A Genetic Algorithm Approach to DNA Microarrays Analysis of Pancreatic Cancer

Nicolae Teodor MELITA[1], Irinel POPESCU[2], Stefan HOLBAN[3]
[1] *"Politehnica" University of Timisoara, Faculty of Automation and Computers, 2 V. Parvan Blvd.,*
RO-300223 Timisoara, Romania, nt_melita@yahoo.com
[2]*Fundeni Clinical Institute, Center of General Surgery and Liver Transplantation, 258 Fundeni Street,*
RO-022328 Bucharest, Sector 2, Romania, irinel.popescu@icfundeni.ro
[3]*"Politehnica" University of Timisoara, Faculty of Automation and Computers,2 V. Parvan Blvd., RO-*
300223 Timisoara, Romania,stefan@cs.utt.ro

***Abstract*—We address the problem of collecting and analyzing vast amount of information in medicine and biology, in the light of the revolutionary technological evolution during the last decades. Currently, the methods of achieving information challenge our capacity to sort and process that data. However, we use the methods of machine learning to sort and analyze this information. In this comprehensive review we describe an experiment of analyzing DNA microarrays using a Genetic Algorithm for feature selection. We study how we can establish a causal relationship between a pattern of genic expression and the evolution of pancreatic cancer using a Genetic Algorithm.**

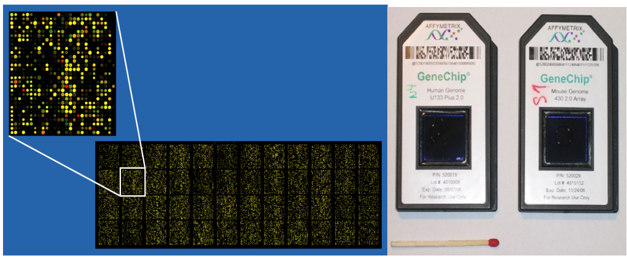
***Index Terms*—DNA Microarrays, Feature Selection, Genetic Algorithm, Suppot Vector Machines, Pancreatic Cancer**

I. INTRODUCTION

In last decades, information technology generated a revolution in medicine, in all areas, from diagnosis techniques to high level surgery procedures. In this context, we witness a spectacular revolution in genetics. The complexity of the research process became so overwhelming, that it is almost impossible these days to develop a breakthrough research in medicine without the collaboration of scientists from completely different fields. We expect that future development will provide us with new diagnostic methods and treatments capable to heal some of the worst prognosticated diseases nowadays.

The Pancreatic Cancer is still a big challenge for the medicine at the moment. The lack of an efficient screening and the unspecific simptomatology make early diagnosis almost impossible in most of the cases. Consequently to late diagnosis, the treatment is inefficient, and we witness a very high rate of mortality.

The DNA microarrays (Figure 1) are glass or plastic chips which immobilize thousands to hundred thousands samples of DNA fragments, cDNA or oligonucleotides, depending of chip construction technology.



a) b)

Figure 1. a) An example of DNA microarray, Stanford technology; b) An example of Affimetrix chip (the source of the image is wikipedia.org, a public domain).

The microarray technology allows the comparison of samples collected from normal and tumor biological probes, in terms of differentially expressed genes. In this manner, we can establish a causal relationship between a pattern of genic expression and the evolution of a malignant process; to find the markers of that specific process. The microarray technology sets the basis for very efficient screening and diagnosing cancer in an early stage of development. It could also expand into a starting point for developing new treatments for various types of cancer.

Our experiment represents a specific step in a more complex research project concerning the pancreatic cancer. The project “Gene Expression Profile and Biomarkers Study Correlated with Clinicopathological Parameters in Pancreatic Cancer” (GENOPACT) is a Romanian National Grant, CEEX 56/2005, developed by the Department of Surgery within the Fundeni Clinical Institute in collaboration with several academic and research institutes. The aim of the GENOPACT project is to discover a group of markers for the pancreas cancer, which will increase the efficiency of diagnosing the disease in early stages. Finding an optimal subset of differentially expressed genes is a very important task in achieving this goal.

