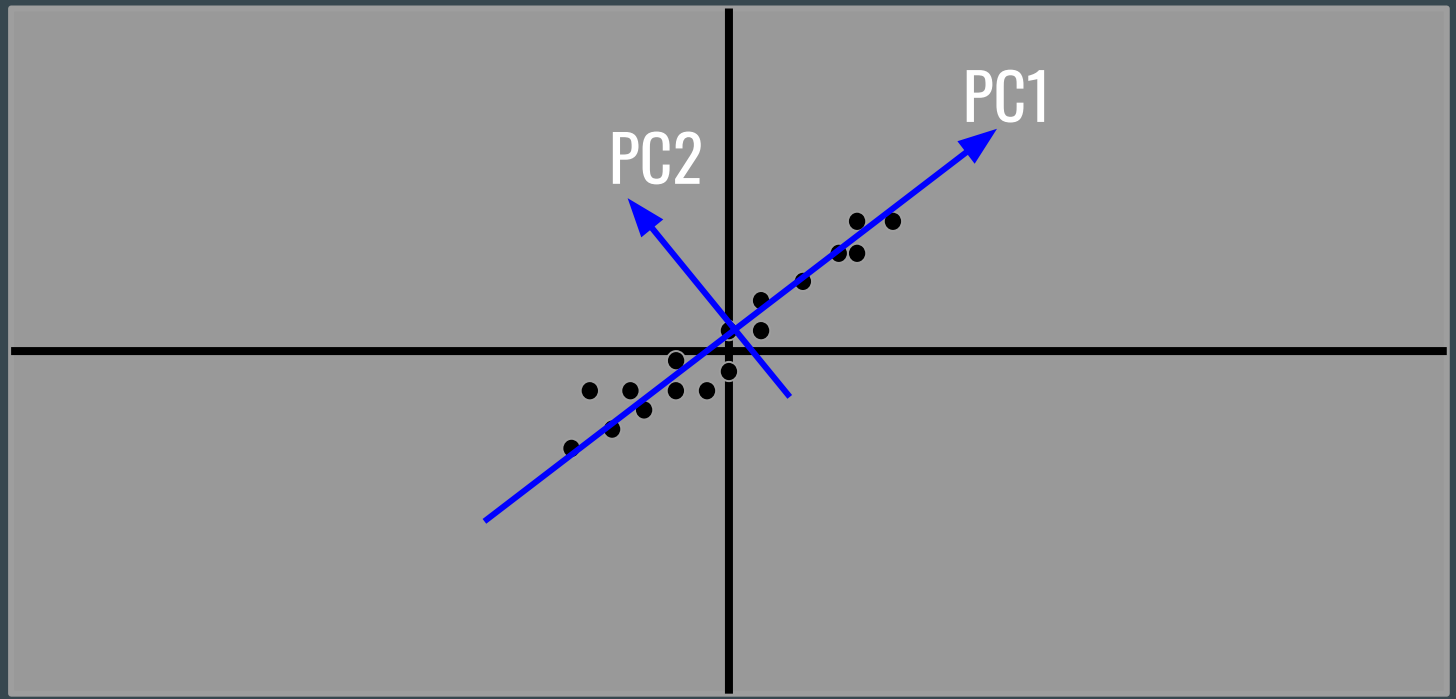
$$\Downarrow \textbf{PCA}$$

$$< \mathrm{PC}_1, \mathrm{PC}_2, \mathrm{PC}_3 >$$

Principal Components

- The result of PCA is a list of principal components, or the **directions of maximal variance in the data set**, such that all principal components are orthogonal
- Principle components are named in order of significance, being how much variability of the data they describe (PC1 , PC2 , ... , PCN)
- PCs can aid in data visualization, by plotting the data points on the new axes of the principal components

