Project Report
CSE 591-Computational Molecular Biology
Instructor: Dr.Chitta Baral

Comparison of Classification Techniques of Gene Expression Profiles of Breast Cancer.

Submitted

By

NagaDeepa Vuppaladadium
Shubhra Gupta

May 12, 2003

Abstract:

DNA Microarrays Technology can detect the expression levels of thousands of genes at a time in different tissues or cells in different conditions. The genes can be functionally classified by applying machine learning algorithm to the microarrays data. In this report, we have examined two supervised machine learning algorithms - SVMs (Support Vector Machine), and C4.5 (decision tree) – for their abilities to classify relapse and non-relapse breast cancer patient samples using data collected from a microarray with ~ 25,000 genes synthesized by inkjet technology. Two kinds of Feature Selection methods namely, Gene Sampling and Wilxocon Rank Test were performed in order to select significant set of genes. The classification was performed on both datasets. Additionally we have chosen a set of genes from the NCBI website which are previously known to have been involved in Breast Cancer. The classification was performed on the datasets with and without these genes.

The results indicated that SVM outperformed C4.5 in the classification of relapse versus non-relapse patient samples for data set obtained through Wilcoxon Test. Whereas C4.5 performed better than SVM for Gene Sampling dataset. SVM did not make use of genes that were previously known to be important for Breast Cancer, in its classification. But on the other hand, Decision Tree selected some of these genes for its classification process. By studying the functional annotation of the genes selected by the feature selection methods and those selected by classifiers further conclusions can be drawn.

1 Introduction:

Breast Cancer is the second major cause of cancer death in American women, with an estimated 44,190 lives lost (290 men and 43,900 women) in the US in 1997. The treatment for breast cancer is often not easily determined. Breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome. Patients are often likely to develop a situation after their initial diagnosis when the disease relapses within an interval of 5 years. Such a condition is called a Relapse State. The patients who are from any disease beyond a period of 5 years are in Non-Relapse state.

A DNA microarray was used to do analysis on primary breast tumors of 117 young patients and machine-learning algorithm was applied to evaluate the significant genes involved in the classification process. Out of the 117 samples, 44 samples are from patients who have survived the disease for 5-year interval. 34 samples are from patients who developed distant metastases and the remaining samples are from patients who developed breast cancer from two kinds of mutations BRCA1 (18) and BRCA2 (2) and additionally 19 patient profiles.

1.1 Support Vector Machines (SVMs)

SVMs are considered a supervised computer learning method because they exploit prior knowledge of gene function to identify unknown genes of similar function from expression data. SVMs avoid several problems associated with unsupervised clustering methods such as hierarchical clustering methods and self-organizing maps. The basic idea of SVM is to construct a hyper plane as the decision surface by maximizing margin of separation between positive and negative examples.

1.2 Decision Trees (C4.5)

It uses fixed sets of attributes, and creates a decision tree to classify an instance into a fixed set of class-labels. At every step, if the remaining instances are all of the same class, it predicts that class, otherwise, it chooses the attribute with the highest "information gain" and creates a decision based on that attribute to split the training set into one subset per discrete value of the feature, or two subsets based on a threshold-comparison for continuous features. It recursively does this until all nodes are final, or a certain user-specified threshold is met. Once the decision tree is built, C4.5 prunes the tree to avoid over fitting, again based on a user-specified setting.

2 Objective:

The purpose of our project is to perform two supervised algorithms over two datasets (Obtained by different feature selection criteria) and verify the performance of the classifier if some known set of important genes were removed from both datasets.

One of the feature Selection methods was used by the original experimenters and its our attempt to verify the advantage or disadvantage of that feature selection method. We want to verify if a better method of feature selection can improve the accuracy of classification.

The second part of our project is to perform some analysis on the significant genes used by the decision tree classifier.