II. PROBLEM STATEMENT

The problem we are addressing here is how we process the information provided by microarrays in order to achieve knowledge. Nowadays, the methods of machine learning and statistics are key factors of the research. The number of

Digital Object Identifier 10.4316/AECE.2008.02008

probes immobilized on a single array grows every year, consequently, the complexity of the analysis increases.

We are interested in finding a group of differentially expressed genes that characterize the process in pancreatic cancer. Using Affymetrix HG-U133 Plus 2 arrays, we will compare samples collected from normal and tumor cells, derived from patients diagnosed with pancreatic cancer.

The main drawbacks of microarray technology are the background noise and the insufficient sensitivity. It is very difficult to distinguish between the genes that are causally involved in the process of interest, and the genes that are differentially expressed as consequences of another process.

We will use Machine learning techniques to overcome these problems. Our goal is to select, from all the differentially expressed probes, a subset of probes that we can use to discriminate very well between the normal and tumor samples. However, the machine learning techniques give an image of the problem, but further biological validation is needed to draw solid scientific conclusions.

We will use a Genetic Algorithm (GA) to select a subset of differentially expressed genes from the microarray data, we will study how efficient this subset proves to be in discriminating between the normal and tumor samples, and we will briefly inspect the biological significance of our experiment.

III. SURVEY OF THE LITERATURE

Because we are dealing with a relatively new interdisciplinary field, the literature is devised between all the research fields involved. We are interested in a better understanding of our dataset, so we want to know about the methods of biotechnology for creating microarrays and providing data to be analyzed (Causton, Quackenbush & Brazma, 2003 [1]). Other approaches focus on the bioinformatics' point of view on methods of collecting and analyzing data (Dov Stekel, 2003 [2]). The books that focus on the specific machine learning methods help in developing an image of how the algorithms work, their strong and weak points (Ressom, 2007 [3]; Duda, P. E. Hart and D. G. Stork, 2001 [4]; I. Witten and E. Frank, 2005 [5]).

A very helpful set of documents are focused on using the specific software tools that we can use in microarray analysis with emphasis on specific features (Venables & Ripley, 2000 [6], [7], [8]; D. G. Stork and E. Yom-Tov, 2004 [9]; Nicolae Morariu, Sorin Vlad [10]; Sam Roberts [11]; Robert Gentleman, Vince Carey, Wolfgang Huber, Rafael A. Irizarry, Sandrine Dudoit [12]). These books are designed to introduce the researchers in using these software packages fast and effective.

Currently, there are several software packages that offer the tools for our analysis. The experiments presented in this review were performed in R (version 2.6.2), utilizing the Bioconductor Project. The R software and supplementary packages are freely downloadable on the official website: http://cran.r-project.org/. The Bioconductor software, all the additional packages and the documentation are available on the Project's website: http://www.bioconductor.org/.

IV. METHOD

Our experiment is a part of the project Genopact, CEEX 56/2005, developed by a multidisciplinary team, and supported by a group of healthcare providers, academic and research institutes. In this point of the research, we focus on selecting a subset of the probes that are optimal for discriminating between the normal and pancreatic cancer samples. Our analysis aims to restrict the group of genes assumed to have a causal relationship with the pancreatic cancer's evolution.

The GENOPACT dataset consists of 39 pancreatic cancer-normal sample pairs collected from patients diagnosed and monitored at the Center of General Surgery and Liver Transplantation from Fundeni Clinical Institute. The measurements were accomplished using Affymetrix HG-U133 Plus 2 arrays, resulting 78 microarray expression data.

First, we preprocessed the data using 5 algorithms (GCRMA, RMA, PLIER, MAS5, and LIWONG). The GCRMA granted the best results, so we developed our experiment based on this dataset. We assessed the quality of our data benefiting from the affy, affycoretools, affyQCReport, and simpleaffy R packages. The samples found to be problematic were removed. We continued the analysis with a dataset consisting of 70 samples.

