FINAL PROJECT REPORT

Filtering, Unsupervised Clustering and Statistical Analysis (Two Sample T-test) of Cancer Gene Expression Data

By
SHUBHRA GUPTA

CBS 591
DATA MINING
Computational Biosciences
May 6, 2003

Abstract

In this final project report we will see what we have done in this project. We have fixed a cancer data set for our project and we wrote a working program in C for filtering the cancer data set. After filtering the data, we have done unsupervised clustering (K-mean) using software (EPCLUST). Then we compare our results with the work of Laura J. van 't et al. And also we have done two samples T-test of cancer data to know which gene is down regulated in two classes (relapse and non-relapse) of samples. And how many downregulated genes belong to cluster1 and cluster2 and verify our results from NCBI (National Center for Biotechnology Information) using entrez.

1. Introduction

1.1 DNA Microarray

In gene microarray data each chip represents an experiment. An array in microarray data is an orderly arrangement of samples. In general, arrays are described as microarrays. Microarray technology provides an opportunity to analyze the expression level of thousands of genes in a tissue or cell culture sample. Due to these massive amounts of data, the occurrence of meaningless noises and variations are inevitable. Data mining techniques are useful in this regard.

1.2 Application of Microarray

There are two major applications for the DNA microarray technology

- Identification of sequence (gene / gene mutation), and
- Determination of expression level of genes

And others are drug discovery, gene discovery, disease diagnosis and etc.

In microarray data there exist at least two nomenclature systems for referring to hybridization partners. Both use common terms: ***probes*** and ***targets***. **Probe** is the tethered nucleic acid with known sequence, whereas a **target** is the free nucleic acid sample whose identity is being detected.

1.3 Data Preprocessing

Data preprocessing has to perform, to increase the accuracy of the mining task. There are many methods to preprocess data. Like,

Data filtering
Data Cleaning
Data Integration
Data Normalization
Data Reduction

1.4 Needs to Normalize Microarray Data

Differences in labeling efficiency between the two dyes (like cy3 and cy5) will affect different microarray experiments to different extents. Thus to compare microarrays, we need to try to remove the systematic variation (that is differences in labeling efficiency between the two dyes) to bring the data from the different experiments onto a field, where one can analyze it.

1.5 Use of Tool

The purpose of using tool is to find patterns in large data sets. Clustering experiments discover new subtypes of tissue samples. The continued success of the methodology depends on the development of computational tools that can mine the resulting large data sets.

1.6 Some Available Algorithms for Unsupervised Clustering:

Hierarchical methods - Agglomerative and Divisive hierarchical clustering, Partitioning methods – k-means, k medoids, Self-organization map, CLICK, CAST. Some comparison of clustering algorithms is given in Sharan, Elkon and Shamir (2001).

1.7 Downregulated and Upregulated Gene

In microarray chip green fluorescent dye represents reference cDNA → downregulated gene and red fluorescent dye represents reference cDNA → upregulated gene.

To do this project I took the breast cancer microarray data that data is normalized data.

1.8 Hereditary Breast Cancer

The majority of breast cancers are not due to inherited factors. Doctors think that inheriting a faulty gene causes only small number of breast cancers (about 5-10%). In the last several years' two main genes have been identified, known as BRCA1 and BRCA2 (breast cancer 1 and breast cancer 2). If there is a mutation (fault) in BRCA1 (chromosome 17) or BRCA2 (chromosome 13) it can increase a person's chances of getting breast cancer, but does not mean it will definitely occur. There have some other breast cancer genes, which are known, but some other have not yet been found. We each have two copies of approx 30,000 genes. One copy of each gene inherits from our father and one copy from our mother. Each gene contains a different package of information to tell our bodies how to grow and work. Suppose this information is in the form of a series of letters. If there is a mistake in the series of letters, this results in a mutation in the gene. There are several from of breast cancer like, Ductal carcinoma in situ (DCIS) is an early form, secondary form of breast cancer, etc. The main concern to diagnosis is if one would find a lump or other change in their breasts. There are many treatments for breast cancer like Chemotherapy, radiotherapy, clinical trails, etc…
Each year 182,000 women are diagnosed with breast cancer and 43,300 die. The vast majority of breast cancers occur in women over the age of 50.

A report from the **National Cancer Institute (NCI),** one woman in eight in the United States either has or will develop breast cancer in her lifetime. In addition, 1,600 men will be diagnosed with breast cancer and 400 will die this year.