3 Dataset Description:

This dataset has 2 classes. "Relapse" and "Non-Relapse"
The dataset contains 34 samples of "Relapse" and 44 samples of "Non-Relapse".
There are 24,481 genes for each of the sample.
The dataset has Log10 Intensity ratio for each gene-for each sample.
The dataset has p-value for each of the gene-for each sample.

4 Methodology

There are a few tasks that had to be performed before classification and feature selection They are Normalization, Handling Missing values.

4.1 Normalization:

The dataset contained 24,481 genes data for each the Log 10 ratio value had been given and also the p-value. So the dataset has been normalized.

4.2 Handling Missing Values:

The dataset contained several missing values. So, we had to format the data and introduce a value of "Zero" for all the missing values.

4.2.1 Why 0?

The Log base 10 ratio values essentially showed the difference between the experiment and the control. we need to do feature selection based on the Log base 10 ratio- as would be seen in later sections. We also needed to perform statistical analysis.

We observed that the missing values were rather continuously placed, either for the sample or for the gene. So that meant that introducing a mean value would mean that that particular gene or sample was significant enough to that extent.

So we best decided to ignore such kind of a sample or gene and so introduced a 0.

5. Feature Selection.

Since the number of genes is around 24,481, its clear that all the genes are not meant for the purpose of classification.
So it has to be filtered.
We have performed filtering using two types.

1. Using **Wilcoxon rank test and FDR** at 50% (False Discovery Rate) to select the significant genes.
2. Using a general **Gene Sampling** method (*<u>as done by the authors</u>*)
   - Atleast a two fold difference and
   - A p-value less than 0.01

   For more than 5 samples out of the 78 samples.

5.1 Gene Sampling:

We created an interface and wrote a program, which would select those genes that would satisfy the stated condition in above section.
The program has been written in ASP and VBScript, additionally other tools involved in doing this were Front Page, HTML, MS Access Database. The web-interface (Figure1) allows the user to give the required p-value and also the kind of fold difference desired.

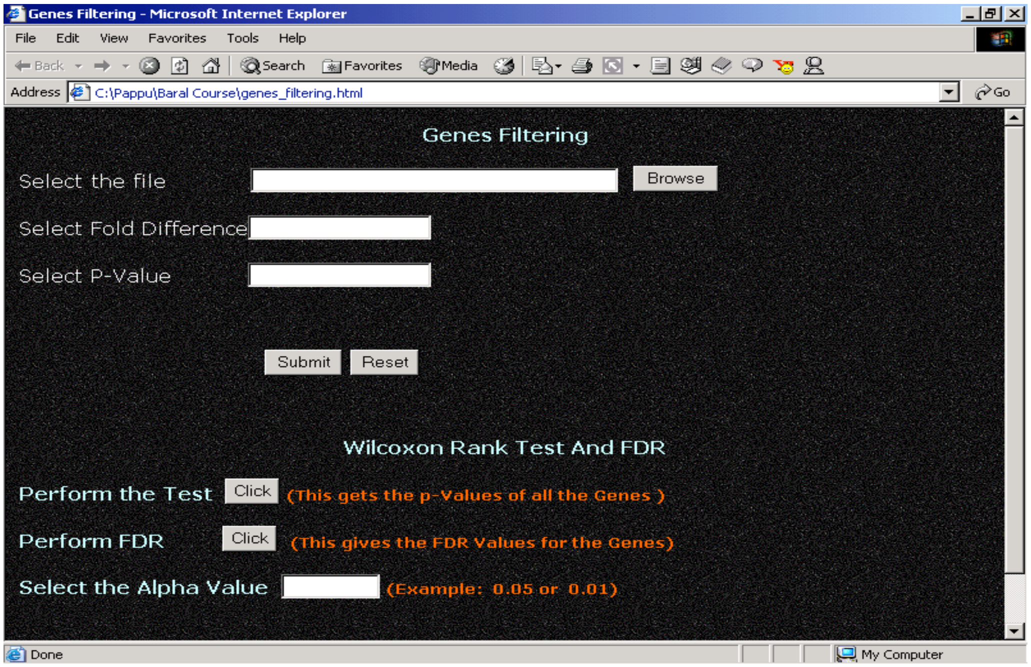

Genes Filtering - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address C:\Pappu\Baral Course\genes_filtering.html
Go
Genes Filtering
Select the file
Browse
Select Fold Difference
Select P-Value
Submit
Reset
Wilcoxon Rank Test And FDR
Perform the Test
Click
(This gets the p-Values of all the Genes)
Perform FDR
Click
(This gives the FDR Values for the Genes)
Select the Alpha Value
(Example: 0.05 or 0.01)
Done
My Computer