We utilized the genefilter R package to apply a non-specific filter on the dataset, removing the probes with IQR across the samples on the log base 2 < 0.5. Furthermore, the data was filtered using the moderated t-statistics computed with the limma[13] package. The p-value=8e-09 was found to be the cut-off where the Affymetrix controls were not differentially expressed anymore. The dataset was filtered for probes with log fold change>2.0 which were differentially expressed at p-value<8e-09. The result was a new dataset with 365 features.

Finally, we used a genetic algorithm to select the best features from a dataset with 365 probes and 62 samples. The 62 samples were randomly selected, with equal proportion of normal and tumor samples. The other 8 samples, 4 of each class, were kept separately, for consecutive validation of the results. The fitness function for the genetic algorithm was implemented upon a linear discriminant classifier (LDA). The genetic algorithm was set to minimize the error rate of the linear discriminant classifier, computed using 10-fold cross validation. Our aim was to determine which probes in our dataset are the most valuable for predicting the samples' class, rather than finding the smallest subset of features that can perfectly separate the normal and tumor arrays, on this specific dataset.

After we ran the Genetic Algorithm with 200 iterations, over the training set with 62 samples and 365 probes, 45 features (Table 1) appeared with a frequency more than 18% in the optimal selected features subsets. We applied the GA implementation provided in the genalg package.

We used unsupervised, and then supervised machine learning methods to evaluate our results. We focused on the full dataset containing 54675 features and 70 samples, the dataset with 365 features and 70 samples, resulted following the filtering step, and the dataset with 45 features and 70 samples, outcome of the genetic algorithm. We wished to test if the smallest dataset, with just 45 features is efficient in discriminating the tumor from normal samples. We were also interested to compare the performance of well-known efficient classifiers on the two datasets.

TABLE 1. THE MOST FREQUENT GENES IN THE GA OUTPUT

	GeneAbbreviation	Frequency (%)	GeneName
1	FN1	65	fibronectin 1
2	GPRC5A	56	G protein-coupled receptor, family C, group 5, member A
3	CDH11	56	cadherin 11, type 2, OB-cadherin (osteoblast)
4	VCAN	49	versican
5	OLR1	47	oxidized low density lipoprotein (lectin-like) receptor 1
6	SULF1	46	sulfatase 1
7	NNMT	44	nicotinamide N-methyltransferase
8	WISP1	44	WNT1 inducible signaling pathway protein 1
9	RARRES1	42	retinoic acid receptor responder (tazarotene induced) 1
10	ASPN	36	asporin
11	FBN1	35	fibrillin 1
12	COL1A1	35	collagen, type I, alpha 1
13	MXRA5	34	matrix-remodelling associated 5
14	COL3A1	34	collagen, type III, alpha 1
15	COL8A1	33	collagen, type VIII, alpha 1
16	BNC2	32	basonuclin 2
17	CDNA FLJ38181 fis, clone FCBBF1000125	32	NA
18	ITGBL1	31	integrin, beta-like 1 (with EGF-like repeat domains)
19	CALD1	30	caldesmon 1
20	RAB31	30	RAB31, member RAS oncogene family
21	MMP7	27	matrix metallopeptidase 7 (matrilysin, uterine)
22	PALLD	27	palladin, cytoskeletal associated protein
23	COL12A1	27	collagen, type XII, alpha 1
24	INHBA	27	inhibin, beta A
25	DPYSL3	25	dihydropyrimidinase-like 3
26	SEMA3C	23	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C
27	TMEPAI	23	transmembrane, prostate androgen induced RNA
28	CDNA FLJ38472 fis, clone FEBRA2022148	23	NA
29	RUNX1T1	22	runt-related transcription factor 1; translocated to, 1 (cyclin D-related)
30	C5orf13	22	chromosome 5 open reading frame 13
31	IGFBP3	22	insulin-like growth factor binding protein 3
32	BICD1	22	bicaudal D homolog 1 (Drosophila)
33	TGM2	21	transglutaminase 2
34	COL1A2	21	collagen, type I, alpha 2
35	FBXO32	21	F-box protein 32
36	MFAP2	20	microfibrillar-associated protein 2
37	BGN	20	biglycan
38	HOP	19	homeodomain-only protein
39	ITGA2	19	integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor)
40	RAB34	19	RAB34, member RAS oncogene family
41	FER1L3	19	fer-1-like 3, myoferlin (C. elegans)
42	PRRX1	19	paired related homeobox 1
43	TGFBI	18	transforming growth factor, beta-induced, 68kDa
44	ZNF532	18	zinc finger protein 532
45	FXYD5	18	FXYD domain containing ion transport regulator 5

