Accuracy of PCA for Cancer detection applied to micro array data

by Nasser Abbasi

Project supervisor: Dr C.H. Lee

Mathematics department, CSUF

Goal of the study

- Apply a mathematical technique for pattern recognition called Principal Component Analysis (PCA) to the detection of primary liver and bladder cancer using actual medical micro array data obtained from public databases such as Stanford SMD and NCBI GEO.
- Evaluate accuracy of PCA in detection of primary liver and bladder cancer.

Phases of using PCA

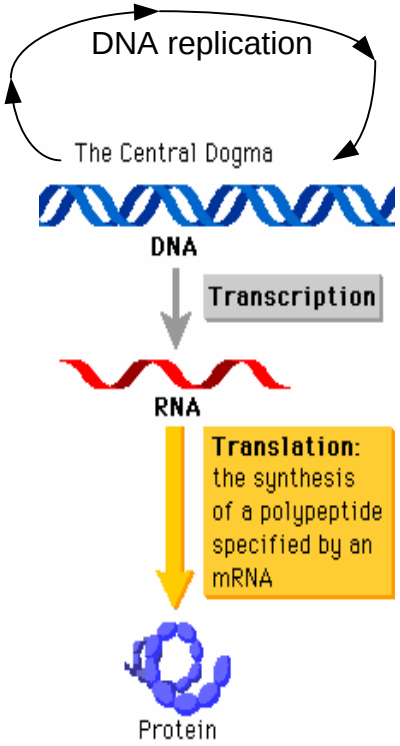
1. The first phase: Called the training phase. We use PCA to obtain the *dominant signal* from a collection of signals. This collection of signals will be called the PCA working set. The dominant signal is the one which correlates the most with all the signals in the working set. This signal will be called the eigensignal.
2. The second phase: Called the detection phase. Determine the projection of the input signal against the eigensignal. Is the projection positive or negative? How large is the projection?
3. Validate result for accuracy: Knowing the correct type of the input signal, determine the accuracy of the detection phase.

Before going into the details of the project and the mathematics of PCA, we take a 2 minutes break and give a short introduction about DNA and Genes

Why genes are important

Central dogma of molecular biology

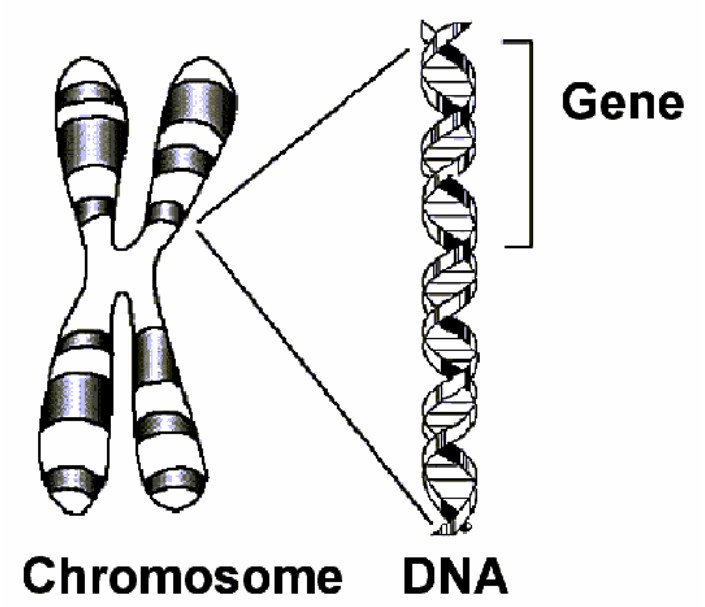
DNA makes RNA makes Protein



http://phschool.com/science/biology_place/biocoach/imagestranslation/centdogtl.gif

Genes are special regions in DNA

- Each human cell contain 46 chromosomes.
- Each chromosome is of different length.
- Total of 3 billion base-pairs in the DNA spread among the 46 chromosomes.
- A Gene is special region in the DNA which is able to encodes protein



Some facts about DNA

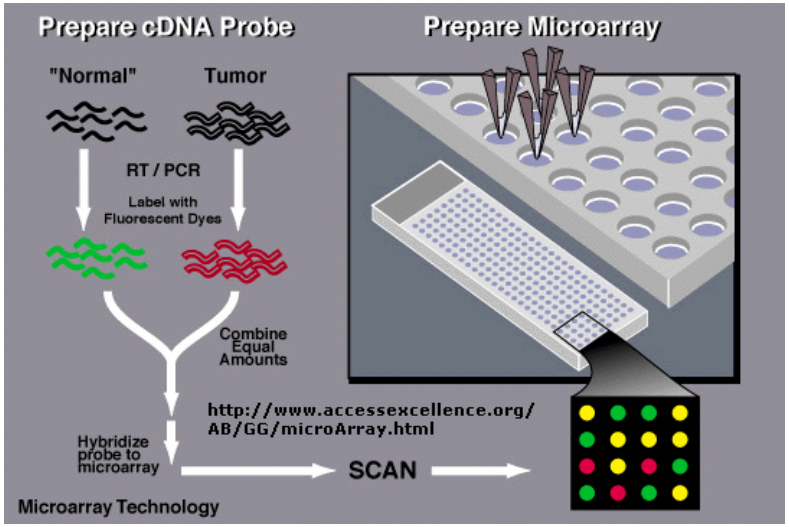
- DNA is the main central molecule from which all the cell functions originate.
- Each human cell contain the same DNA.
- Each human cell contain about 3 billion base pairs, which are spread out in 46 different chromosomes.
- The chromosomes are not all the same size, some are much longer than others. There are 6 billion nucleotides in each human cell. (each base-pair is 2 nucleotides).
- When a human cell divides and new cell is created, a new 6 billion nucleotides are made in the process.
- The human body contain large amount of cells, some estimate is at 100 trillion cells. Hence the human body contain in it 100 trillion times 6 billion nucleotides or 600,000,000,000,000,000,000,000 nucleotides.

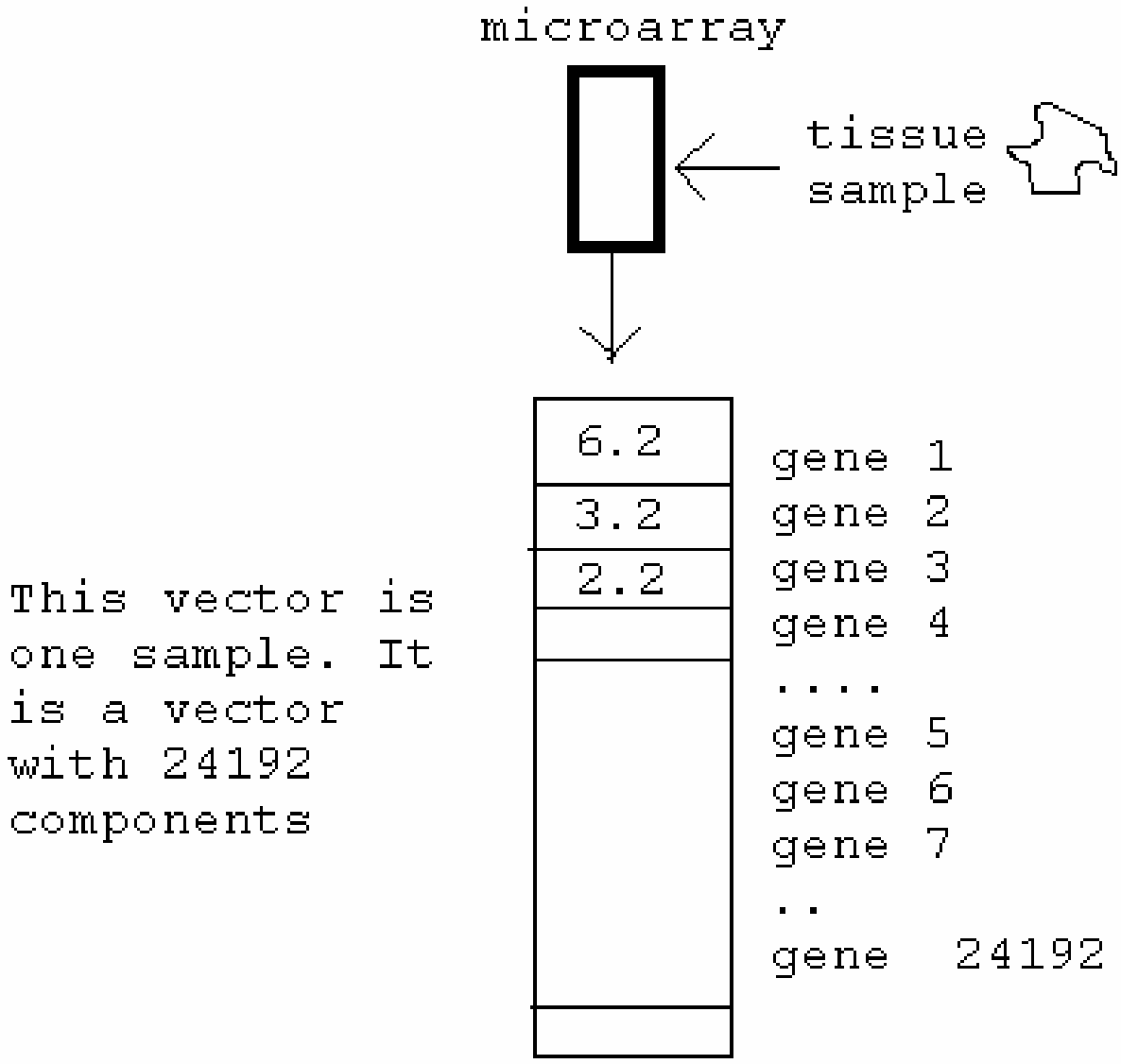
Genes and cancer

- Cancer occurs when human cells divide and duplicate without control. In a normal cell, specific protein and enzymes control the life cycle of a cell by controlling the production of new cells. Since each specific protein is made by specific gene(s), knowing which genes are on or off in a cancerous cell, and how active that gene is gives an indication of the gene role in the cancer.
- By obtaining a sample from part of the body which is known to have cancer, and if by some process we are able to determine which genes are turned on and how active these genes and then compare these genes activities in a cancer-free sample of the same part of the body, we can then predict if these genes contributed to the cancer or note.