Figure 1: Interface for Genes Filtering.

5.1.1 Brief Description:

The Web Interface looks like in Figure 1. It is programmed to perform:
1. Genes Filtering
2. Wilcoxon Rank Test And FDR.

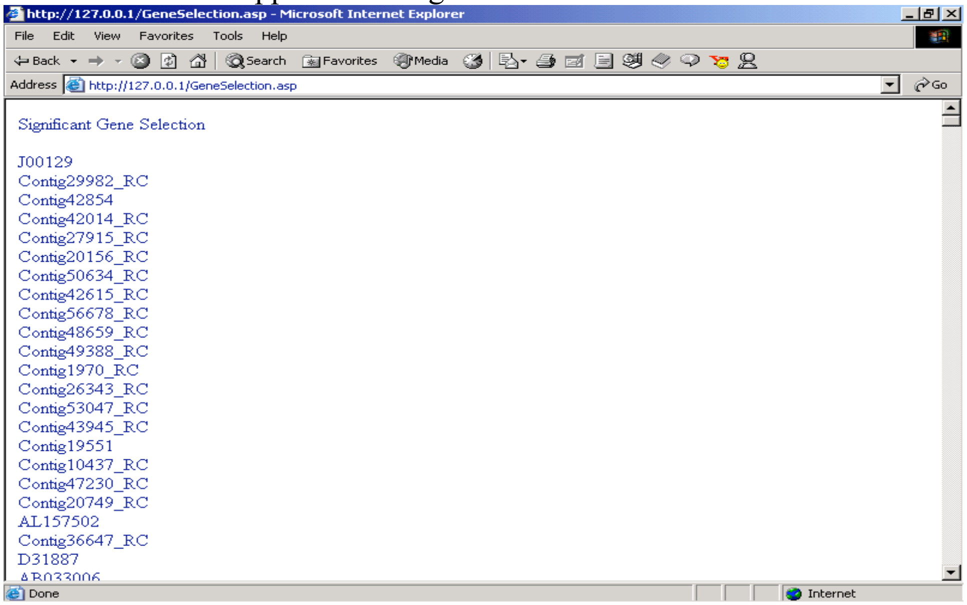
5.1.2 Fold Difference:

If the dataset contains Log Base 10 ratio values, then the 2-fold difference is 0.3. And A Log 2 Base ratio value will be 1; similarly for other (higher) folds.
So based on this condition, the user shall input the fold difference.

5.1.3 P-Value:

The dataset might contain p-values for each gene and for each sample. In that case the user shall the least p-value they would like to select. For instance 0.05 or 0.01

A typical Example would be:
Have those genes selected for which the Fold difference should be at least 2 and a p-value less than 0.01.
The results would appear as in Figure 2.


http://127.0.0.1/GeneSelection.asp - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Media
Address
http://127.0.0.1/GeneSelection.asp
Go
Done
Internet

Significant Gene Selection

J00129
Contig29982_RC
Contig42854
Contig42014_RC
Contig27915_RC
Contig20156_RC
Contig50634_RC
Contig42615_RC
Contig56678_RC
Contig48659_RC
Contig49388_RC
Contig1970_RC
Contig26343_RC
Contig53047_RC
Contig43945_RC
Contig19551
Contig10437_RC
Contig47230_RC
Contig20749_RC
AL157502
Contig36647_RC
D31887
AB033006

Figure 2: Significant Genes List

5.2 Wilcoxon Rank Test and FDR

The code for the Wilcoxon and FDR has been written in Excel Sheet. The following are the steps followed.