In this project report section 2 include objective of project. Section 3 gives knowledge about data set. Section 4 is about algorithm of filtering and contain C program. Section 5 gives about algorithm of k-mean and what software I used to do clustering. Section 6 includes results about clustering. Section 7 gives which gene belongs to cluster 1 and which gene belongs to cluster 2 after filtering. Section 8 represents graphical results of clustering. Section 9 is why Laura J. van 't et al used hierarchical clustering. T-test and results of T-test is given in section 10. Types of cluster algorithm one can use are given in section 11. Section 12 includes references.

2. Objective

The major goal of these studies is to identify a subset of informative genes for class prediction as well as to uncover classes that were previously unknown (class discovery).

3. Dataset

I have breast cancer data from the following website
http://www.rii.com/publications/2002/vantveer.htm

Dataset given in above link is looks like below data, first column contain gene name, second column contain sample1 with two values log10 (ratio) and P-value. This is similar for 77 samples.

	Sample 1,>5 yr survival(150 months) age 43 erp yes		**Sample 2,>5 yr survival(77 months) age 44 erp yes**		**Sample 3,>5 yr survival(128 months) age 41 erp no**		**Sample 4,>5 yr survival(156 months) age 41 erp yes**	
Systematic name	**Log10(ratio)**	**P-value**	**Log10(ratio)**	**P-value**	**Log10(ratio)**	**P-value**	**Log10(ratio)**	**P-value**
Contig45645_RC	-0.048	9.32E-01	-0.243	5.00E-01	0.081	9.08E-01	-0.215	4.95E-01
Contig44916_RC	-0.005	9.91E-01	-0.09	7.93E-01	-0.035	9.64E-01	-0.207	6.42E-01
D25272	0.102	8.10E-01	-0.152	7.35E-01	0.409	8.71E-01	-0.158	5.95E-01
J00129	-0.448	3.59E-01	-0.48	1.79E-01	-0.568	6.40E-02	-0.819	2.66E-01
Contig29982_RC	-0.296	3.67E-01	-0.512	1.38E-01	-0.411	3.37E-01	-0.267	7.99E-01
Contig26811	0.055	9.04E-01	-0.189	6.89E-01	1.339	6.35E-01	0.229	7.77E-01
D25274	-0.147	5.17E-02	0.108	7.81E-02	0.011	9.28E-01	0.059	5.35E-01
Contig36292	0.085	8.98E-01	-0.105	7.91E-01	0.792	6.07E-01	-0.018	9.88E-01
Contig42854	-0.1	1.64E-01	-0.031	6.27E-01	-0.398	3.91E-02	0.023	8.45E-01
Contig34839	0.081	8.77E-01	-0.105	7.85E-01	0.336	7.54E-01	-0.064	9.69E-01

This is the work of Laura J. van 't veer et al (1998) on breast cancer data. In the paper of Laura J. van 't veer et al (1998) breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome. The strongest predictors for metastases, e.g., lymph node status and histological grade of each sample, fail to accurately classify breast tumors according to their clinical behavior. Chemo- or hormonal therapy reduces the risk of distant metastases by approximately 1/3rd, although 70-80% of the patients would have survived without this treatment.

Here in our data we used DNA microarray analysis on primary breast tumors of 77 young patients and applied unsupervised clustering to identify a gene expression profile. This gene expression profile could be helpful in clinical parameters to predict the outcome of disease.

4. Work done

4.1 Data filtering

Filtering is the process of deciding which genes in a microarray experiment have significantly varying expression across conditions. We want to know which gene has different expression in relapse and non-relapse class. E.g. filter out those genes whose similar expression in relapse and non-relapse class.

I did filtering of breast cancer data. In the filtering process I have taken 77 primary breast cancer samples in which 33 patients who developed distant metastases within 5 years that is **relapse** and 44 patients who continued to be disease-free after a period of at least 5 years that is **non-relapse** (all attributes) with 24,483 human genes (instances). Some 3037 genes significantly found from the group of 77 samples using twofold difference >= 0.3 (that is log10 intensity ratio) and P-value < 0.01 in at least four tumors (samples) in my dataset.