Visualization



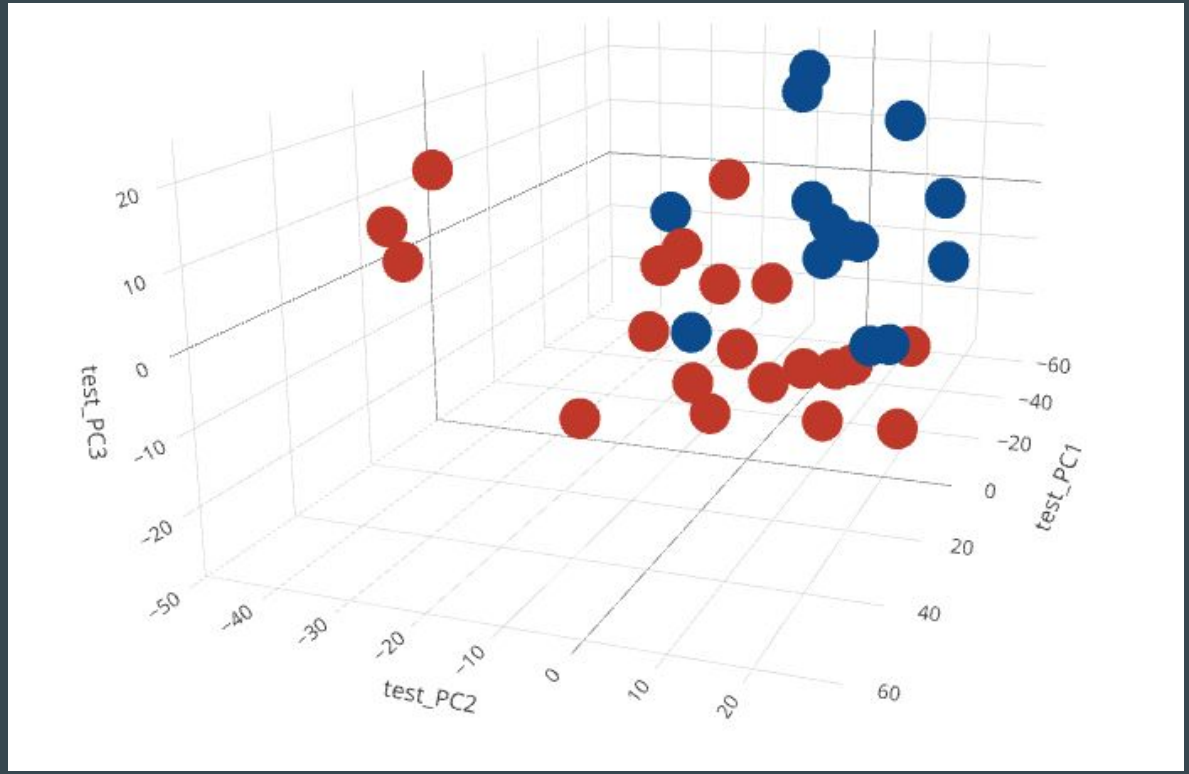
PCA - How it Works

- PCA is a procedure that uses an orthogonal transformation to reduce dimensions, and this is done by finding the **eigenvectors of the covariance matrix of the gene expression levels**
- If A is the covariance matrix of the data, then the eigenvector z_1 and eigenvalue λ_1 would form the transformation

$$Az_1 = \lambda_1 z_1$$

- The eigenvalue with the largest absolute value will indicate that the data have the largest variance along its vector

PCA: Principle Components Graphed



Possibility of Further Class Seperation

- In our goal of trying to classify the two different sets, we want the gene expression vectors of patients with different cancer types to be as separate as possible
    - From the previous visualization, it doesn't seem immediately clear that the classes are entirely separable
- A part of the problem we are seeing is that the visualized PCA results were for all ~7000 genes, where not nearly that many are responsible for the difference in conditions between the two clinical groups
- This problem can be solved by statistically evaluating the genes between the two groups, and determining which genes are **differentially expressed** between the two groups

Background: T Test

- The T test is used to test if there is a **difference in the mean expression level of a single gene** between the two groups
    - Can be applied to each gene to determine all genes that are differentially expressed between the two groups
- The test statistic for the T test:

$$T = \frac{\text{signal}}{\text{noise}} = \frac{\text{diff between group means}}{\text{variability of groups}} = \frac{|\bar{x_1} - \bar{x_2}|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