5.2.1 Wilcoxon Rank Test

1. Find the ranks for each of the gene-for each sample.
-There are 78 samples. So each would sample would be assigned a rank between 1-78 for that gene. This would be carried out for all the genes (24,481). Then,

2. Sum the ranks for each gene.
   -So we will get 24,481 ranks.
3. Find the Mean and Std Dev.
4. Find the Mean and Std Dev with respect to the gene's rank.
5. Find the Z statistic for each gene.
6. Find the P-value for each gene.

5.2.2 FDR

The p-value obtained may contain type 1 errors, which implies that it's possible that we might falsely reject some genes that which are actually required. So we also performed the FDR rule and selected the significant genes by controlling the False Discovery Rate at 50%.

The Null Hypotheses: There is a significant difference between the two samples for gene i.
Alternate Hypothesis: There is no significant difference between the two samples for the gene i.

Steps:

1. Arrange the p-values in ascending. So the smallest p value would mean "the most significant gene" and the largest p-value would mean "the least significant gene"
2. Apply an alpha of 0.5 and use the Benjamini and Hochberg Method to control the FDR.
   *Starting from the largest p-value $p_{(m)}$, compare $p_{(i)}$ with 0.5*i/m. Continue as long as $p_{(i)} > 0.5*i/m$. Let k be the first time when $p_{(k)}$ is less or equal to 0.5*k/m, and reject the hypotheses corresponding to the smallest k p-values.*
3. Select the Significant genes based on the Benjamini and Hochberg Method criteria.

6. Important Genes Involved in Breast Cancer from NCBI:

I selected the genes importantly involved in Breast Cancer from NCBI. These are the genes, which several authors have found to be importantly involved in breast cancer. These have been termed as 'Known' Genes.

This would be a good way of seeing if the Feature Selection included the genes that are important (as found by other authors) for the analysis of Breast Cancer.

7. Classification using SVM and Decision Trees

The Datasets have been classified using SVM and Decision Trees under various conditions.
Tools Used:
SVM: *SVMLight*
Decision Tree: C4.5

The formats have been changed appropriately in order to carry the classification.

8 Results:

The results obtained from Feature Selection and Classification is outlined below followed by discussion.

8.1 Feature Selection Results:

Result1: Using the Gene Sampling Method of Feature Selection, we obtained 4498 Genes out of 24,481.

Result 2: Using Wilcoxon Rank Test and FDR at 50% we obtained 4025 Genes out of 24,481.

	Original Set	Gene Sampling	Wilcoxon Rank Test	NCBI
Number of Genes Selected	24481	4498	4025	301

8.2 Important- Known Breast Cancer Genes Results:

Result 3: From the NCBI Website, we obtained 313 Genes importantly involved in Breast Cancer. These are termed as 'Known' Genes.

Result 4: 124 Genes of Gene Sampling Dataset belonged to the 'Known' Genes.

Result 5: 52 Genes of Wilcoxon and FDR belonged to 'Known' Genes.

Result 6: 24 Genes of Result 4 and Result 5 were in common.

8.3 Classification Results:

Result 7:

Classification of Relapse and Non Relapse Cancer Samples Using All Genes (With the known Genes And Gene Sampling)					
Classification Technique	TP	TN	FP	FN	Error Rate (%)
SVM	7	4	8	7	41.3
Decision Tree	9	7	4	6	38.35

Result 8:

Classification of Relapse and Non Relapse Cancer Samples Using All Genes (With the known Genes and Wilcoxon Rank Test)					
Classification Technique	TP	TN	FP	FN	Error Rate (%)
SVM	8	10	2	6	30.76
Decision Tree	7	4	8	7	57.69

Result 9:

Classification of Relapse and Non Relapse Cancer Samples Using All Genes (Without the Known Genes and Gene Sampling)					
Classification Technique	TP	TN	FP	FN	Error Rate (%)
SVM	7	4	8	7	57.69
Decision Tree	8	7	4	7	42.3