We analyzed each dataset with the unsupervised methods, Divisive Analysis and Partitioning Around Medoids. The diana and pam implementations respectively, offered by the cluster R package, were employed for this task. For both methods we carried out the experiments using the Euclidean distance. We also applied Multidimensional Scaling (the sammon version implemented in the MASS R package) and PCA on the datasets. Some results of the unsupervised learning phase are presented in the Appendix A (Figure 2-12).

The 8 samples excluded from the GA step were tested with a linear discriminant classifier trained on the same 62 samples set, that was presented to the GA, but with only 45 features.

We continued our analysis, illustrating the performance of two classifiers over the filtered dataset with 365 features and the one with 45 features, generated consequently to GA output analysis. For this purpose, we preferred the support vector machines (SVM) with linear kernel function, and the regularized discriminant (RDA) offered by the MLInterfaces R package. The performance of classifiers over each dataset was evaluated using 5-fold cross validation.

The results of both the unsupervised and supervised learning steps were encouraging, so it became interesting to check if our results could gain biological sense. We tested for significant pathways in our dataset using the procedure offered by the R package globaltest.

V. RESULTS

The performance of the linear discriminant classifier, trained with 62 samples, over the testing set with 8 samples is presented in Table 2.

TABLE 2. THE PERFORMANCE OF LINEAR DISCRIMINANT CLASSIFIER

	Testing set with 45 features
LDA	predicted given normal tumor normal 4 0 tumor 1 3

Both classifiers performed better on the smaller dataset, with only 45 features. The results we got for the SVM with linear kernel and RDA are presented in the Table 3. We notice that the supervised learning results were in agreement with the beliefs we had after analyzing the unsupervised learning results. The dataset with 45 features is more efficient in predicting the samples' class with linear kernel SVM or RDA classifiers.

TABLE 3. THE SVM AND RDA PERFORMANCES

	Dataset with 365 features	Dataset with 45 features
SVM	Predicted given normal tumor normal 30 5 tumor 3 32	predicted given normal tumor normal 31 4 tumor 0 35
RDA	Predicted given normal tumor normal 31 4 tumor 4 31	predicted given normal tumor normal 30 5 tumor 0 35

The test for significant pathways on the 45 features dataset showed that, even the dataset contains a very small number of genes, at lest five KEGG pathways are differentially expressed between the tumor and normal samples. The significant differentially expressed pathways in the 45 features dataset are presented in Table 4.

TABLE 4. THE DIFFERENTIALLY EXPRESSED PATHWAYS

	KEGG code	Pathway Name
1	04060	Cytokine-cytokine receptor interaction
2	04350	TGF-beta signaling pathway
3	04810	Regulation of actin cytoskeleton
4	05222	Small cell lung cancer
5	04514	Cell Adhesion Molecules

The pathways found significant in the dataset with 365 features are shown in Table 5.

TABLE 5. THE DIFFERENTIALLY EXPRESSED PATHWAYS

	KEGG code	Pathway Name
1	04060	Cytokine-cytokine receptor interaction
2	04350	TGF-beta signaling pathway
3	04810	Regulation of actin cytoskeleton
4	04540	Gap junction
5	05214	Glioma
6	05218	Melanoma
7	04010	MAPK signaling pathway
8	05222	small cell lung cancer
9	01430	Cell junction

VI. CONCLUSION

1. Both classifiers were able to predict the correct class of the samples better on the dataset with just 45 features. These results encouraged us to believe that these features are very important for predicting the cancer samples. Additional validation on new samples is needed to confirm our result.

2. Most of the genes outputted by the genetic algorithm are known to be related or involved in different types of cancers. However, further biological validation is needed to prove our results.