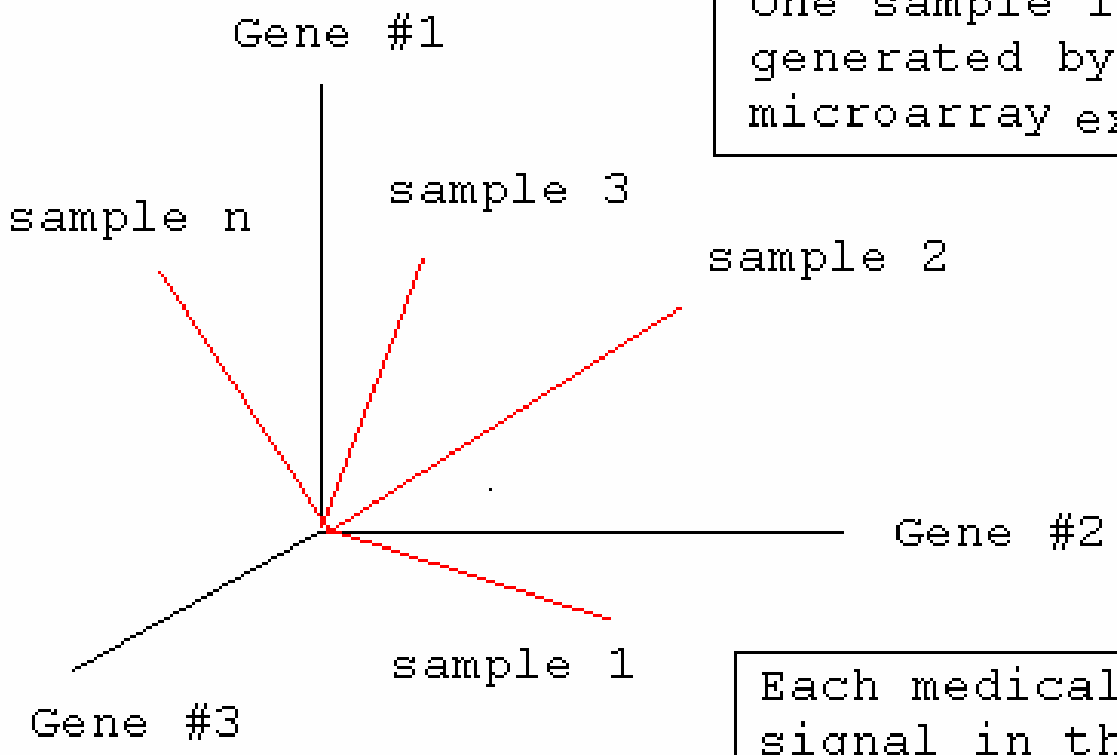
What is micro array?

- A plate which contain collection of spots. Some plates have up to 24,000 spots on them
- When plate is manufactured, in each spot a specific cDNA probe is fixed which attracts an mRNA for a specific gene.
- Sample to be analyzed is poured into the plate.
- mRNA from the sample swims and finds the correct spot to attach to.



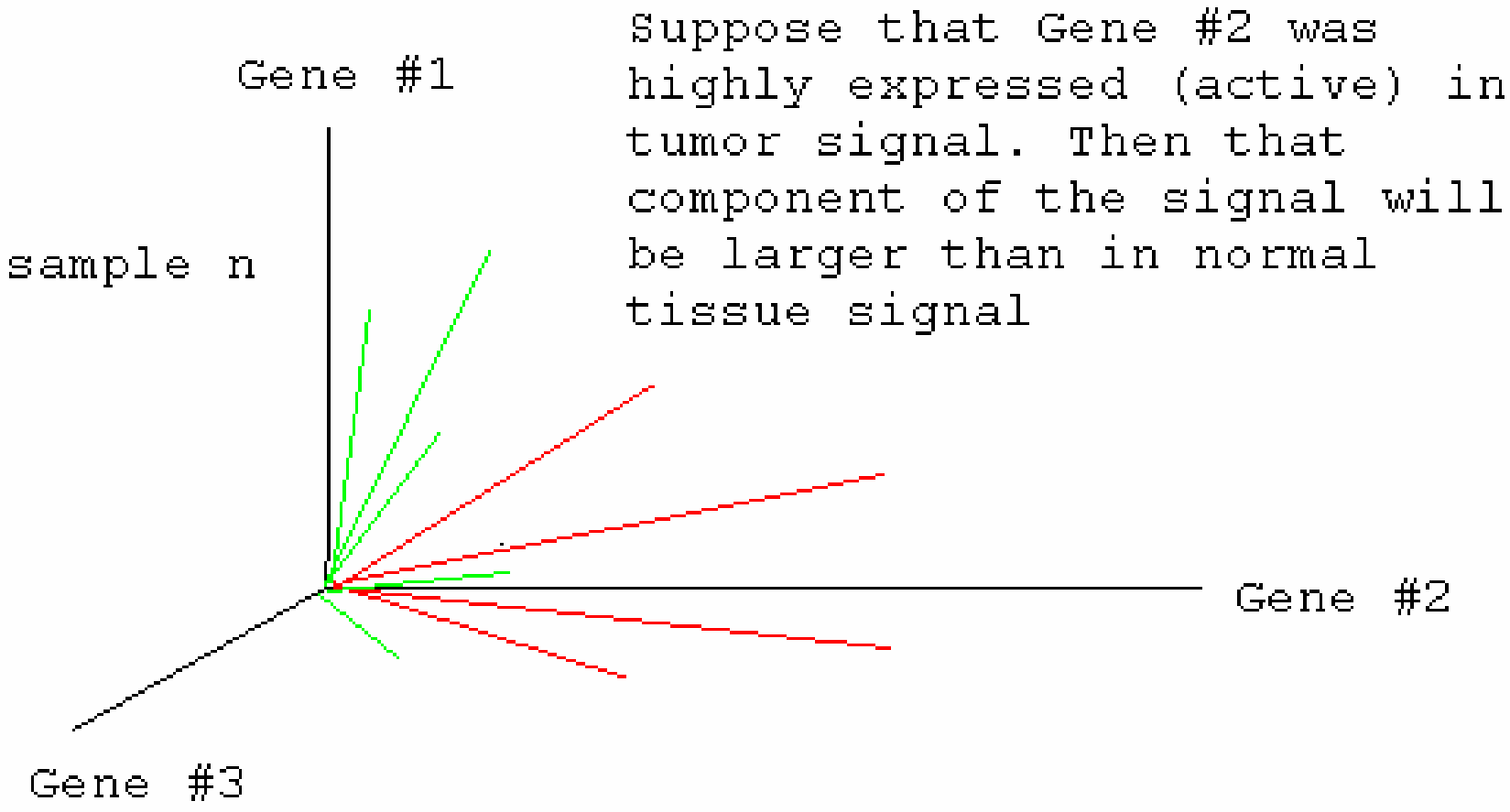

microarray
tissue
sample
6.2
3.2
2.2
gene 1
gene 2
gene 3
gene 4
....
gene 5
gene 6
gene 7
..
gene 24192
This vector is one sample. It is a vector with 24192 components

One microarray generates one signal/vector in the space of genes

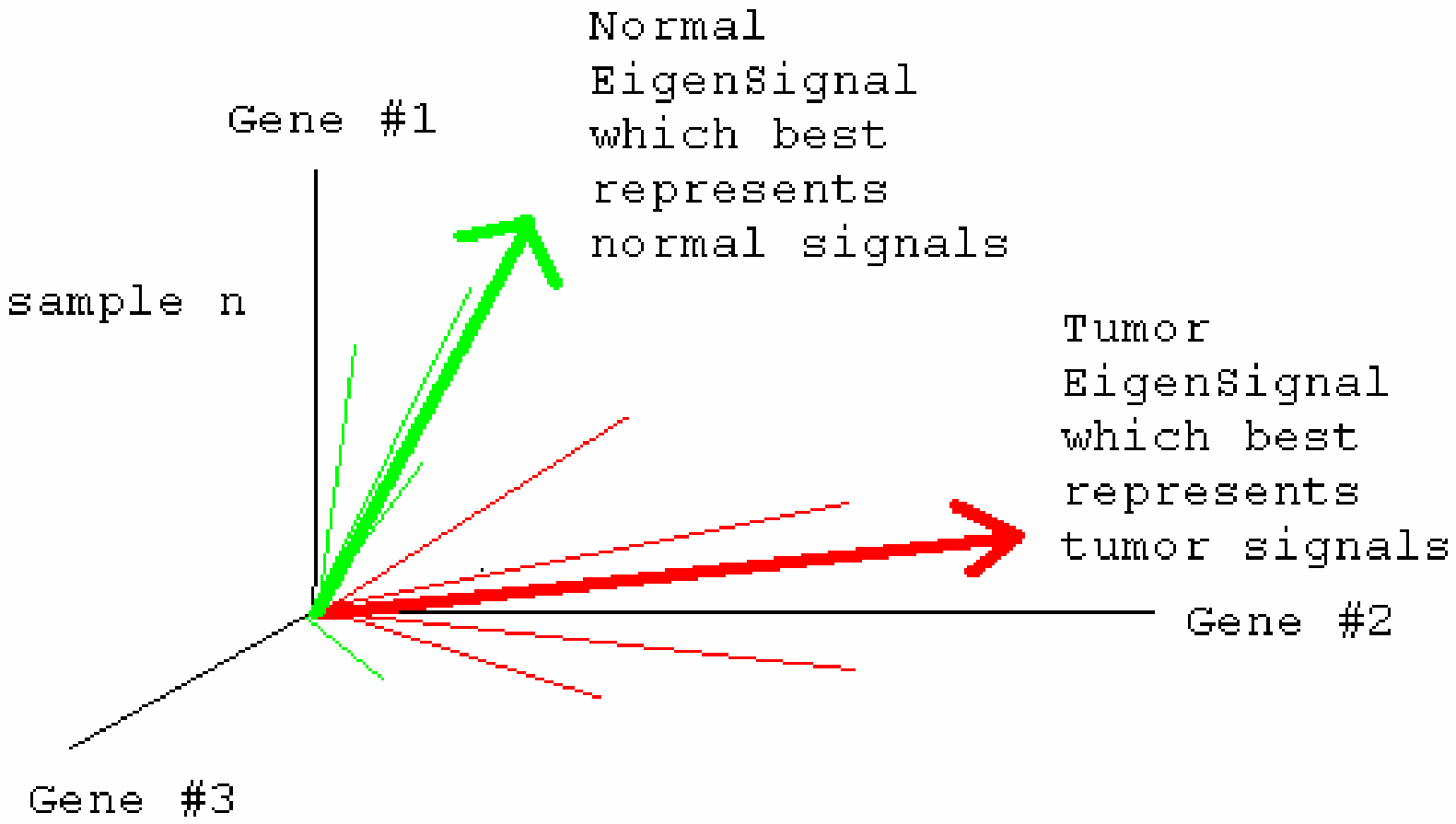


One sample is generated by one microarray experiment

Each medical sample is a signal in the space of human genes. The dimension of this space is about 30,000 genes (the number of genes in a human cell DNA)