4.2

To do filtering of cancer data, I wrote a program in C, which follows the following steps:

Filtering (data file, genes, samples)
Input: Raw data of 77 samples (tumors) with twofold difference & P-values and one column of Systematic name (Systematic name given to each gene or sequence) total 155(attributes) and 24,483 genes (instances)
Output: Get those genes that satisfied the condition of twofold and P-value at least in four samples

Algorithm:
Step 1*:* Initialization and data reading.
Step 2: Compare the cell value of each sample for twofold difference >= 0.3 and P-value < 0.01 in at least 4 tumors.
Step 3: Select the rows satisfying the condition in step 2.
Step 4: Print the data in a separate file.

This is the actual program in C for filtering process.

```
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <ctype.h>
#include <float.h>
#define M 25000
#define N 100
/* This is breast cancer normalization program */
main( )
{
      FILE *in,*out;
      int i,j;
      int rows,cols;
      int count=0;
char*  a[M][N] ;
char file[25],file1[25];
printf("Enter the input data file name:");
scanf("%s",file);
printf("Enter the output data file name:");
scanf("%s",file1);
printf("Enter the no of rows:");
scanf("%d",&rows);
printf("Enter the no of cols:");
scanf("%d",&cols);

/*_______input- output_____________________________*/
in=fopen(file,"r");
out=fopen(file1,"w");
for(i=0;i<rows;i++)
{
   for(j=0;j<cols;j++)
        {
      a[i][j] = (char *) malloc(100* sizeof(char));
      fscanf(in,"%s",a[i][j]);
      }
}
/*
//for debugging purposes to make sure it reads proper
   for(i=0;i<rows;i++)
{
   for(j=0;j<cols;j++)
        {
      printf("%s",a[i][j]);
      }
}
*/

fprintf(out,"The normalized data is\n");
for(i=2;i<rows;i++){
      count = 0;
for(j=0;j<(cols-1)/2;j++){
    if(atof(a[i][2*j+1])>=0.3 &&  atof(a[i][2*j+2])<0.01)
    {
```

```
            count=count++;
    }

                    }
if(count >=4){
for(j=0;j<cols;j++){
      fprintf(out,"%s ",a[i][j]);}
fprintf(out,"\n");
                    }
                    }

fclose(in);
fclose(out);
}
```

In the filtering process I also used MS excel, MS access.

These are some genes that I got after filtration.

Gene name
J00129
Contig42014_RC
Contig27915_RC
Contig20156_RC
Contig50634_RC
Contig56678_RC
Contig48659_RC
Contig49388_RC
Contig1970_RC
Contig53047_RC
Contig19551
Contig10437_RC
Contig47230_RC
Contig20749_RC
AL157502
Contig36647_RC
AB033006
AB033007
AB033025
Contig40673_RC
Contig17345_RC
AB033035
AF227899
AB033049
Contig67229_RC
Contig3396_RC
AB033066
Contig46243_RC
Contig26077_RC
U45975

Contig43679_RC
AB033073
AB033086
NM_003004
NM_003010
NM_003012
NM_003014
Contig29226_RC
NM_003022

4.3 log ratios

Suppose A/B=10, A/B=1 and A/B= -10 then log of these values are log A/B=1, log A/B=0 and log A/B = -1. Using log one can decrease the difference between values.

5. Data Mining-Clustering

Clustering is often used for discovering classes, patterns or structure in data. Cluster analysis for grouping together genes or samples with similar expression patterns. Objects in the same cluster are similar to each other than objects in different clusters. Applications of clustering algorithms to microarray data are for the purpose of functional genomic, cell classification and disease diagnostics.

Pattern recognition methods can be divided into two categories: **supervised and unsupervised**.

5.1 In unsupervised clustering I used k-mean clustering algorithm.

Unsupervised methods – K- Means
Input: The number of clusters k and a database with n objects.
Output: A set of cluster that minimizes the squared-error criterion.
Complexity O (nkt), n= number of objects, k= number of cluster and t= number of iteration

Method:
Step1: Randomly select k object from the data points, each of which initially represents a cluster mean or center
Step2: for remaining objects, an object is assigned to the cluster to which it is most similar, based on the (Euclidean or correlation) distance between the object and the cluster mean
Step3: Then compute the new mean for each cluster
Step4: repeat this process until no change in cluster mean
This process is used to minimize the squared-error criterion of a set k clusters.

5.2 Choose k-mean

This is the clustering algorithm, which I understood very well in the class and also class labels are not present in the training data.