3. The pathways found to be differentially expressed between the tumor and normal samples, in the 45 features dataset, are notoriously involved in different malignant processes. This fact encourages us to believe that our findings have biological meaning.

4. We conclude that our approach is successful in selecting the most significant genes for predicting the samples' class. We have reasons to believe that, in the next steps of the project we can establish a very specific subset of genes causally related with the evolution of pancreatic cancer.

APPENDIX A

Unsupervised Learning Results:

Dataset 1 (70 samples, 54675 features)

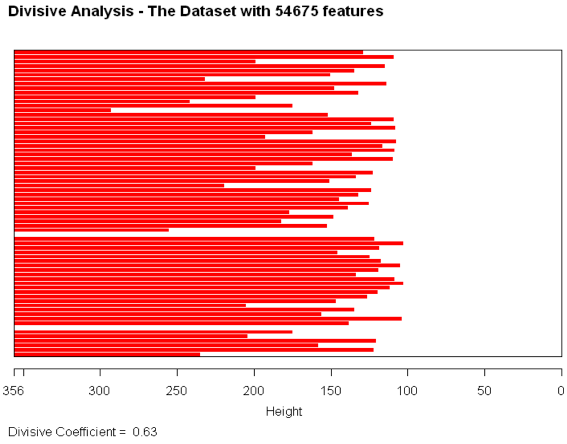


Figure 2. Divisive Analysis.

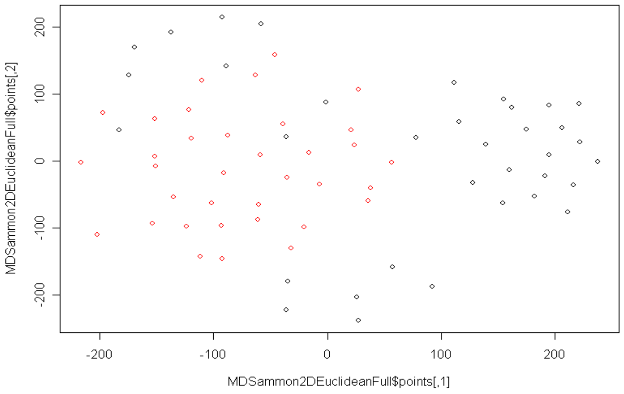


Figure 3. Multidimensional Scaling.