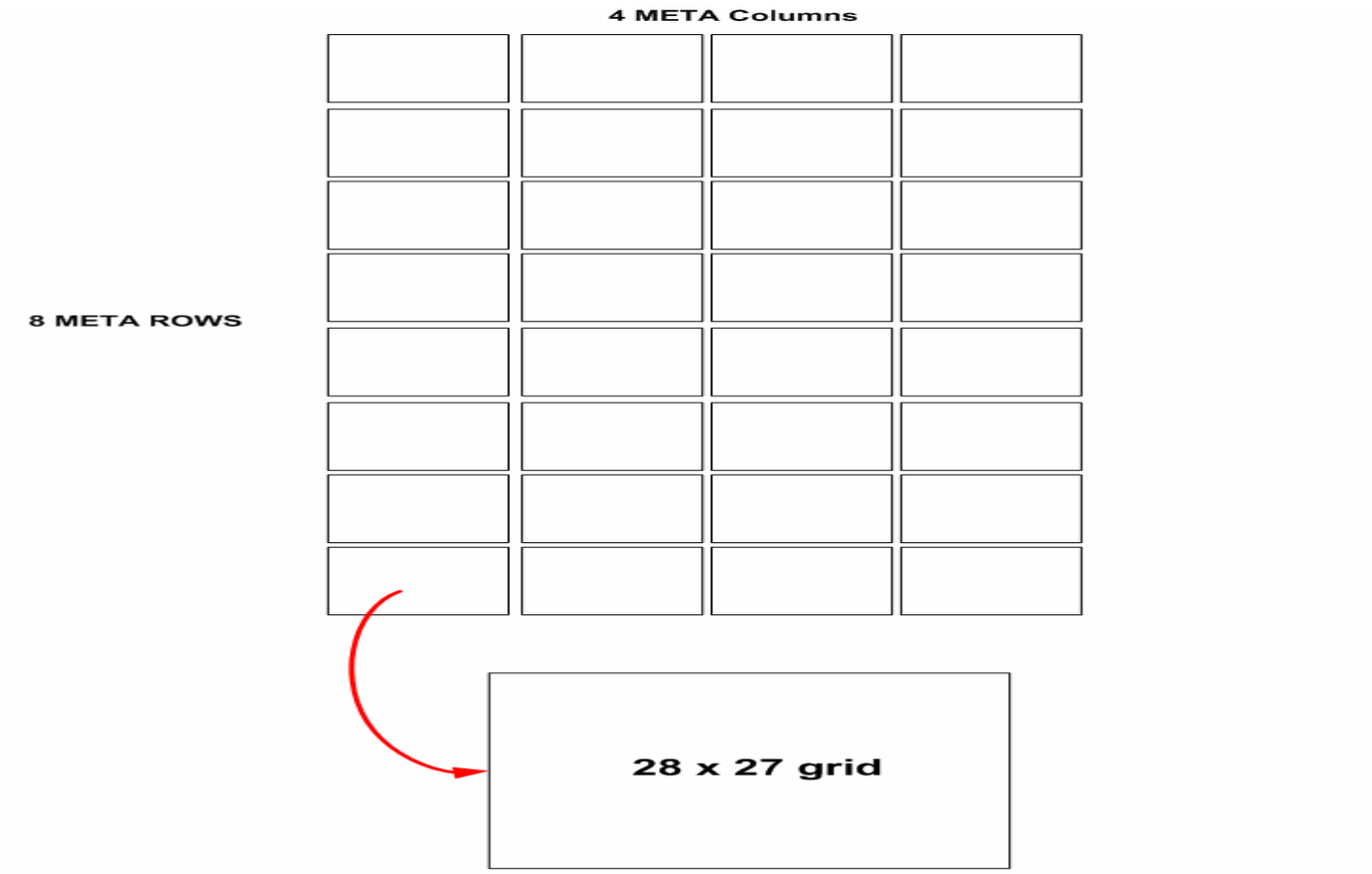
Suppose that Gene #2 was highly expressed (active) in tumor signal. Then that component of the signal will be larger than in normal tissue signal



Finding PCA dominant modes from population samples

Newer microarray technology contain the whole human genom on a chip

Layout of microarray used for project



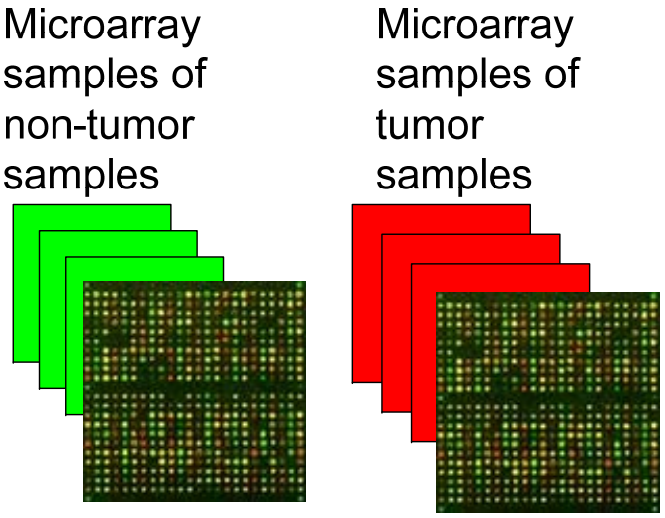
Chen Xin et all, 2002 paper, microarray physical layout.
Total number of spots (4x8)x(28x27)= 24,192 genes

Use PLATFORM record associated with Sample (GSM) files to track spot locations. Use the above to regenerate microarray images

By Nasser Abbasi
040607
Platform_description.vsd

Data used in project

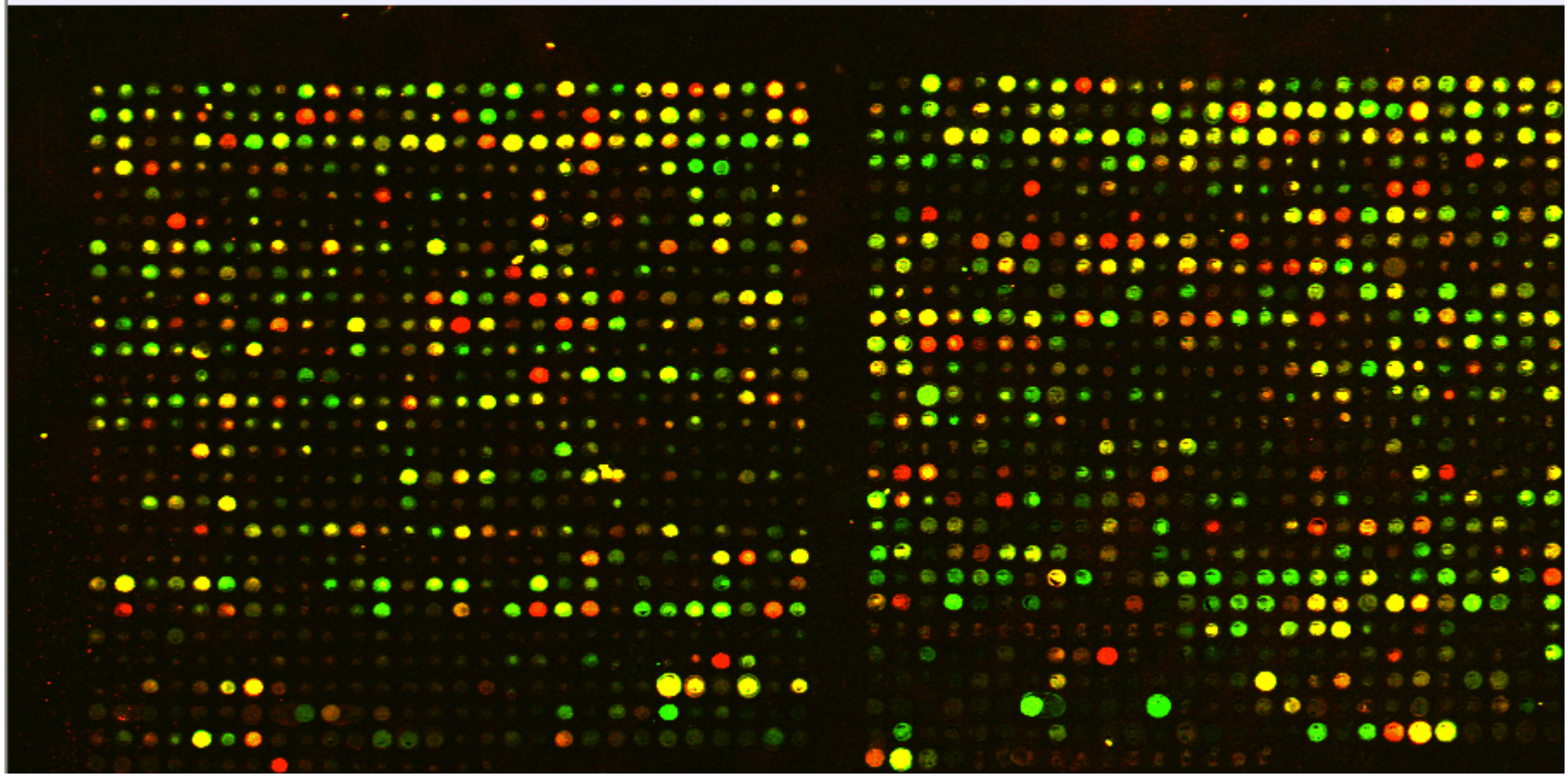
- Bladder and Liver data.
- Samples which contain micro array data from known normal and cancerous bladder and liver data.
- Each sample is a vector of length n. (in this study, length was about 6,000 genes)
- Each component of a vector is a value which represent how much a specific gene is expressed in the sample

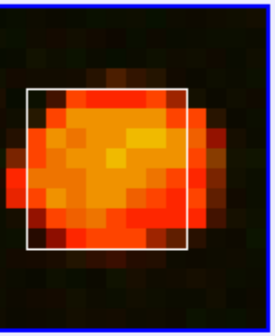


Address http://smd.stanford.edu/cgi-bin/search/QuerySetup.pl

10029	Adenoma (HK69)	Adenoma	Liver	shan092	default		XINCHEN
10018	Adenoma (HK73)	Adenoma	Liver	shan095	default		XINCHEN
7350	Adenoma (SF25)	Adenoma	Liver	shak109	default		XINCHEN
20974	BMP4 12hrs	ES cells	cell line	shbq090	default		XINCHEN
23594	BMP4 24hrs	ES cells	cell line	shcq166	default		XINCHEN
20976	BMP4 3days	ES cells	cell line	shbq092	default		XINCHEN
20152	BMP4 3hrs	ES cells	cell line	shbq024	default		XINCHEN
16552	BMP4 48hrs	ES cells	cell line	shat107	default		XINCHEN