* If the test statistic falls above the critical value on the distribution, it indicates that the means of the two groups are statistically significantly different

Applying the T-test to Microarray Data

- The T-test can be useful in microarray analysis to analyze if gene expression levels are different for a single gene from two groups of individuals
- This test is not ideal, as applying the t-test with α=.05 significance level to a list of 10,000 genes will accept approximately 500 genes that appear to be differentially expressed even if they are random, just by chance

Multiple Comparisons: Family Wise Error Rate

- Family-Wise Error Rate (FWER) - Probability of having a Type 1 error (false positive) in any comparison in the data set
- We would like to control this overall rate of false positives, and there are some common ways to do so
    - The Bonferroni and Šidák corrections for multiple comparisons are common multiple comparison corrections, though in our case of thousands of tests are too conservative
    - Even a more robust technique called FDR tends to be too conservative

Selecting DE Genes: SAM

- A special technique just for the application to Microarrays was designed by a group at Stanford called **Significance Analysis of Microarrays** (SAM)
  - Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu
- Combines classical hypothesis testing approaches with a re-sampling or bootstrapping approach and tends to make the test less conservative while also taking the dependencies into account
  - Taking the dependencies into account is very important as some gene expression levels **directly impact** the expression levels of other genes

Preprocessing: Fold Change Ratio

- **Fold Change Ratio** is the simplest and most intuitive approach to finding genes that are DE

$$\left| \frac{\text{mean group 2 expression - mean group 1 expression}}{\text{mean group 1 expression}} \right| \geq 1.5$$

- This method often used because it is simple and intuitive, however it has important disadvantages
    - The fold threshold is chosen arbitrarily and may often be inappropriate (no genes or non-DE genes can be selected)

Hypothesis Testing: SAM

- Microarrays tend to have low signal to noise ratios for genes with low expression levels (low intensity values have higher variance, high intensity values have lower variance)
    - Constant fold change threshold for all genes will introduce false positives at low expression levels and miss true positives at high expression levels
- SAM calculates a gene-by-gene variance which will allow for the selection of the appropriate genes independently of their expression levels
- In order to avoid as many hypothesis tests as possible, sam has a **filtering step** before any analysis, removing genes that have a fold-change ratio less than a provided threshold

Hypothesis Testing: SAM

- The test statistic used in SAM is the relative difference statistic:

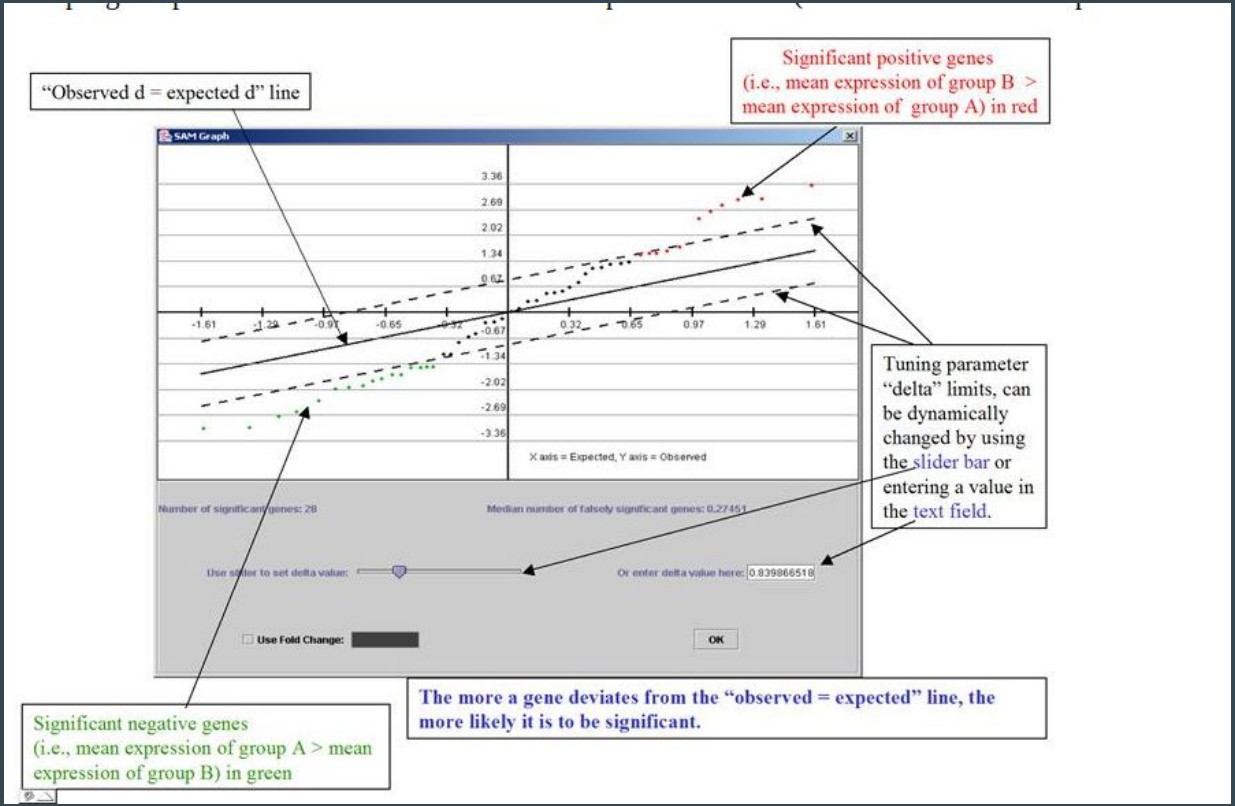
$$d(i) = \frac{\overline{x}_1(i) - \overline{x}_2(i)}{s(i) + s_0}$$

where $\overline{x}_1(i)$ is the mean expression of class ***1***, $\overline{x}_2(i)$ is the mean expression of class ***2***, $s(i)$ is the standard deviation of the gene considering both classes, and s_0 is the "fudge factor", **used to dissociate the variance of $d(i)$ with the expression level of gene *i***

Controlling for FDR in SAM

- SAM uses **permutations of the data** to estimate the percentage of genes identified just by chance
- FDR in SAM:
  - Fix a threshold for differentially expressed genes
  - Count how many genes are reported as differentially expressed in each random permutation of the data
  - Calculate the median number of false positives across all permutations
  - Calculate the FDR as the median number of false positives divided by the number of genes selected as differentially expressed

SAM Results



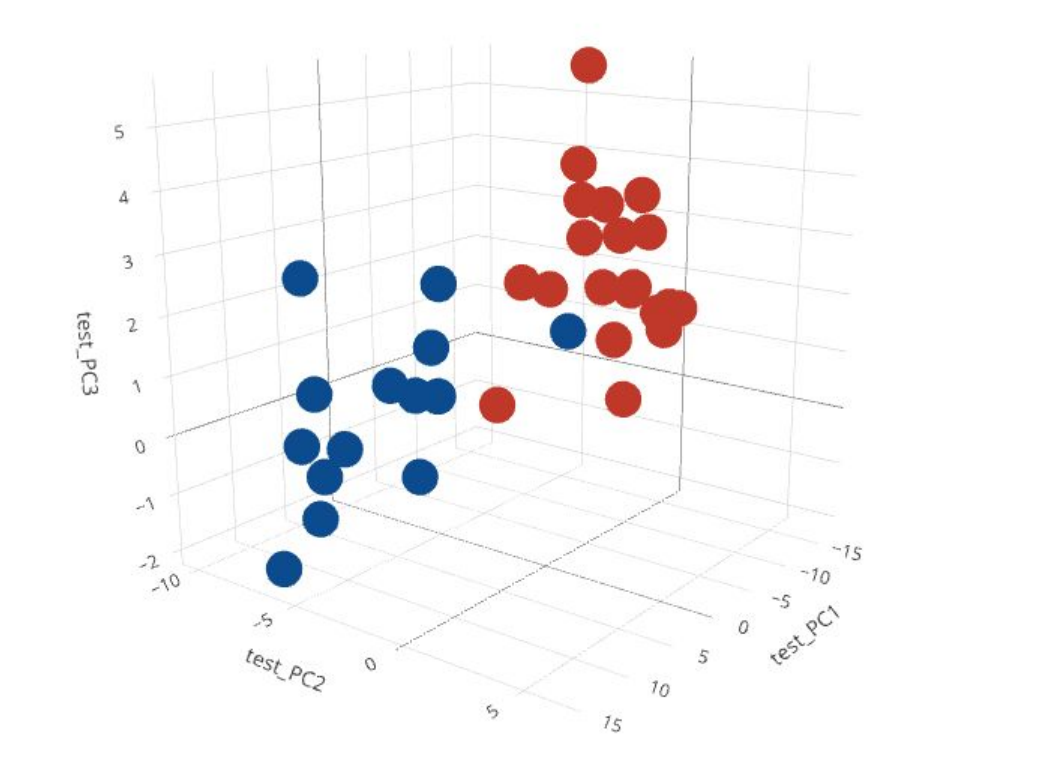
[7]