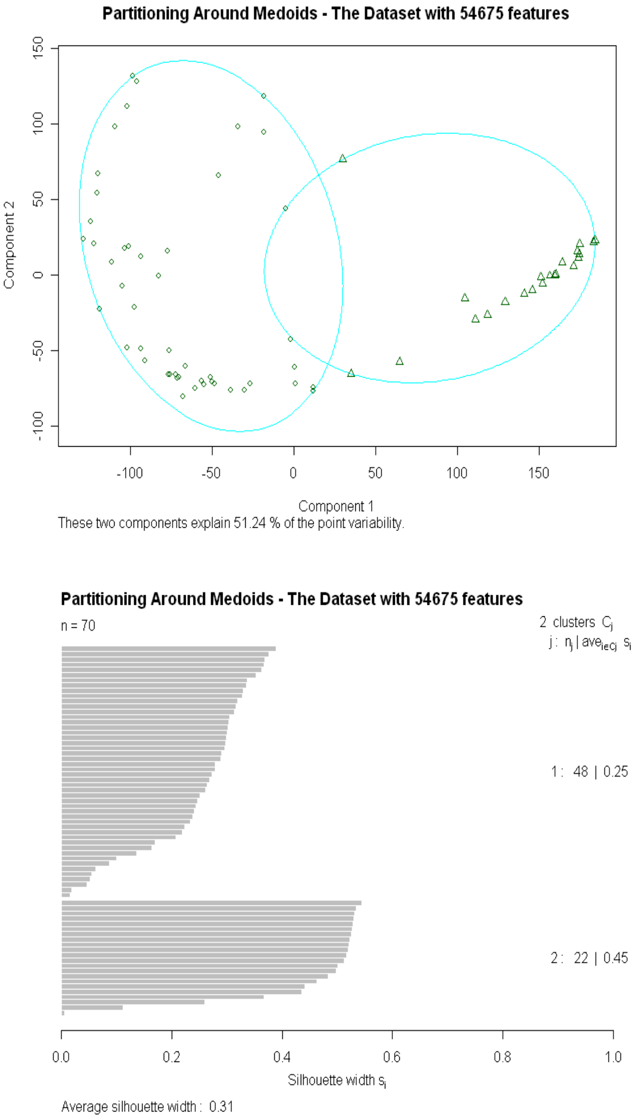


Figure 4. Partitioning Around Medoids.

Dataset 2 (70 samples, 365 features)

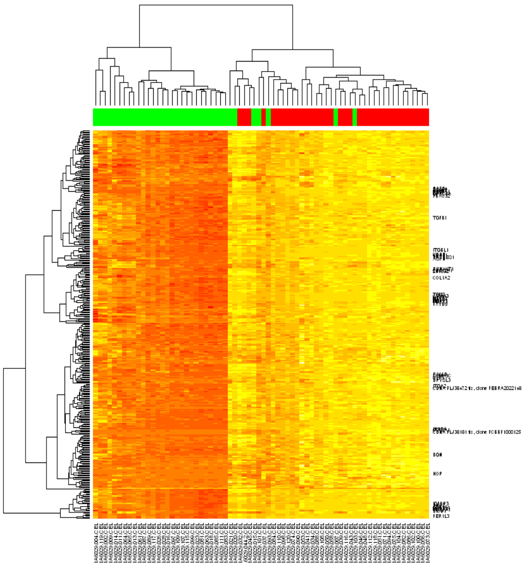


Figure 5. Heatmap and Dendrogram.

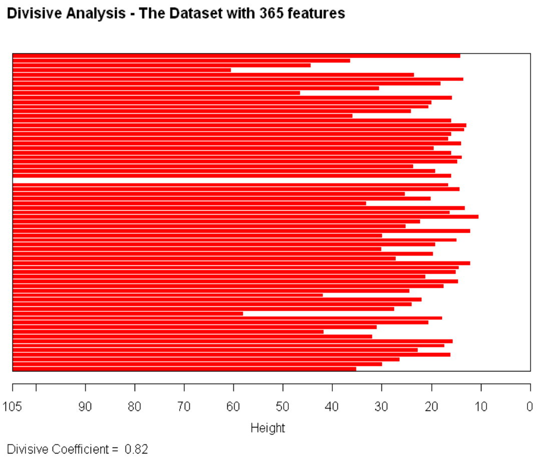


Figure 6. Divisive Analysis.

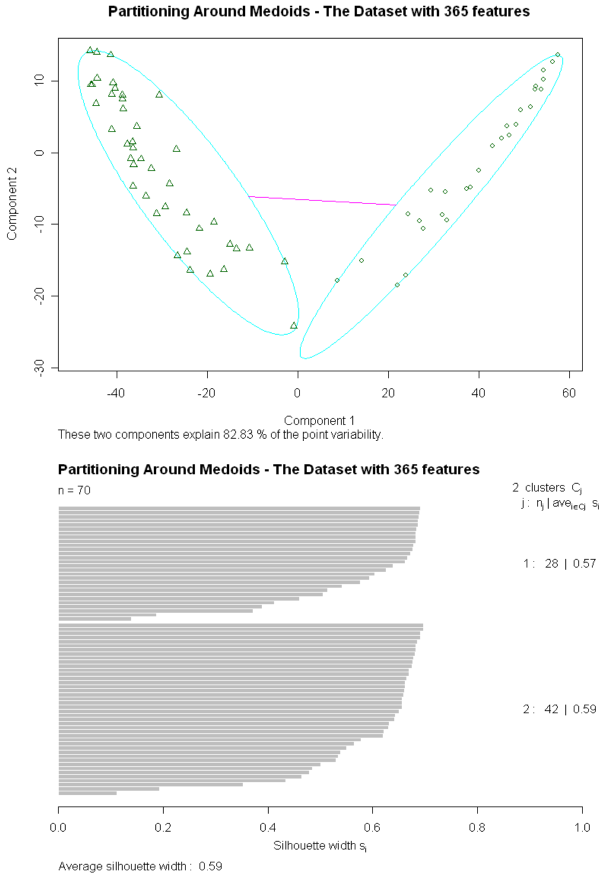


Figure 7. Partitioning Around Medoids.