"Adenoma (HK69)"





k spot to see in array context

ogical Information	
ne ID	IMAGE:773344
e Symbol	SLC16A2
e Name	Solute carrier family 16, member 2 (monocarboxylic acid transporter 8)
ster ID	Hs.75317
ession	AA425395
uslink ID	6567

expression history of this entity

Spot	72
SUID	118449
Ch1 Intensity (Mean)	441
Ch1 Background (Median)	53
Ch1 Net (Mean)	388
Ch2 Intensity (Mean)	3814
Ch2 Net (Mean)	3736
Ch2 Background (Median)	78
Ch2 Normalized Background (Median)	59
Ch2 Normalized Net (Mean)	2830
Ch2 Normalized Intensity (Mean)	2889
Regression Correlation	.98
Spot Flag	0
% CH1 PIXELS > BG + 1SD	100
% CH2 PIXELS > BG + 1SD	100
Regression Ratio	9.091
Number of Background Pixels	263
Number of Spot Pixels	52
Box Top	75
Box Left	273
Box Right	281
Box Bottom	83
Sum of Ch2 Foreground Pixel Intensities	
Sum of Ch1 Foreground Pixel Intensities	
Log(base2) of R/G Normalized Ratio (Mean)	2.867
G/R (Mean)	.104

Ch1 Intensity (Median)	
Ch1 Net (Median)	
% of saturated Ch1 pixels	
Std Dev of Ch1 Intensity	
Std Dev of Ch1 Background	
Ch2 Net (Median)	
Ch2 Intensity (Median)	
% of saturated Ch2 pixels	
Std Dev of Ch2 Intensity	
Std Dev of Ch2 Background	
Normalized Ch2 Net (Median)	
Normalized Ch2 Intensity (Median)	
Diameter of the spot	
R/G Mean (per pixel)	
R/G Median (per pixel)	9.85
% CH1 PIXELS > BG + 2SD	
% CH2 PIXELS > BG + 2SD	
R/G (Median)	
Std Dev of pixel intensity ratios	
Sum of mean intensities	
Sum of median intensities	
X coordinate (whole array, in microns)	
Y coordinate (whole array, in microns)	
Log(base2) of R/G Normalized Ratio (Median)	
R/G Normalized (Median)	
Ch2 Signal-to-Noise Ratio	
Ch1 Signal-to-Noise Ratio	

Mathematics of PCA

How to find the eigensignal
(dominant signal)

How to obtain the eigen signal?

Let $\vec{V}_i$ be a collection of signals. Called snapshots $\{V_1, V_2, \cdots, V_N\}$

Let $\vec{\Phi}$ be the eigen signal that we wish to generate.

It is natural to assume

$$\vec{\Phi} = \sum_{i=1}^{N} w_i \vec{V}_i$$

where w_i are scalar quantities called the weights. Our goal is to find these weights.

Since $\vec{\Phi}$ is the signal which is the most representative of $\{V_1, V_2, \cdots, V_N\}$, hence $\vec{\Phi}$ must be such that it makes the following sum a maximum

$$S = \sum_{i=1}^{N} \left(\left\langle \vec{V}_i, \vec{\Phi} \right\rangle \right)^2$$

Hence we want to maximize S where

$$S = \sum_{i=1}^{N} \left\langle \vec{V}_i, \vec{\Phi} \right\rangle^2 = \sum_{i=1}^{N} \left(\left\langle \vec{V}_i, \sum_{j=1}^{N} w_j \vec{V}_j \right\rangle \right)^2$$

But

$$\langle a, \{b + c + \cdots\} \rangle = \langle a, b \rangle + \langle a, c \rangle + \cdots$$

Hence

$$S = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} w_j \left\langle \vec{V}_i, \vec{V}_j \right\rangle \right)^2$$

Carrying out the above multiplication will lead to

$$S = \vec{w}^T [\theta]^2 \vec{w}$$

Where $[\theta]$ is the matrix of covariance between each pairs of signals $\left\langle \vec{V}_i, \vec{V}_j \right\rangle$

To see that the above is true, let us work it out for 2 signals $\vec{V}_1, \vec{V}_2$

$$
\begin{aligned}
S &= \sum_{i=1}^{2}\left(\sum_{j=1}^{2} w_j \left\langle \vec{V}_i, \vec{V}_j \right\rangle\right)^2 \\
&= \sum_{i=1}^{2}\left(w_1 \left\langle \vec{V}_i, \vec{V}_1 \right\rangle + w_2 \left\langle \vec{V}_i, \vec{V}_2 \right\rangle\right)^2 \\
&= \left(w_1 \left\langle \vec{V}_1, \vec{V}_1 \right\rangle + w_2 \left\langle \vec{V}_1, \vec{V}_2 \right\rangle\right)^2 + \left(w_1 \left\langle \vec{V}_1, \vec{V}_1 \right\rangle + w_2 \left\langle \vec{V}_1, \vec{V}_2 \right\rangle\right)^2 \\
&= w_1^2 \left\langle \vec{V}_1, \vec{V}_1 \right\rangle^2 + w_2^2 \left\langle \vec{V}_1, \vec{V}_2 \right\rangle^2 + 2 w_1 w_2 \left\langle \vec{V}_1, \vec{V}_1 \right\rangle \left\langle \vec{V}_1, \vec{V}_2 \right\rangle \\
&\quad + w_1^2 \left\langle \vec{V}_1, \vec{V}_1 \right\rangle^2 + w_2^2 \left\langle \vec{V}_1, \vec{V}_2 \right\rangle^2 + 2 w_1 w_2 \left\langle \vec{V}_1, \vec{V}_1 \right\rangle \left\langle \vec{V}_1, \vec{V}_2 \right\rangle
\end{aligned}
$$

$$
= \begin{pmatrix} w_1 & w_2 \end{pmatrix}
\begin{pmatrix}
\left\langle \vec{V}_1, \vec{V}_1 \right\rangle & \left\langle \vec{V}_1, \vec{V}_2 \right\rangle & \cdots & \left\langle \vec{V}_1, \vec{V}_n \right\rangle \\
\left\langle \vec{V}_2, \vec{V}_1 \right\rangle & \left\langle \vec{V}_2, \vec{V}_2 \right\rangle & \cdots & \left\langle \vec{V}_2, \vec{V}_n \right\rangle \\
\vdots & \vdots & \ddots & \vdots \\
\left\langle \vec{V}_n, \vec{V}_1 \right\rangle & \left\langle \vec{V}_n, \vec{V}_2 \right\rangle & \cdots & \left\langle \vec{V}_n, \vec{V}_n \right\rangle
\end{pmatrix}^2
\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}
$$

$[\theta]$ is a symmetric positive definite (when each signal V_i is distinct). This is the case in this project. Hence maximizing the above sum is the same as maximizing

$$S = \vec{w}^T [\theta] \vec{w}$$

where

$$[\theta] = \begin{pmatrix} \langle \vec{V}_1, \vec{V}_1 \rangle & \langle \vec{V}_1, \vec{V}_2 \rangle & \cdots & \langle \vec{V}_1, \vec{V}_n \rangle \\ \langle \vec{V}_2, \vec{V}_1 \rangle & \langle \vec{V}_2, \vec{V}_2 \rangle & \cdots & \langle \vec{V}_2, \vec{V}_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \vec{V}_n, \vec{V}_1 \rangle & \langle \vec{V}_n, \vec{V}_2 \rangle & \cdots & \langle \vec{V}_n, \vec{V}_n \rangle \end{pmatrix}$$

If we pick $\vec{w}$ to be an eigenvector of $[\theta]$ then the condition to maximize S will be found as follows (we can easily show this is true only if $\vec{w}$ is eigenvector) .

$$\begin{aligned} S &= \vec{w}^T [\theta] \vec{w} \\ &= \vec{w}^T \lambda \vec{w} \\ &= \lambda \|\vec{w}\|^2 \end{aligned}$$

Hence S is maximum when λ is the largest eigenvalue $\lambda_{\max}$ and $\vec{w}$ is the eigenvector associated with this $\lambda_{\max}$.

Hence the weights w_i are the coordinates of the largest eigenvector $\vec{w}$ of $[\theta]$. Now that we have found the weights, we have found $\vec{\Phi}$

$$\vec{\Phi} = \sum_{i=1}^{N} w_i \vec{V}_i$$

We normalize the eigensignal as well

$$\vec{\Phi} = \frac{\vec{\Phi}}{\left\| \vec{\Phi} \right\|}$$

How to use $\vec{\Phi}$

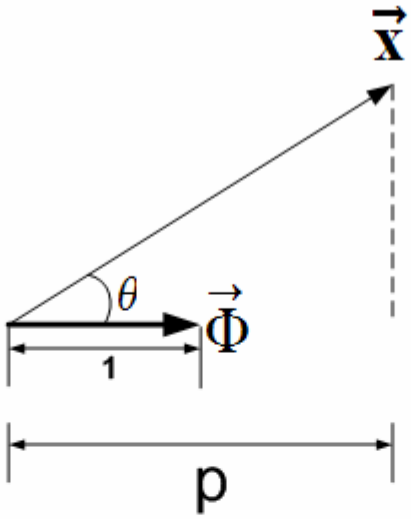
Given an arbitrary signal $\vec{x}$ that we wish to find how closely related it is to the the population $\{V_1, V_2, \cdots, V_N\}_c$, then we obtain $\vec{\Phi}_c$ from that population as described above, and then evaluate how closely correlated $\vec{x}$ is to $\vec{\Phi}_c$ by evaluating

$$p = \frac{\left\langle \vec{x}, \vec{\Phi}_c \right\rangle}{\left\langle \vec{\Phi}_c, \vec{\Phi}_c \right\rangle} = \frac{||\vec{x}|| \left\| \vec{\Phi}_c \right\| \cos\theta}{\left\| \vec{\Phi}_c \right\|^2} = \frac{||\vec{x}|| \cos\theta}{\left\| \vec{\Phi}_c \right\|}$$

But since $\vec{\Phi}_c$ is normalized, we can just write

$$p = \left\langle \vec{x}, \vec{\Phi}_c \right\rangle = ||\vec{x}|| \cos\theta$$