SAM Results

- The way to control for FDR in SAM is through the **delta** parameter
- SAM results allow for a table to be printed that shows how many genes are selected, and various FDR statistics associated with a particular delta value

```
> delta.table[, c(1,4,5,6)]
          delta # called  median FDR 90th perc FDR
 [1,] 0.000000000     6075 0.698307579    0.72730048
 [2,] 0.001398253     6074 0.697802938    0.72692453
 [3,] 0.005593013     6054 0.696129606    0.72572040
 [4,] 0.012584279     6020 0.693434472    0.72431769
 [5,] 0.022372052     5919 0.688862509    0.72268888
 [6,] 0.034956332     5802 0.683164332    0.72052691
 [7,] 0.050337118     5703 0.676347952    0.71694599
 [8,] 0.068514410     5577 0.668886925    0.71369533
 [9,] 0.089488210     5447 0.660944601    0.71013889
[10,] 0.113258515     5258 0.651490830    0.70660486
[11,] 0.139825328     5101 0.641292930    0.70193982
[12,] 0.169188646     4692 0.610085040    0.68580425
[13,] 0.201348472     4328 0.577394822    0.66765624
[14,] 0.236304803     3754 0.534950516    0.64823651
[15,] 0.274057642     3379 0.492852701    0.62216388
[16,] 0.314606987     3040 0.451124946    0.59473223
[17,] 0.357952838     2695 0.404419275    0.56110381
[18,] 0.404095197     2423 0.357245264    0.52219938
[19,] 0.453034061     2128 0.311267016    0.48249925
[20,] 0.504769432     1889 0.267768428    0.44309299
[21,] 0.559301310     1722 0.229481780    0.39912345
[22,] 0.616629694     1460 0.189206156    0.35675929
[23,] 0.676754585     1245 0.154167519    0.31444128
[24,] 0.739675983     1084 0.124987124    0.27358293
[25,] 0.805393886      865 0.096589279    0.23494690
[26,] 0.873908297      736 0.075679100    0.19942466
[27,] 0.945219214      589 0.057506809    0.16613078
[28,] 1.019326638      481 0.042251365    0.13301356
[29,] 1.096230568      394 0.032476914    0.11080359
[30,] 1.175931004      318 0.023669819    0.08994531
[31,] 1.258427948      274 0.016482487    0.07142411
```

PCA: Principle Components Graphed



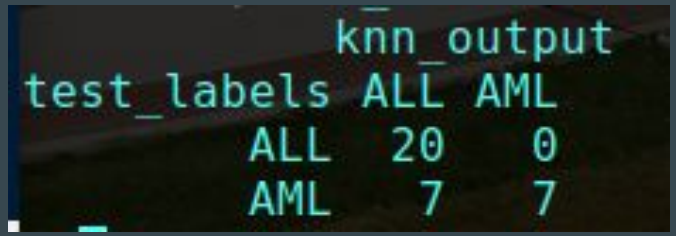
Machine Learning: K-Nearest Neighbor Classifier