5.3 Software

For k-mean clustering I used software EPCLUST (**E**xpression **P**rofile data **CLUST**ering and analysis). This software can do data clustering, visualization, and analysis for numeric (e.g. gene expression data) as well as sequence data. This is a web-based tool. Using EPCLUST one can do hierarchical and K means clustering. Jaak Vilo, Patrick Kemmeren and Misha Kapushesky wrote this tool.

5.4 Working of EPCLUST

This tool can do filtering of your data set. But I have already done filtering with c program. So have done only clustering from that tool.

Algorithm of EPCLUST:
Input: Data file, value of k, Distance (Euclidean or Pearson correlation)
Output: cluster data according to similarity
Step1: upload data file
Step2: Select clustering process (k-mean)
Step3: Define value of k
Step4: Select distance (correlation)
Step5: Get cluster according to value of k

6. Evaluation of work

Results:
1.
I found two groups of cluster based on relapse and non-relapse class. In these groups distance between genes within cluster by correlation and distance from center to genes was Euclidean.
In my experiment I got **1512** genes in cluster1 and **1525** genes in cluster 2.

Laura et al did hierarchical clustering on 98 samples and 5,000 genes
a) Found a group of downregulated genes which contains
- ER-α gene and genes co-regulated with ER.
b) Another gene cluster was associated with lymphocytic infiltrate, which contains those
- Genes expressed primarily by B and T cells.

2.
I compared my result with the work of Laura J. Van't Veer Et al.
I don't know about all the genes but whatever genes they gave in their paper.
I compare those and found all **genes of ER-∝ gene belonging to cluster 1** and all **genes, which primarily expressed by B and T cell belonging to cluster 2**.

So in that way my results are consistence.

3.

Conclusion: clustering detect two subgroup of breast cancer; differ in ER status and lymphocytic infiltration.

7. These are some genes from cluster1 and cluster2.

CLUSTER_0001
J00129
Contig42014_RC
Contig27915_RC
Contig20156_RC
Contig50634_RC
Contig53047_RC
Contig19551
Contig10437_RC
AB033006
Contig17345_RC
AB033035
AB033066
NM_003004
NM_003012
Contig29226_RC
NM_002300
NM_003034
NM_003035
NM_002308
Contig54839_RC

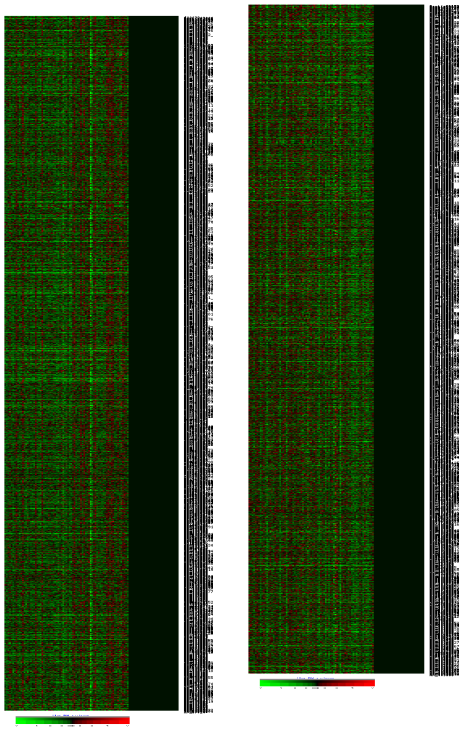
CLUSTER_0002
Contig56678_RC
Contig48659_RC
Contig49388_RC
Contig1970_RC
Contig47230_RC
Contig20749_RC
AL157502
Contig36647_RC
AB033007
AB033025
Contig40673_RC
AF227899
AB033049
Contig67229_RC
Contig3396_RC
Contig46243_RC
Contig26077_RC
U45975