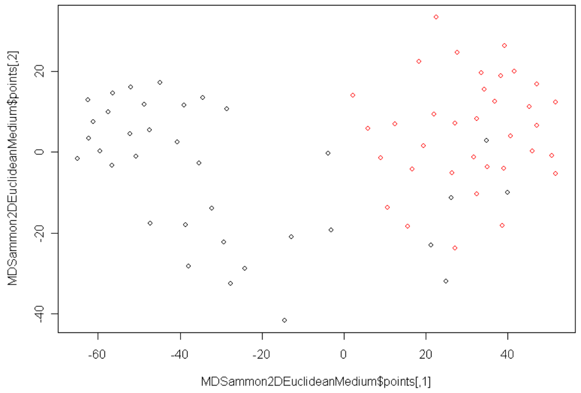


Figure 8. Multidimensional Scaling.

Dataset 3 (70 samples, 45 features)

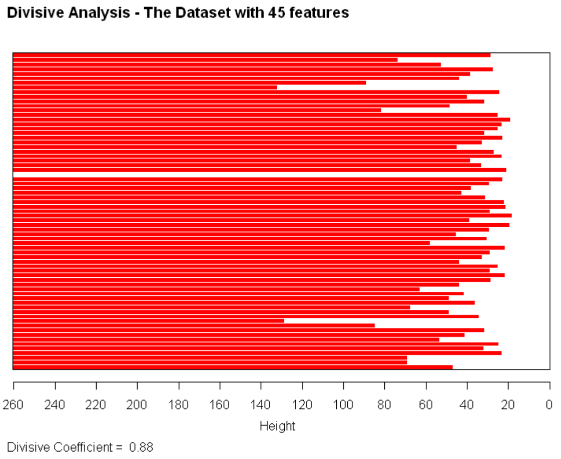


Figure 9. Divisive Analysis.

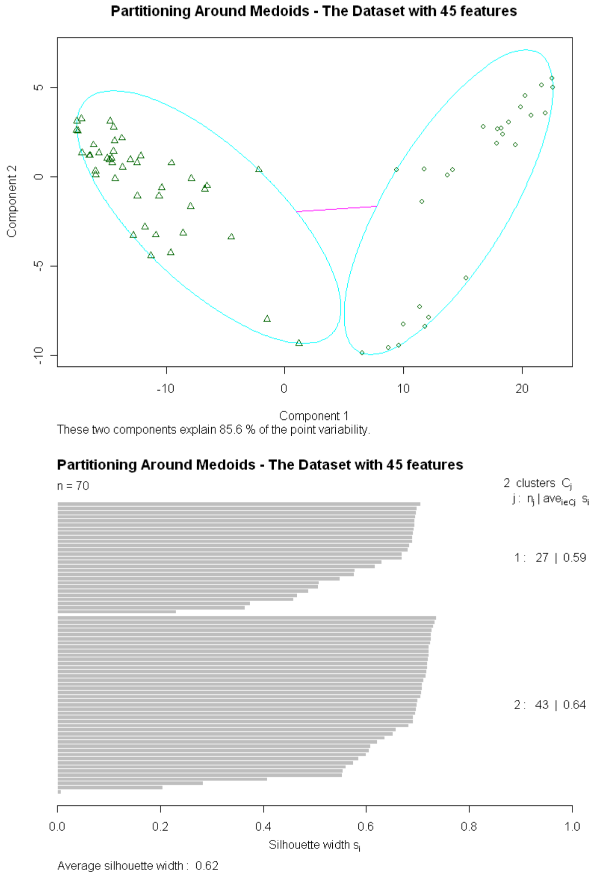


Figure 10. Partitioning Around Medoids.