The more positive this p is, the more likely that $\vec{x}$ belongs to population $\vec{V}_i$. The more negative p is, the less likely it is from being part of the population.



$$p = \frac{\left\langle \vec{x}, \vec{\Phi}_c \right\rangle}{\left\langle \vec{\Phi}_c, \vec{\Phi}_c \right\rangle} = \frac{\|\vec{x}\| \left\| \vec{\Phi}_c \right\| \cos\theta}{\left\| \vec{\Phi}_c \right\|^2} = \frac{\|\vec{x}\| \cos\theta}{\left\| \vec{\Phi}_c \right\|}$$

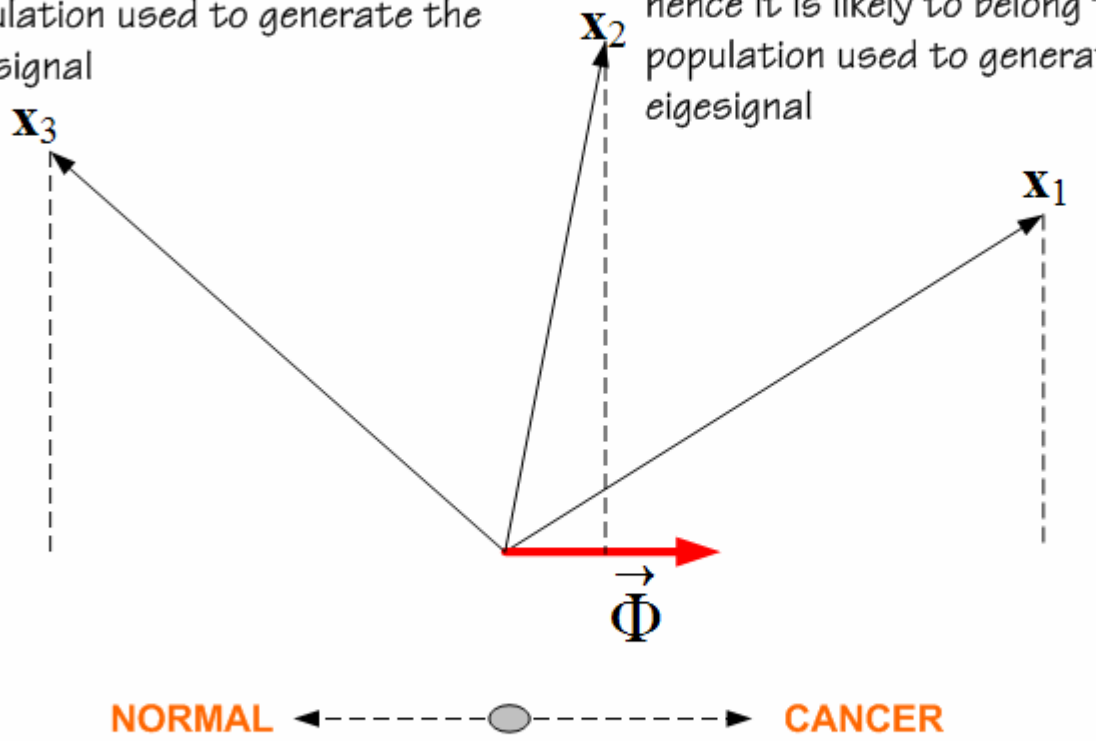
But since $\vec{\Phi}_c$ is normalized, we can just write

$$p = \left\langle \vec{x}, \vec{\Phi}_c \right\rangle = \|\vec{x}\| \cos\theta$$

This signal has large projection against the eigensignal, but it is a negative projection, hence it more less likely to belong to the population used to generate the eigesignal

This signal has smaller projection against eigensignal, but still positive, hence it is likely to belong to the population used to generate the eigesignal

While this signal has larger positive projection against eigensignal, hence more likely to belong to the population used to generate the eigesignal



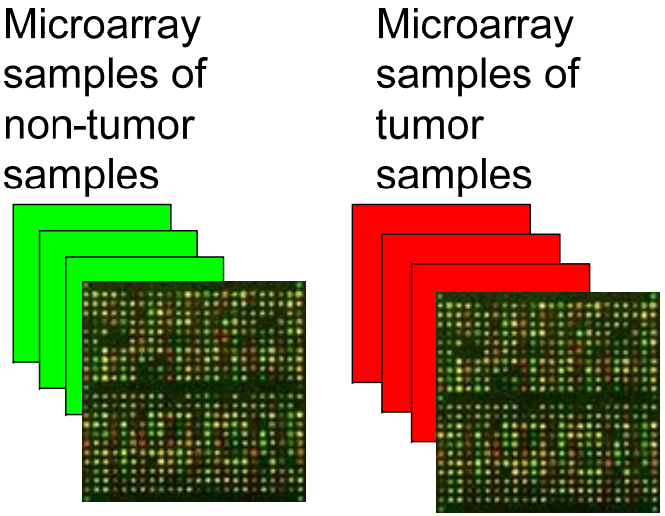
The larger the projection is on either side, the more certain one about its accuracy

Future analysis can incorporate this property in the form of a confidence measure related to the length of projection to make the PCA prediction results more certain

How to use $\vec{\Phi}$ for detection of cancer?

We obtained 2 sets of populations $\{\vec{C}_i\}$ which is the set of cancer signals, and $\{\vec{N}_j\}$ which is the set of normal signals. This was done for both bladder and liver.

From each one of the above 2 populations, we obtain the eigen signal for each: $\vec{\Phi}_C, \vec{\Phi}_N$

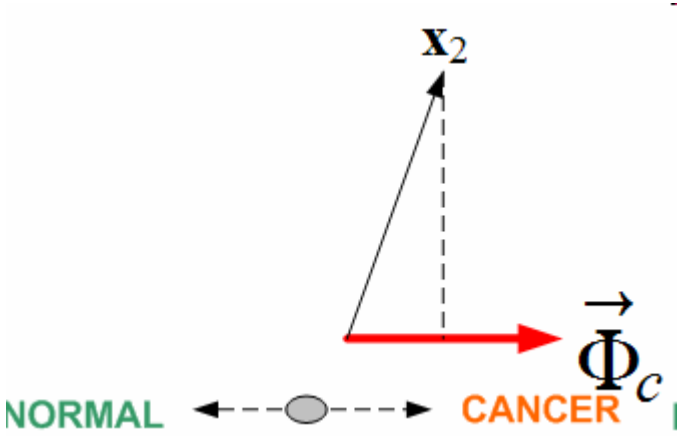


How to use $\vec{\Phi}_c, \vec{\Phi}_N$ for cancer detection

- 3 different algorithms used to determine if an arbitrary input sample is cancerous or not.
- Each algorithm was compared for accuracy.
- First algorithm uses $\vec{\Phi}_C$
- Second algorithm uses $\vec{\Phi}_N$
- Third algorithm is heuristic and uses a combination of $\vec{\Phi}_c, \vec{\Phi}_N$