In order to classify a new patient by their genetic sample:

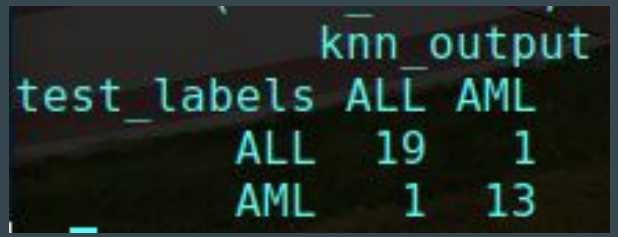
1. The distance between the new patient and all other patients in the training set are calculated (based on the principal components of the previously found differentially expressed genes)
2. The samples are then ordered closest to furthest from the new patient and the k closest samples are retained
3. The class that is most highly represented in the k closest neighbors is chosen to be predicted class of the new patient

Application: KNN

Top 3 PCs of All Genes



Top 3 PCs of DE Genes



Questions?

References

[1] Drăghici Sorin. Statistics and Data Analysis for Microarrays: Using R and Bioconductor. Chapman and Hall, 2012.

[2] Vong, Andrew. “3D DNA Strand Rendering in MAYA.” av_designs, 27 Mar. 2015, www.andrewvong.com/3d-dna-strand-rendering-in-maya/.

[3]“Ribonucleic Acid (RNA).” RNA Extraction, Quinnipiac University, qu.edu.iq/bt/wp-content/uploads/2015/12/RNA-extraction.pdf.

[4] charity-stanton. “What Organic Molecule Is DNA?” SlideServe, 12 Nov. 2014, www.slideserve.com/charity-stanton/what-organic-molecule-is-dna.

[5] Catherine. “Role of MRNA in Protein Synthesis.” Study.com, Study.com, 2018, study.com/academy/lesson/role-of-mrna-in-protein-synthesis.html.

[6] Klimas, Lisa. “Gene Expression and the D816V Mutation.” Mast Attack, www.mastattack.org, 9 Aug. 2017, www.mastattack.org/2014/06/gene-expression-and-the-d816v-mutation

[7] Cui, Yan. “Data File: Stanford_Large.Txt.” Microarray Data Analysis II, compbio.uthsc.edu/microarray/lecture2.htm.]

Biological Background

- **Deoxyribonucleic acid** (DNA) is the building block of life and is responsible for carrying instructions via units of **genes**, which regulate growth, development, function and reproduction of cells
- DNA follows a base-4 code and is composed of **oligonucleotides**, also known as **bases**
  - [C] - Cytosine
  - [G] - Guanine
  - [A] - Adenine
  - [T] - Thymine
- Not all pairs are possible though, the only possible **base pairs** are one of the pairs A-T or C-G

Biological Background

- These bases cannot just pair and form strands on their own though, they must be attached to a phosphate to allow for this
    - When a base is paired with a phosphate, we now call be base a **nucleotide**
- Now we can speak of **genes**, being contiguous sequences of nucleotides capped with **start and stop codons** to demarcate the beginning and end of the gene

[3]

Background - Covariance Matrix

- A covariance matrix of a data set is a matrix that expresses how each feature varies with every other feature
    - As such it captures the shape of the data set

$$\begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \dots & cov(x_1, x_n) \\ cov(x_1, x_2) & cov(x_2, x_2) & \dots & cov(x_2, x_n) \\ . & & & . \\ . & . & & . \\ . & & . & . \\ cov(x_1, x_n) & \dots & & cov(x_n, x_n) \end{bmatrix}$$

PCA in R

- Two functions are available for PCA in R, being **prcomp()** and **princomp()**
  - Difference in how PCA is calculated
- **Princomp()** follows the method that we have described
- **Prcomp()** uses a process called singular value decomposition (svp), with very little difference in output, and is a preferred computational method