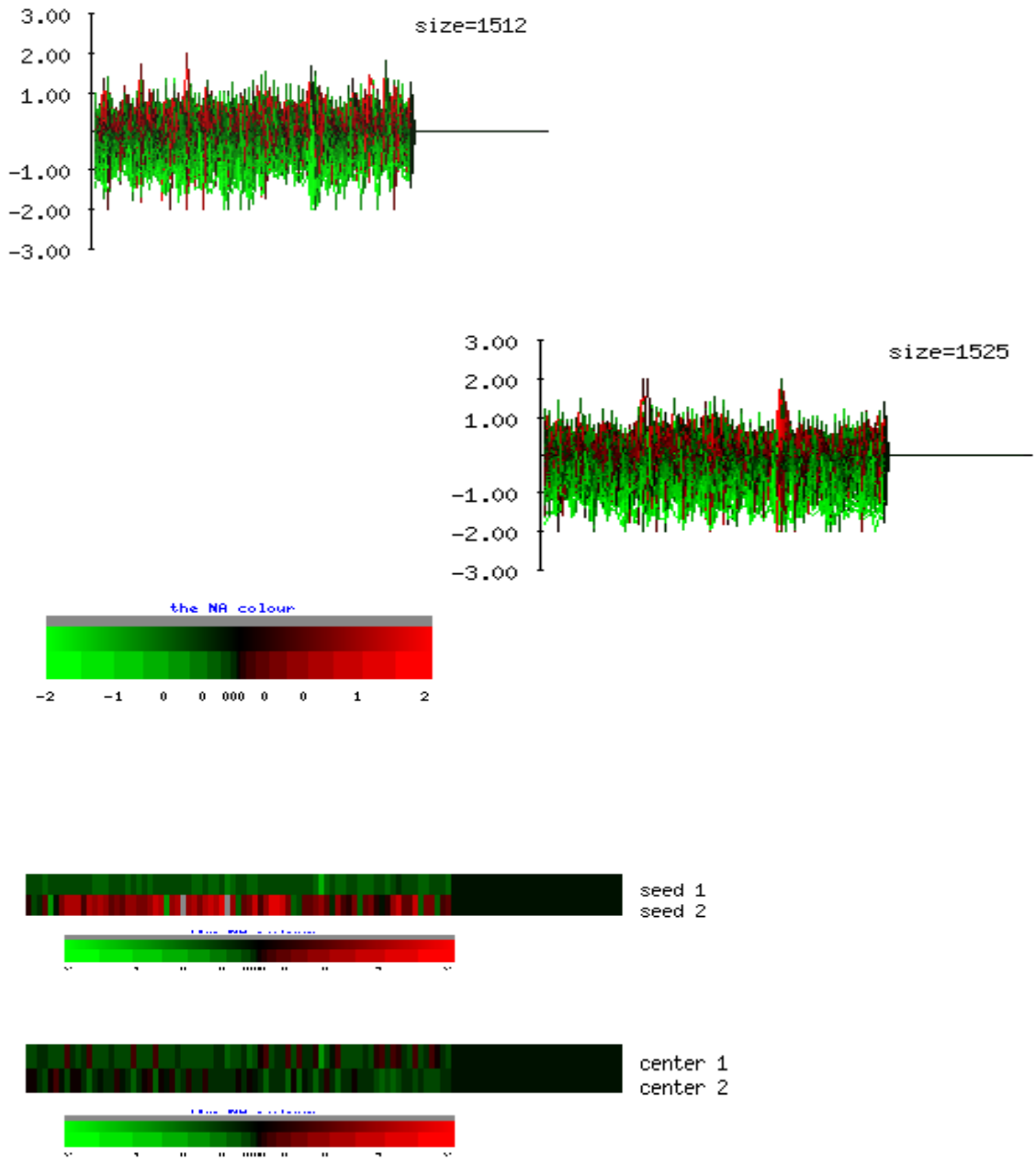
ELEMENTS

Obj	Cluster	dist from center	Silhouette (if exists)
O:1	2	4.088646	
O:2	1	1.826677	
O:3	1	2.069855	
O:4	1	2.290362	
O:5	1	1.887579	
O:6	1	4.160309	
O:7	1	1.418852	
O:8	1	2.050042	
O:9	1	2.141933	
O:10	2	6.527531	
O:11	1	4.061175	
O:12	1	2.502386	
O:13	2	4.031596	
O:14	1	3.088104	
O:15	1	2.537169	
O:16	2	5.107786	
O:17	1	2.639123	
O:18	1	1.811228	
O:19	1	2.874770	
O:20	1	2.789006	

8.

These are some clustering graph. First graph is representing cluster 1 and second one is representing cluster2.





9. They used hierarchical clustering

Hierarchical clustering algorithms are among the oldest and most popular clustering method. They fall into the class of unsupervised clustering and visualization methods. These techniques are useful when labels are unavailable. Examples include attempts to identify (yet unknown) sub-classes of tumors or work on identifying the clusters of genes. Probably it is because of this reason the authors (Laura J. Van't Veer Et al, 2002) used this approach. Hierarchical clustering approach is to differentiate between some kinds of tumor types for the purpose of choosing the most appropriate chemotherapy treatment.