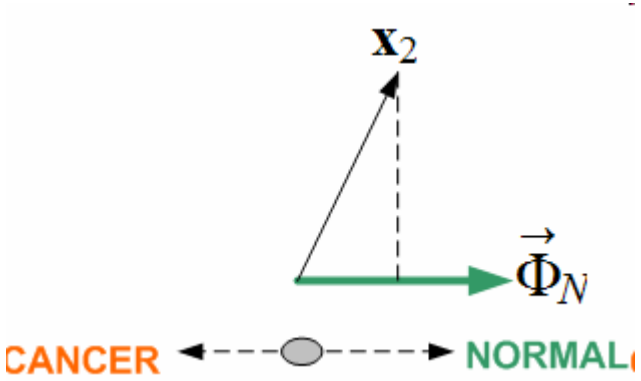
Algorithm 1

if $p_c(\vec{x}) > 0$ then $\vec{x}$ is cancerous else normal



Algorithm 2

if $p_n(\vec{x}) > 0$ then $\vec{x}$ is normal else cancerous



Algorithm 3

$IF\ p_c(\vec{x}) > 0\ OR\ p_c(\vec{x}) - p_n(\vec{x}) > 0$

Cancerous

ELSEIF $p_n(\vec{x}) > 0$ THEN

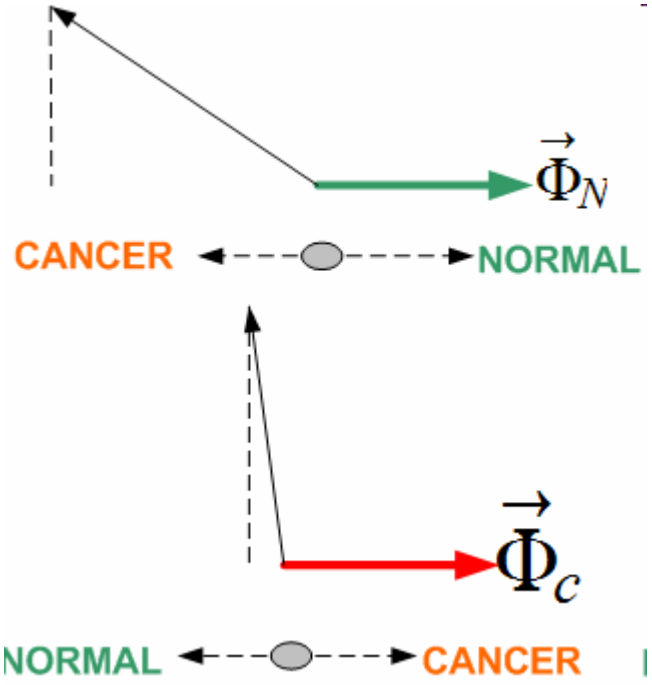
normal

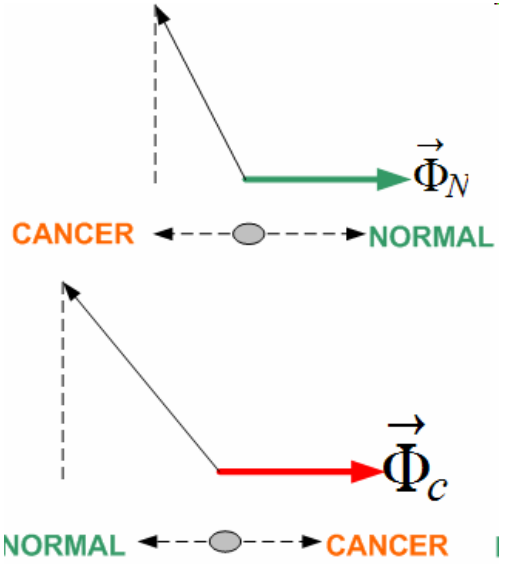
ELSE

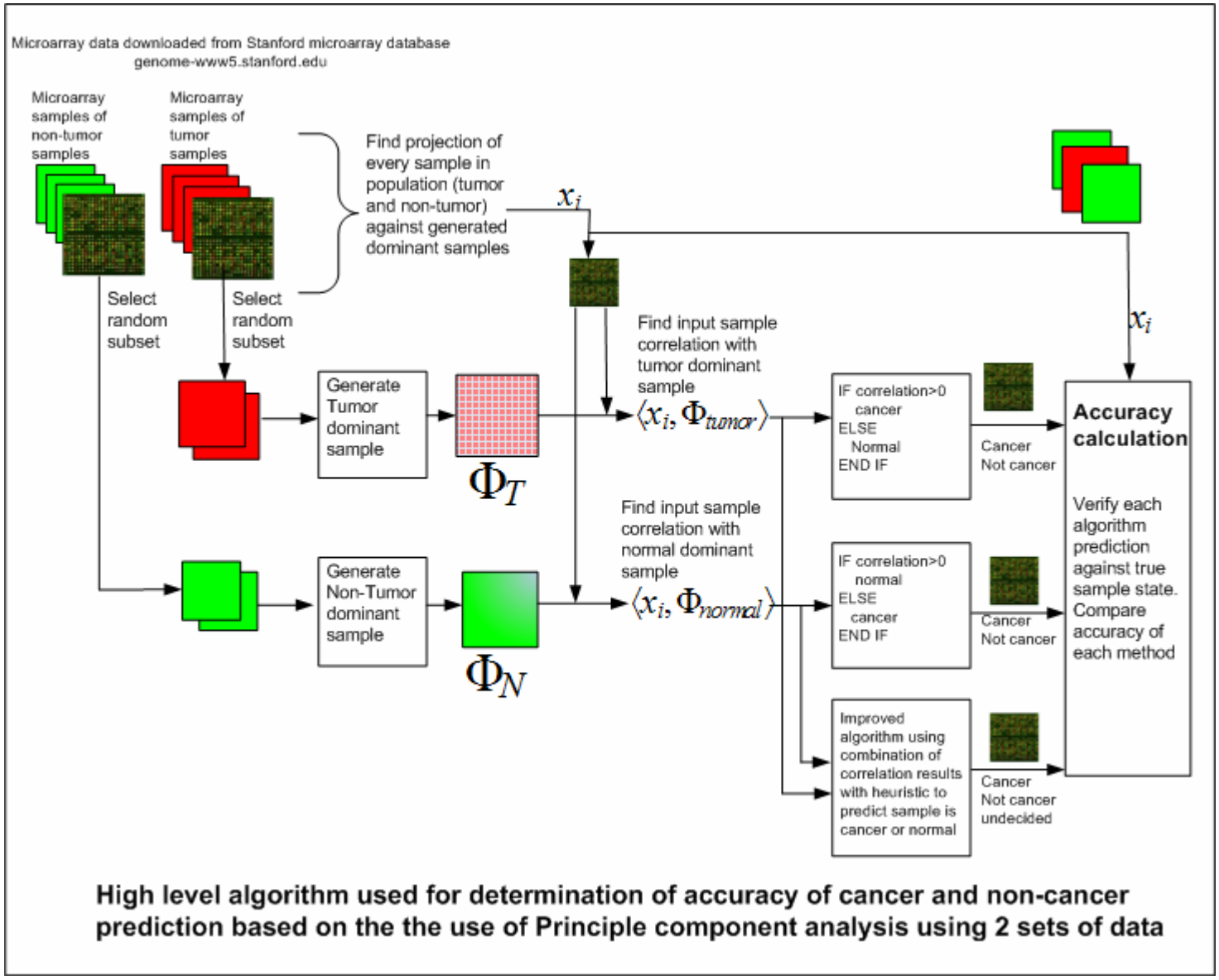
Unable to decide

ENDIF

Tumor







High level algorithm used for determination of accuracy of cancer and non-cancer prediction based on the the use of Principle component analysis using 2 sets of data

Using more than one eigensignal for finding correlation of arbitrary signal to population

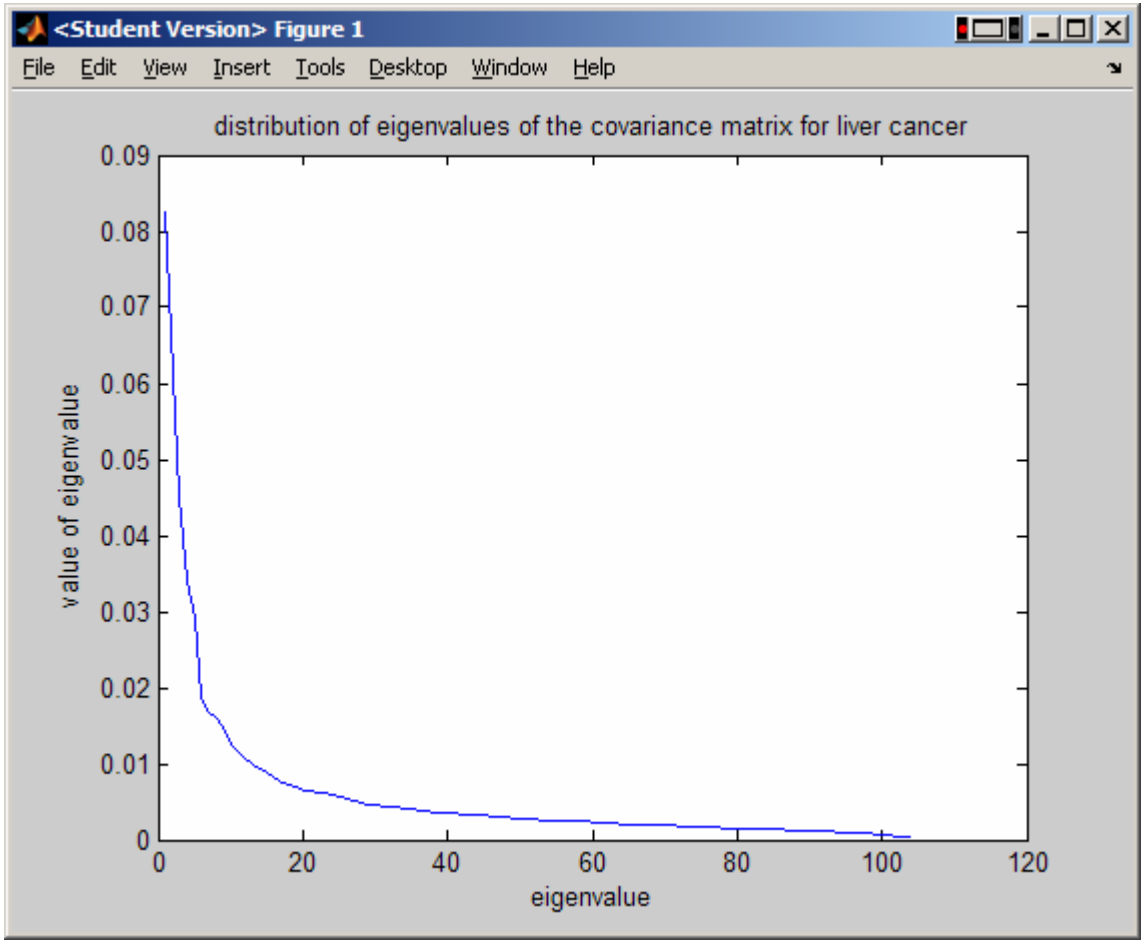
Only the first few eigensignals need be considered. Most of the information content is concentrated in the first (primary) eigensignals

K>> nSamples

104

[u,lam] = eig(theta);

plot(flipud(diag(lam)))



In this study, most of the power was found to be concetrated in the first eigenvectors (corresponding to the first few eigenvalues)

To find projection of an arbitrary signal againt more than one eigensignal:

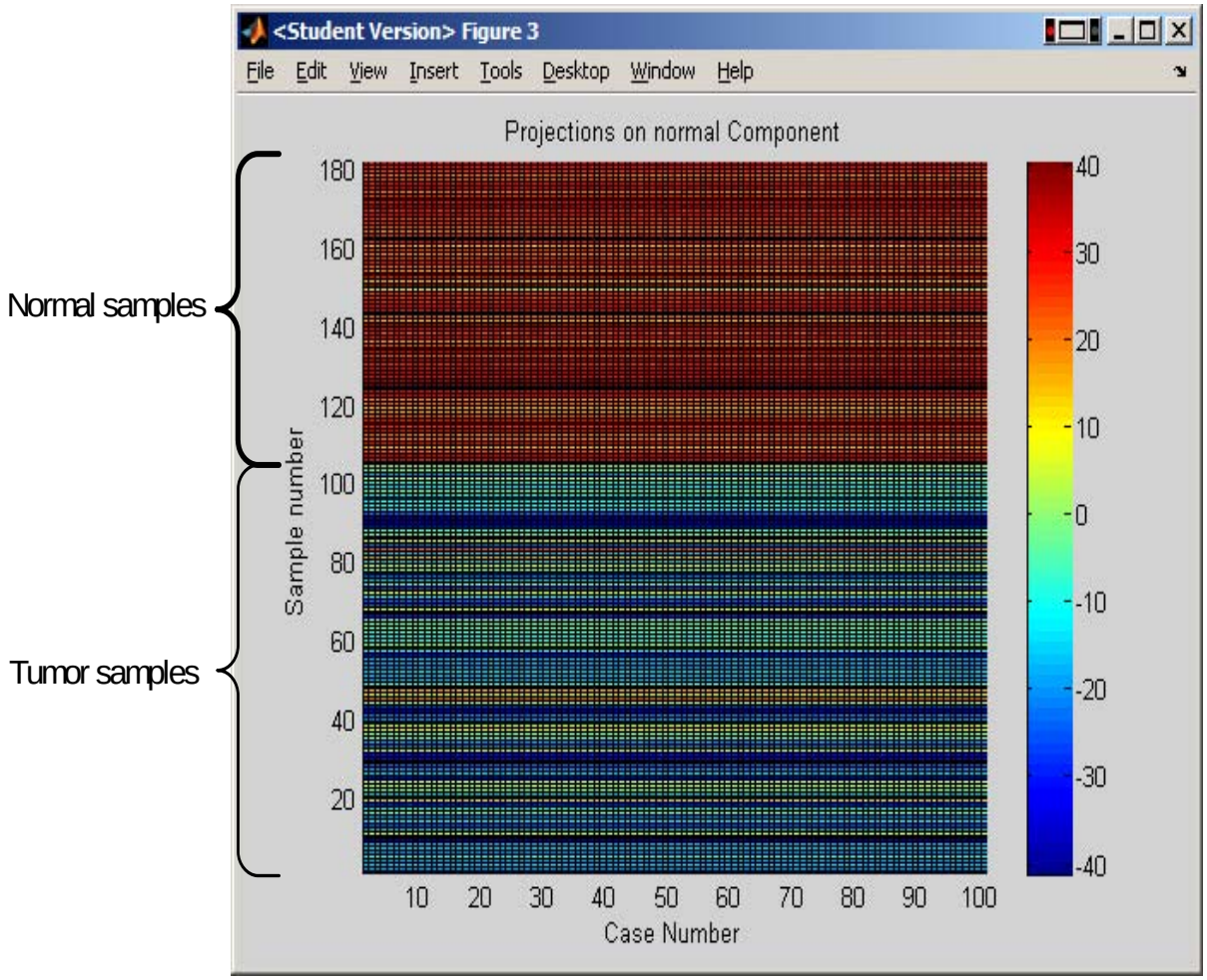
$$p(\vec{x}) = \frac{\langle \vec{x}, \vec{\Phi}_{c_1} \rangle}{\langle \vec{\Phi}_{c_1}, \vec{\Phi}_{c_1} \rangle} + \frac{\langle \vec{x}, \vec{\Phi}_{c_2} \rangle}{\langle \vec{\Phi}_{c_2}, \vec{\Phi}_{c_2} \rangle} + \cdots + \frac{\langle \vec{x}, \vec{\Phi}_{c_k} \rangle}{\langle \vec{\Phi}_{c}, \vec{\Phi}_{c_k} \rangle}$$