PCA in R

- **Prcomp()** options:
  - center: 0-center the data
  - scale: scale the data to have unit variance
- Usage:

```
gene_pca = prcomp(cleaned_data, scale=T, center=T)
```

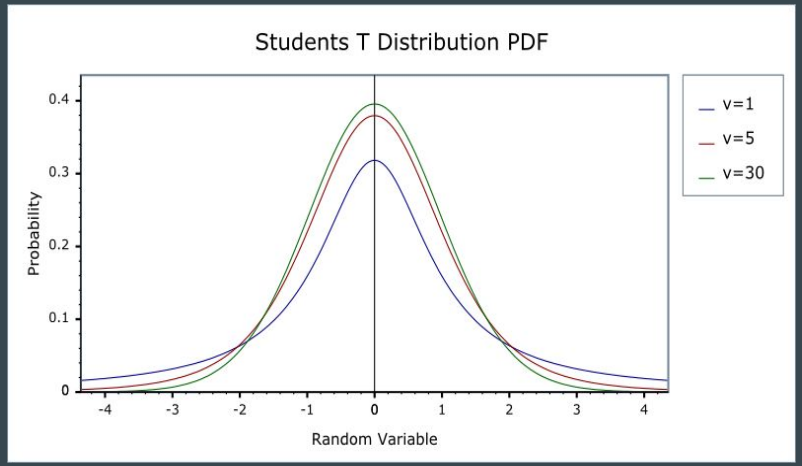
where cleaned_data is a matrix of samples x features

Background: Hypothesis Testing

- Hypothesis Test - A statistical test in which a null hypothesis (H_0) is assessed through the use of a test statistic, comparing this test statistic to a critical value
  - In most cases if the test statistic is larger than the critical value, then H_0 is rejected.
- P Value - The probability of rejecting a null hypothesis H_0 when it is true due to random chance / sampling error
- Significance Level - A significance level α is the acceptable p value such that H_0 is rejected in a statistically significant manner
- Critical Value - A quantity derived from the statistical distribution of the test statistic and the accepted p value.
- Test Statistic - A quantity derived from a sample, compared to a critical value to test for the truth value of the null hypothesis

Background: T Test

- Student's T distribution - A statistical distribution that can be used in comparing means of normally distributed data sets in situations with unknown standard deviation σ and a small sample size. This distribution comes from dividing a Normal Distribution with a Chi-Square Distribution



[5]

Multiple Comparisons: Corrections

- Bonferroni notes that for small *p*, Eq. 4 can be approximated by taking only the first two terms of the binomial expansion of $(1-p)^R$, resulting in the Bonferroni correction for multiple comparisons

$$\alpha_c = \frac{\alpha_e}{R} \quad \text{(Eq. 6)}$$

- Unfortunately, both mentioned corrections require a very small significance level for any reasonable R value.
- These corrections are sufficient but not necessary for declaring a gene differentiable between control and patient groups

Multiple Comparisons: Step-Wise Correction

- The Holm's step-wise group of methods allow less conservative adjustments of the *p*-values, ordering genes in increasing order of their *p*-value and making successive smaller adjustments
- Procedure:
  - Choose the experiment-level significance α_e
  - Order the genes in the increasing order of individual *p*-values
  - Compare the *p*-values of each gene with a threshold that depend on the position of the gene in the list of ordered values. The thresholds are as follows:

    $\frac{\alpha_e}{R}$ for the first gene, $\frac{\alpha_e}{R-1}$ for the second gene, and so on
  - Let *k* be the largest *i* for which $p_i < \frac{\alpha_e}{R-i+1}$. Reject the null hypothesis for *i* = 1, 2, ... , k

Multiple Comparisons: False Discovery Rate (FDR)

- The FDR correction procedure allows for some dependencies between variables, while the previous methods act on the assumption of independence
- FDR correction procedure:
  - Choose the experiment-level significance α_e
  - Order the genes in the increasing order of individual *p*-values
  - Compare the *p*-values of each gene with a threshold that depends on the position of the gene in the list of ordered values. The thresholds are as follows:

    $\frac{1}{R}\alpha_e$ for the first gene, $\frac{2}{R}\alpha_e$ for the second gene, and so on
  - Let *k* be the largest *i* for which $p_i < \frac{i}{R}\alpha_e$. Reject the null hypothesis for *i* = 1, 2, ... , k

Multiple Comparisons: Permutation Correction

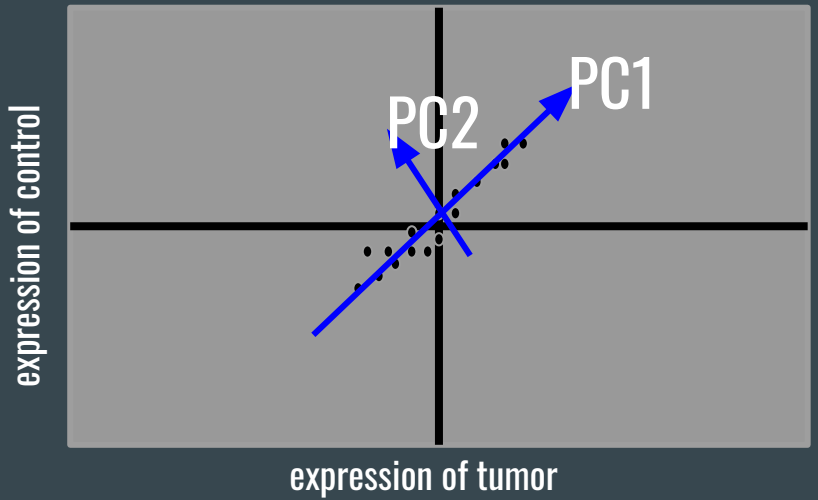
- The Westfall and Young (W-Y) step-down correction is a more general method that adjusts the *p*-value whole taking into consideration the possible correlations
- This method permutes the classes individuals thousands of times, running a T-test after every permutation
- The *p*-value for a gene *i* will the the proportion of times the value of *t* calculated for the real labels is less than or equal to the value of *t* calculated for a random permutation

$$p\text{-value for gene } i : \frac{\text{number of permutations for which } t_j^{(b)} \geq t_i}{\text{total number of permuatations}}$$

where $t_j^{(b)}$ are the calculated t-values from gene j and permutation b

Caution

- In Microarray analysis and other fields, variance along one axis may be expected
- In DNA Microarrays, this expected variance comes along in the form of expression level, it is known that different genes will express at different levels, but we are interested in the **ratio** of the expression levels in comparative analysis



Hypothesis Testing: Gene Filtering

- Common filters ensure that
  - There are at least a few samples in the group that have significant expression values
  - The ratio of maximum / minimum intensity is at least 1.5 (the gene is variable)

Moderated t-Statistic

- Another possible method to selecting differentially expressed genes is to make the assumption that all genes share a common variance
- Under this assumption, a pooled variance can be calculated
    - Given that there is now more observations to estimate this common variance, our estimate will now be much more reliable
    - An Empirical Bayes model can be used
- However, there is not an overall consensus that this is a valid assumption to make, and the originating paper for SAM cited the opposite case, being that variance is not even just expression level dependent, but entirely different on a gene by gene basis

$$(\text{Expression levels of all genes}) \in \mathbb{R}^{7000} \Rightarrow (\text{Principal Components}) \in \mathbb{R}^3$$

Hypothesis Testing: SAM

- Uses a "relative difference" statistic d_i for gene i
  - Very similar to a t statistic with equal variance, except that the "gene-specific scatter" (std. dev. of the difference) s_i in the denominator is offset by a "fudge factor" s_0
  - This s_0 is an "exchangeability" factor chosen to minimize the coefficient of variation of d_i
  - **This small positive exchangeability factor stabilizes d_i for genes with low expression levels**
- A permutation test is used to assess significance of d_i, as well as estimate the FDR