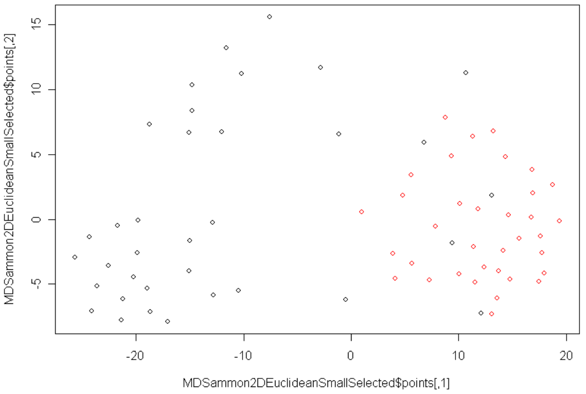


Figure 11. Multidimensional Scaling.

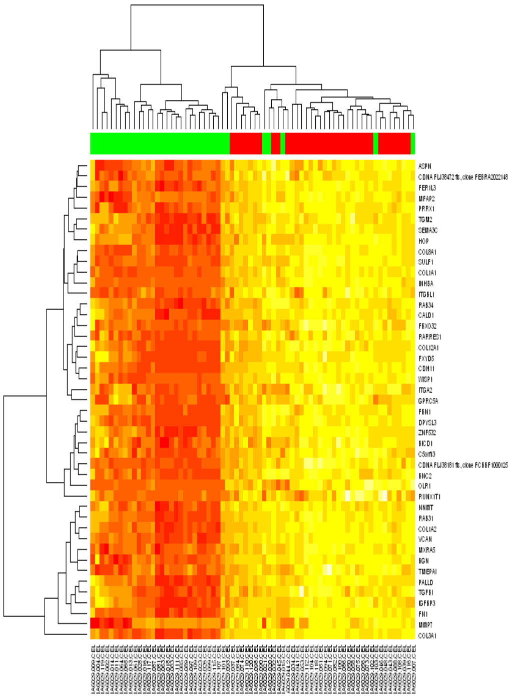


Figure 12. Heatmap and Dendrogram.

REFERENCES

[1] Helen Causton, John Quackenbush, Alvis Brazma (2003). Microarray Gene Expression Data Analysis: A Beginner's Guide, Blackwell Publishing Professional.

[2] Dov Stekel, (2003), Microarray Bioinformatics, Cambridge University Press.

[3] H. Ressom (2007), Lecture Notes, Georgetown University.

[4] R. O. Duda, P. E. Hart and D. G. Stork, (2001), Pattern Classification, Second Edition, Wiley.

[5] I. Witten and E. Frank, (2005), Data Mining (2nd Ed.), Morgan Kaufmann.

[6] W. N. Venables, D. M. Smith & the R Development Core Team (2006), An Introduction to R.

[7] William N. Venables and Brian D. Ripley (2002). Modern Applied Statistics with S. Fourth Edition, Springer, New York.

[8] William N. Venables and Brian D. Ripley, (2000) S Programming. Springer, New York.

[9] D. G. Stork and E. Yom-Tov, (2004), Computer Manual in MATLAB to Accompany Pattern Classification, Second Edition, Wiley.

[10] Sam Roberts, (2005), Using Genetic Algorithms to Select a Subset of Predictive Variables from a High-Dimensional Microarray Dataset.

[11] Nicolae Morariu, Sorin Vlad, (2007), Using Pattern Classification and Recognition Techniques for Diagnostic and Prediction, Advances in Electrical and Computer Engineering, Vol.7.

[12] Robert Gentleman, Vince Carey, Wolfgang Huber, Rafael A. Irizarry, Sandrine Dudoit (2005), Bioinformatics and Computational Biology Solutions using R and Bioconductor, Springer, New York.

[13] Smyth, G. K., (2004), Linear models and empirical Bayes methods for assessing dierential expression in microarray experiments, Statistical Applications in Genetics and Molecular Biology,Vol.3, No.1, Article 3,http://www.bepress.com/sagmb/vol3/iss1/art3