.

10. Statistical Analysis two sample T-test

I did T-test on each filtered 3037 genes to see whether genes are differentially expressed between relapse and non-relapse samples or not. By T-test one can know which gene is down regulated in non-relapse and which gene is down regulates in relapse. And also it allow us to distinguish, to some extent, between differences of intensity measurement due to gene expression differences between relapse and non-relapse tissues and those due to other factor such as a slightly defective array chip. This would help in disease treatment by looking difference in gene expression.

10.1

I used Two sample t-test formula. In this formula n is the length of one class sample and

$$T = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{nS_X{}^2 + mS_Y{}^2}{n+m-2}}\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{(\bar{X} - \bar{Y})/\hat{\sigma}_{joint}}{\sqrt{\frac{1}{n}+\frac{1}{m}}}.$$

m is the length of other class sample.
T-statistic follows a t-distribution with n+m-2 degrees of freedom.

In my case n = 44 samples of non-relapse and m =33 samples of relapse.
Degree of freedom n+m-2 = 44+33-2 = 75
So I calculated all t-distribution with 75 degree of freedom.

I have done all the calculations in MS excel. Calculate “division factor”, Mean of each gene in groups of non-relapse and relapse samples, difference between mean of each gene in groups of samples, then find out the sums of squares for each gene in groups. Calculate the T values according to above formula. Then compare these t values with 75 degree of freedom at 5% (95%) significance, which is1.664 (from t-distribution table). Collected all genes, which was greater> 1.664 or <-1.664 (because t-distribution is symmetrical).

10.2 Results

1.
In the t-test I got 717 genes, which were satisfying this condition.

2.
Then ranked them according to mean and find out
380 genes is downregulated in relapse

Given below are few of them

Contig53047_RC
Contig19551
AB033006
AB033066
NM_002300
NM_003035

NM_002358
NM_003090
NM_002395
NM_001673
NM_001679
D25328
Contig50122_RC
NM_003108

337 genes are downregulated in non-relapse

Some of genes, which is downregulated in non-relapse

NM_001609
Contig50838_RC
NM_001635
NM_000909
NM_000949
NM_000964
NM_000992
Contig47045_RC
Contig27623_RC
NM_003105
NM_003118
M59979

3.

To see what is the percentage of those downregulated genes in cluster 1 and cluster 2. I picked up **22 downregulated genes from relapse sample and found 18 belong to cluster 1 and 4 belong to cluster 2. So out of 22 genes 4 represent ER-∝ gene class and 18 represents primarily expressed by B and T cell. Same processor I did for non-relapse and found that out of 15 genes, 11 belong to cluster 1 and 4 belong to cluster 2. So out of 15 genes 4 represent ER-∝ gene class and 11 represents primarily expressed by B and T cell.**

4.

We can't say for all 717 downregulated genes because one need to check for each and every gene, but **based on these results, I can say that the major part of downregulated genes belong to cluster1.**

5.

To verify these results I searched on NCBI database by using the accession number of each 22 and 15 genes **and found most of these genes have tissue type B cell** and some of ER –ά genes.

As in our result cluster 1 contains most of these genes, which expressed by B and T cell. So this is again verifying that our results are reliable.

11. Which clustering algorithms one can use

There is no definite answer. Research is going on. But guess is that it depends upon data set.

12. References

[1] Laura J. van 't veer, Hongyue Dai, Marc J. van de Vijver, yudong D. He, Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, Karin van der Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, Rene Bernards & Stephen H. Friend, **Gene expression profiling predicts clinical outcome of breast cancer**, Nature 415 (2002), pp 530 – 535.

[2] W Dubitzky, M. Granzow, D Berrar, **Data mining and machine learning methods for microarray analysis**, Methods of Microarray data Analysis (2002), pp 5-22.

[3] "**Finding groups in data**" **An introduction to cluster analysis,**
Leonard K. and Peter J. R, A Wiley-Interscience Publication, 1990.

[4] Leping Li, Lee G. Pedersen, Thomas A. Darden and Clarice R. Weinberg, **Class Prediction and Discovery based on Gene Expression Data** (2001),

[5] http://www.quantlet.de/~hizir/xcs/html/xcshtmlnode15.html

[6] http://www.cs.uvm.edu/~xwu/kdd/Syllabus.html

[7] http://www.gene-chips.com/GeneChips.html

[8] http://ep.ebi.ac.uk/Docs/epclust/index.html