Accuracy for the diagnosis algorithms for liver and bladder cancers

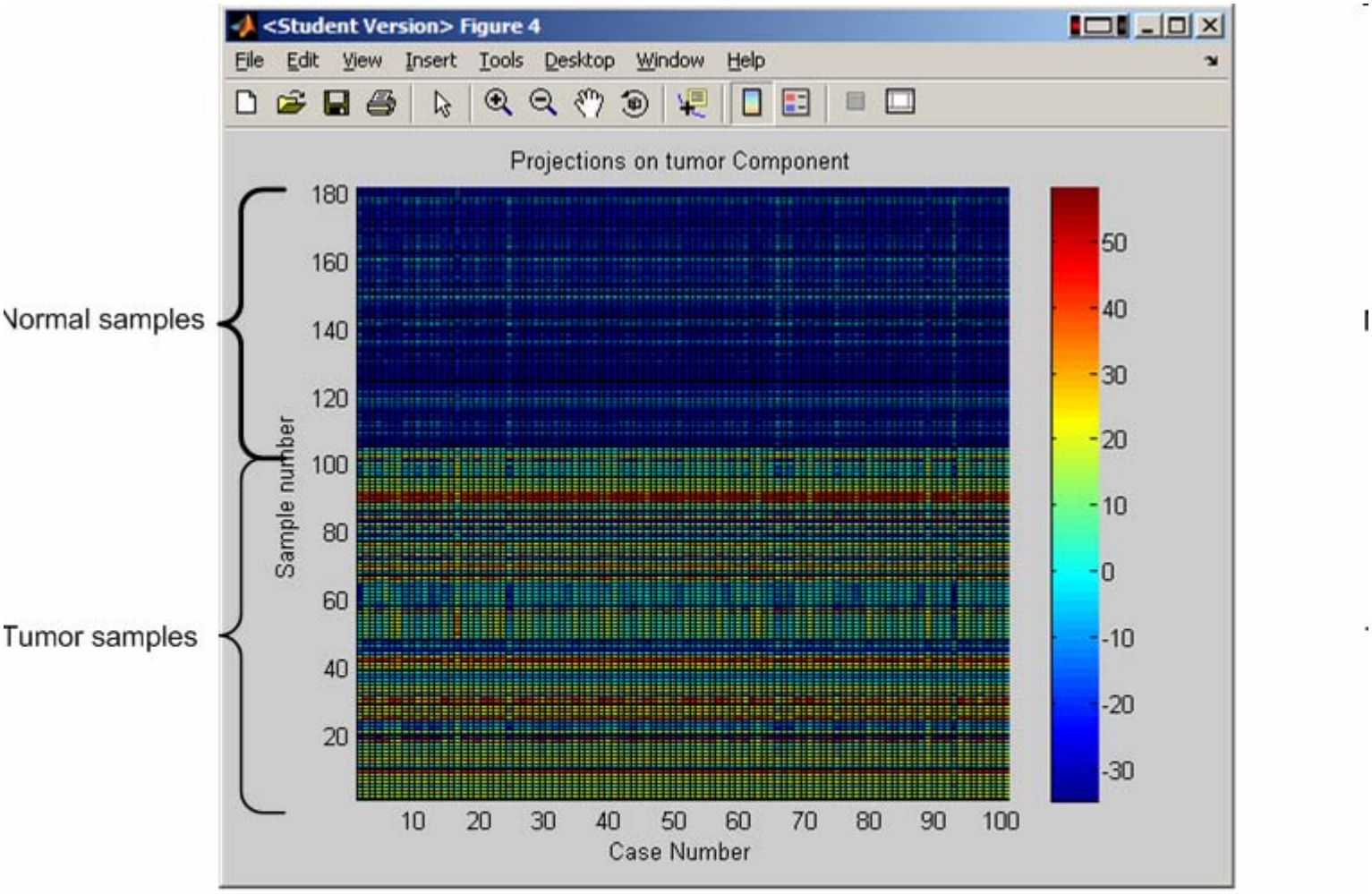
Data set	Accuracy of detection of	Algorithm	One mode	Two modes	Three modes	Four modes	Five modes
Liver	Cancer	1	69.46	82.74	80.58	80.37	78.07
		2	81.47	78.44	81.30	82.61	80.96
		3	80.75	88.54	87.15	89.82	89.54
	Normal	1	99.99	98.91	99.51	99.21	99.63
		2	100.00	96.41	95.11	93.28	90.68
		3	99.99	98.72	98.54	98.44	98.94
Bladder	Cancer	1	57.17	62.15	64.83	68.11	70.51
		2	80.35	77.35	73.20	69.23	70.26
		3	82.35	82.97	83.30	83.81	84.25
	Normal	1	99.95	99.32	99.86	99.95	100.00
		2	100.00	99.50	94.32	93.59	91.59
		3	99.82	99.41	99.71	99.81	100.00

Graphical view of the projection of samples (signals) onto the eigensample (eigensignal) repeated over 100 random experiments

105 known cancer samples and 76 known normal samples used.



Tumor samples correlate negatively with the tumor-free eigensample
While tumor-free samples correlate positively with the tumor-free eignesample

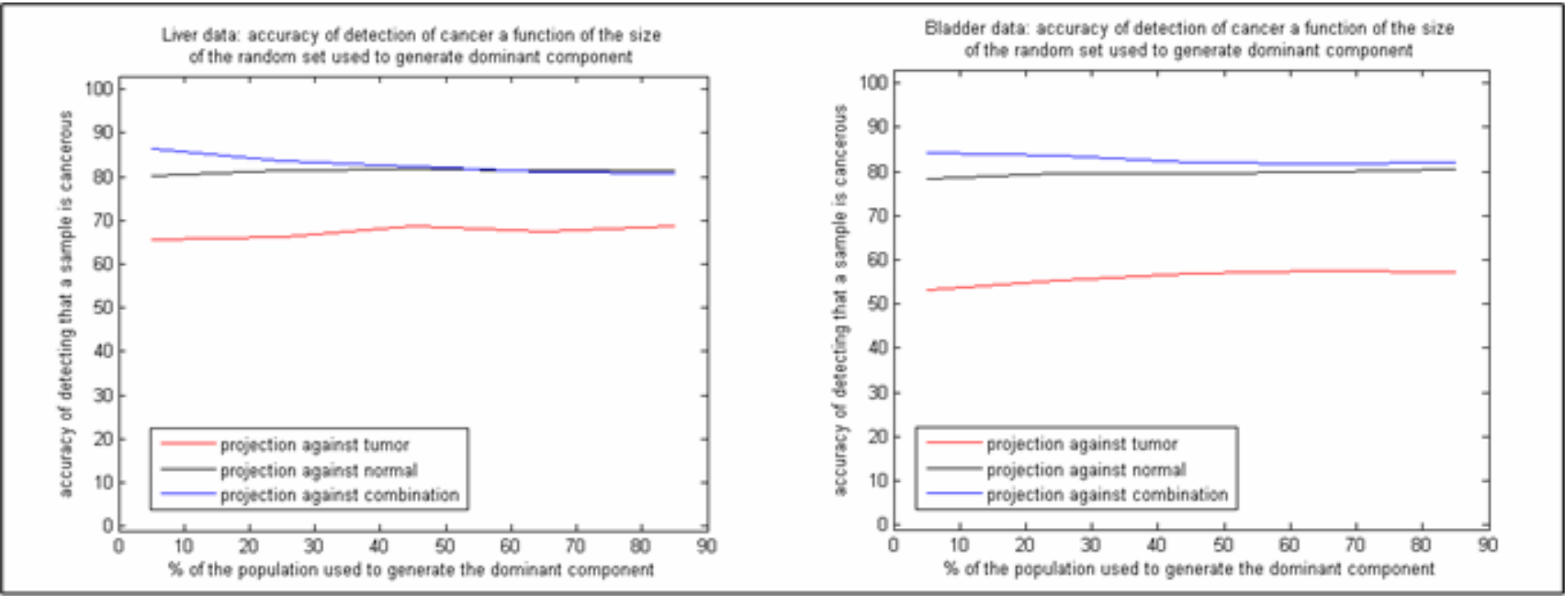


Tumor samples correlate positively with the tumor eigensample While tumor-free samples correlate negatively with the tumor eignesample

The effect of changing the number of signals used to generate the eigensignal on the accuracy of PCA for cancer detection

The population used to generate the eigensignal from is called the **working set**.

We now investigate the effect of changing the working set size on the accuracy of PCA for cancer detection


Liver data: accuracy of detection of cancer a function of the size of the random set used to generate dominant component
accuracy of detecting that a sample is cancerous
% of the population used to generate the dominant component
projection against tumor
projection against normal
projection against combination
Bladder data: accuracy of detection of cancer a function of the size of the random set used to generate dominant component
accuracy of detecting that a sample is cancerous
% of the population used to generate the dominant component
projection against tumor
projection against normal
projection against combination

General final observations

Based on the observations of the projections, we find that cancerous samples do not correlate positively as strongly with the cancerous dominant component when compared to how strongly the cancer-free samples negatively correlate with the cancerous dominant component.

Cancerous samples correlate much strongly, but in the negative sense, with the cancer-free dominant component.

Hence, when attempting to decide if a sample is cancerous or not, it is not recommend to measure the strength of the positive correlation with the cancerous dominant component, but instead one should measure the strength of how negatively the sample correlates with the cancer-free dominant component. The situation with cancer-free samples is different. Cancer-free samples do correlate very strongly in the positive sense with the cancer-free dominant principle component.

Cancer-free samples also correlate very strongly in the negative direction with the cancerous dominant component.

From the above, we conclude that it is best to always correlate the sample to be examined with the cancer-free dominant component since a cancer-free sample will exhibit a strong positive correlation while at the same time a cancerous sample would exhibit a strong correlation but in the negative sense. In other words, both types of samples have stronger correlations with the cancer-free dominant component when looking at the absolute magnitude of the correlation than the case would be if we had used a cancerous dominant component.

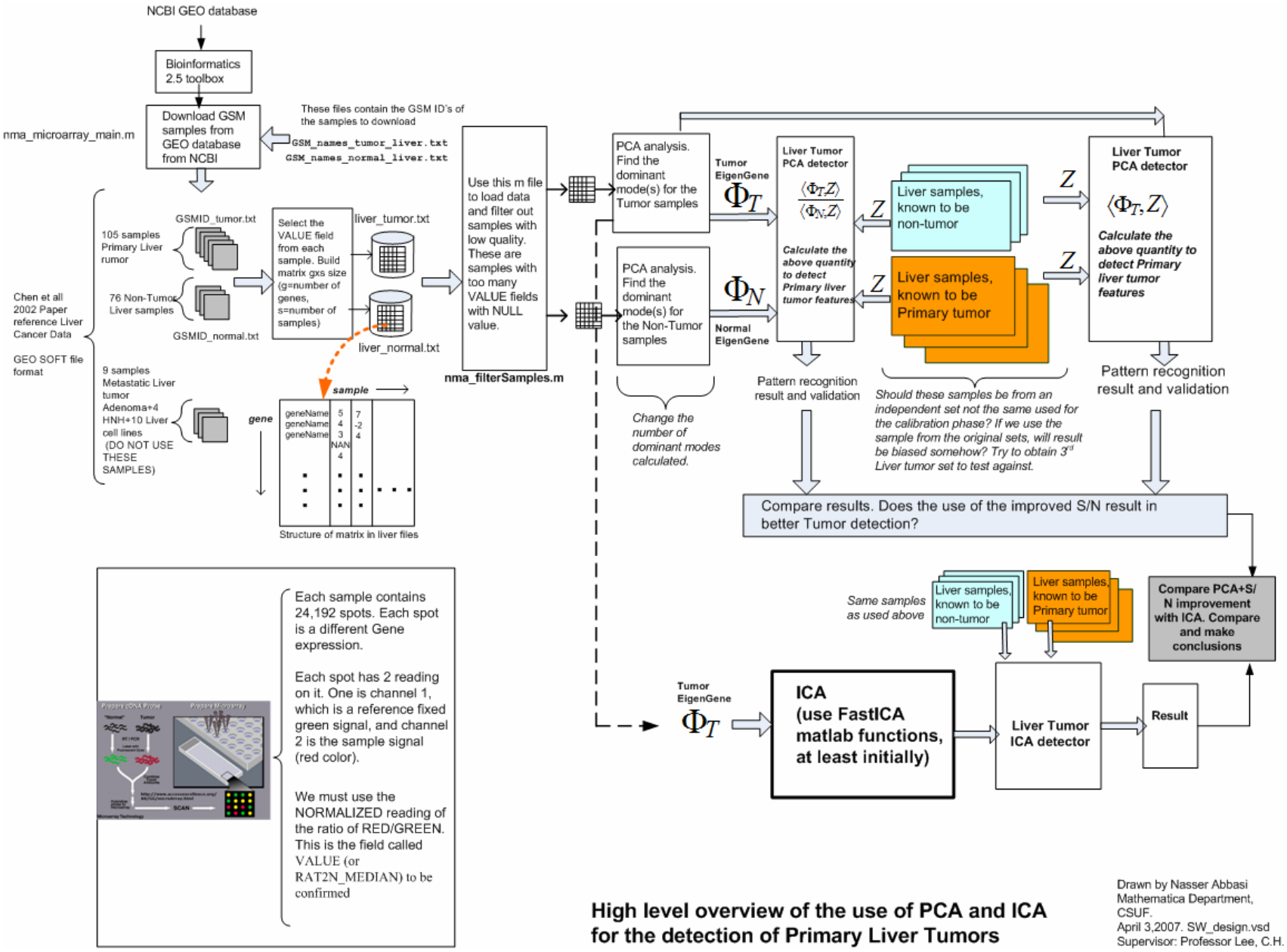
The third algorithm introduces a heuristic algorithmic improvement in the detection of cancer. As a result of this improvement, we were able to improve cancer detection. However, since this improvement in detection is based on a heuristic improvement, more tests are needed against larger set of data.

Study conclusion

- Examining the correlation of an arbitrary tissue sample with the PCA dominant component sample generated from the cancer-free samples produces more accurate results for both cancer and cancer-free detection
- An algorithmic improvement that considers the correlation of a sample against both PCA modes was implemented and was shown to produce more accurate diagnostic results.
- The effect of adding more eigensignals on the accuracy of PCA could not mathematically be analyzed at this time due to lack of time. Some tests showed that adding more eigensignals improved accuracy, while others showed it reduced accuracy. More analysis is needed on this to understand why this happens.
- PCA accuracy improved only slightly by increasing the working set size greatly. This shows that PCA can be effective in extracting dominant features that represent large population from small sample of the population.

Future possible research

- Use SVD for PCA and compare to see if there exist any accuracy improvement.
- Apply ICA (independent component analysis) and compare accuracy of ICA to PCA. See next slide for software flow diagram.
- Apply this analysis to larger set of microarray cancer data from NBCI and Stanford databases


NCBI GEO database
Bioinformatics 2.5 toolbox
nma_microarray_main.m
Download GSM samples from GEO database from NCBI
These files contain the GSM ID's of the samples to download
GSM_names_tumor_liver.txt
GSM_names_normal_liver.txt
Use this m file to load data and filter out samples with low quality. These are samples with too many VALUE fields with NULL value.
nma_filterSamples.m
Chen et all 2002 Paper reference Liver Cancer Data
GEO SOFT file format
105 samples Primary Liver rumor
GSMID_tumor.txt
76 Non-Tumor Liver samples
GSMID_normal.txt
Select the VALUE field from each sample. Build matrix gxs size (g=number of genes, s=number of samples)
liver_tumor.txt
liver_normal.txt
9 samples Metastatic Liver tumor Adenoma+4 HNH+10 Liver cell lines (DO NOT USE THESE SAMPLES)
sample
gene
geneName 5 7
geneName 4 -2
geneName 3 4
NAN 4
Structure of matrix in liver files
PCA analysis. Find the dominant mode(s) for the Tumor samples
Tumor EigenGene
Φ_T
PCA analysis. Find the dominant mode(s) for the Non-Tumor samples
Φ_N
Normal EigenGene
Change the number of dominant modes calculated.
Liver Tumor PCA detector
$\frac{\langle \Phi_T, Z \rangle}{\langle \Phi_N, Z \rangle}$
Calculate the above quantity to detect Primary liver tumor features
Liver samples, known to be non-tumor
Liver samples, known to be Primary tumor
Z
Should these samples be from an independent set not the same used for the calibration phase? If we use the sample from the original sets, will result be biased somehow? Try to obtain 3rd Liver tumor set to test against.
Liver Tumor PCA detector
$\langle \Phi_T, Z \rangle$
Calculate the above quantity to detect Primary liver tumor features
Pattern recognition result and validation
Pattern recognition result and validation
Compare results. Does the use of the improved S/N result in better Tumor detection?
Same samples as used above
Liver samples, known to be non-tumor
Liver samples, known to be Primary tumor
Compare PCA+S/N improvement with ICA. Compare and make conclusions
Tumor EigenGene
Φ_T
ICA (use FastICA matlab functions, at least initially)
Liver Tumor ICA detector
Result
Each sample contains 24,192 spots. Each spot is a different Gene expression.
Each spot has 2 reading on it. One is channel 1, which is a reference fixed green signal, and channel 2 is the sample signal (red color).
We must use the NORMALIZED reading of the ratio of RED/GREEN. This is the field called VALUE (or RAT2N_MEDIAN) to be confirmed
High level overview of the use of PCA and ICA for the detection of Primary Liver Tumors
Drawn by Nasser Abbasi
Mathematica Department,
CSUF.
April 3,2007. SW_design.vsd
Supervisor: Professor Lee, C.H.

Thanks and references

- Thanks to Dr C.H. Lee for his advice during this project..

REFERENCES

- Chen, X., et. al., "*Variation in Gene Expression Patterns in Human Liver Cancers*", Mol Biol Cell. 2002 Jun; 13(6): 1929-39.
- Chen, X., et. al., "*Variation in Gene Expression Patterns in Human Gastric Cancers*", Mol Biol Cell. 2003 Aug; 14(8): 3208-15. Epub 2003 Apr 17.
- N. Abbasi and C.H.Lee "FEATURE EXTRACTION TECHNIQUES ON DNA MICROARRAY DATA FOR CANCER DETECTION". Conference paper. WACBE world congress on bioengineering 2007. Bangkok, Thailand.
- H.V. Ly and H.T. Tran, "*Modeling and Control of Physical Processes using Proper Orthogonal Decomposition*," Computers and Mathematics with Applications, vol. 33 (2001) pp. 223-236.
- D. Peterson and C. H. Lee, "Disease Detection Technique Using the Principal Orthogonal Decomposition on DNA Microarray Data" Proceedings of the 6th Nordic Signal Processing Symposium, NORSIG 2004, Espoo, Finland, pp. 33-36, (2004).
- C. H. Lee and D. Peterson, "A DNA-based pattern recognition technique for cancer detection" Proceedings of the 26th Annual Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 2 (2004) pp. 2956-2959.
- C.H. Lee and M. Vodhanel, "Cancer detection using component analysis methods on DNA microarrays" Proceedings of the 12th Int. Conf. on Biomedical Engineering (2005), Singapore.