Result 10:

Classification of Relapse and Non Relapse Cancer Samples Using All Genes (Without the Known Genes and Wilcoxon Rank Test)					
Classification Technique	TP	TN	FP	FN	Error Rate (%)
SVM	8	10	2	6	30.76
Decision Tree	7	4	8	7	57.69

Result 11:

Classification of Relapse and Non Relapse Cancer Samples Using ONLY Known Genes (Gene Sampling)					
Classification Technique	TP	TN	FP	FN	Error Rate (%)
SVM	6	4	9	7	61.53
Decision Tree	10	6	6	4	38.46

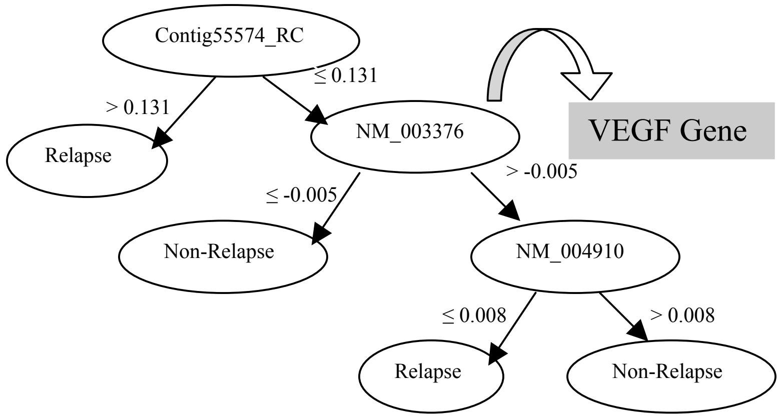
Result 12:

Classification of Relapse and Non Relapse Cancer Samples Using ONLY Known Genes (Wilcoxon Rank Test)					
Classification Technique	TP	TN	FP	FN	Error Rate (%)
SVM	10	4	8	4	46.15
Decision Tree	9	5	7	5	46.2

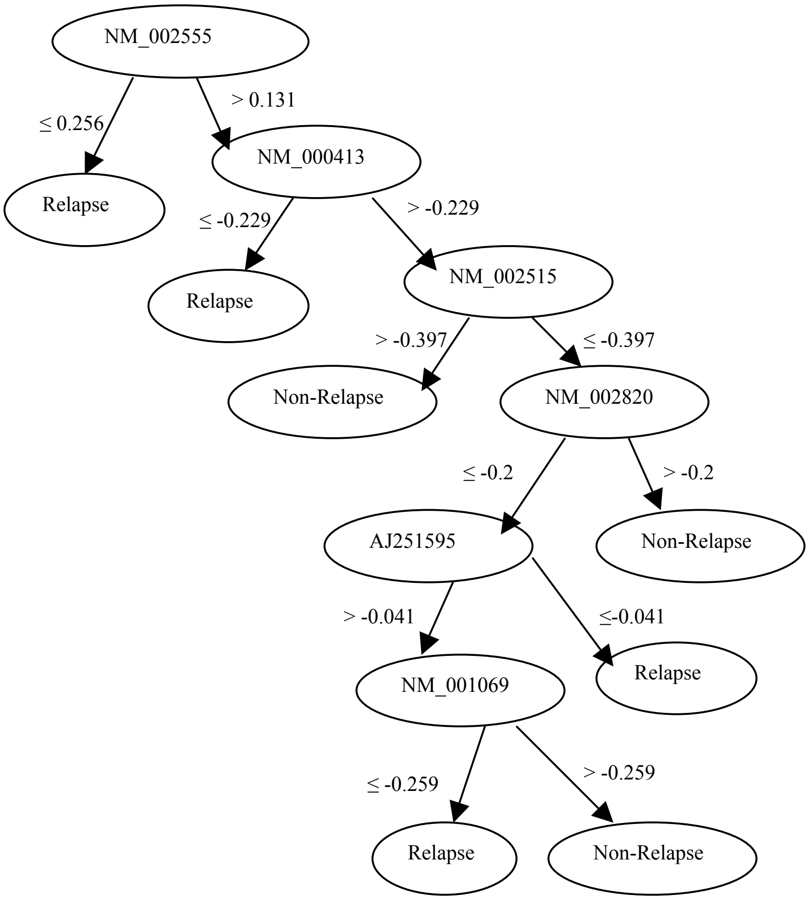
Result 13:

Accuracy Measurement

Classification Technique	With NCBI		Without NCBI		Only NCBI	
	Gene Sampling	Wilcoxon	Gene Sampling	Wilcoxon	Gene Sampling	Wilcoxon
SVM	**42.31**	**69.23**	**42.31**	**69.23**	**38.47**	**53.85**
Decision Tree	**61.15**	**42.15**	**57.69**	**42.3**	**61.53**	**57.69**



Result14: A Decision Tree for classification of Relapse and Non-Relapse patient samples using ALL genes-in Gene Sampling dataset.



Result15: A Decision Tree for classification of Relapse and Non-Relapse patient samples using ONLY the importantly known Genes (From NCBI) from the Gene Sampling dataset.

PS: Other decision trees can be drawn as well, but for explanation we have omitted the rest.

9. Discussion:

The following sections deal with the discussion of the results obtained above.

9.1 SVM Performance:

From Result 13 it can be clearly verified that SVM out performs decision tree when the feature selection is carried out by a statistical method like Wilcoxon Rank Test. The accuracy is as high as 70%.
Also the 'Known' Genes were not a part of classification for either the Gene Sampling Method or Wilcoxon rank test method, which can be noted from the accuracy values that remained exactly the same in both cases. This suggests that the classifier has used a set of genes other than the 'Known' Genes for classification purpose.

This can be cross checked with the Result 11 and Result 12, where the error rate is shown to be as high as 61% and 46.15% when only 'Known' Genes were used for SVM Classification.

9.2 Decision Tree Performance:

Interestingly decision tree has classified with greater accuracy the dataset obtained through Gene sampling unlike the SVM, which could not classify the Gene sampling dataset accurately.

However its performance on Wilcoxon Dataset was very much reduced with an accuracy of just 42.15 in comparison to SVM, which is nearly 70%.
Secondly the decision tree is unable to distinguish the 'known' Genes (From NCBI website) in its classification for Wilcoxon Rank Dataset, which implies that these genes, which are 52 in number, are not forming a part of classification; instead the classifier uses other genes different from the 'known' genes.
Whereas the decision tree could distinguish the 'Known' genes in its classification for the Gene Sampling Dataset. From the *Result 13* it can be observed that the accuracy dropped when the 'Known' were removed from the Gene sampling dataset, suggesting that some 'known' genes (the NCBI genes) was probably involved in the classification of the 78 samples.
This idea is explicated below:

9.3 Genes important in the Decision Tree Classifications of cancer tissues.

In this project the machine-learning algorithm was used not only to accurately classify the samples, but also in knowing the identities of the genes that are important in the process of classifications. These genes may be used as markers for identifications of relapse/non-relapse types of cancers and may shed light on the underlying molecular mechanisms of different types of cancers.

C4.5 can generate decision trees about whether a given sample belongs to an expected group based on the expression levels of a few number of genes.

From the *Result 14* one of the genes selected by the decision tree for classification came from the importantly known set of genes. This gene is the VEGF (Vascular Endothelial Growth Factor).
Upon verifying the Expression of vascular endothelial growth factor and its receptor (Flt-1) in breast carcinoma it is found that VEGF promotes angiogenesis by paracrining in breast carcinoma, and takes part in tumor invasion and lymph node metastasis. Blocking their secretion and effect may act as a new treatment for breast carcinoma.

9.4 Reasoning behind the behavior of SVM and Decision Tree:

Why SVM does so much better on Wilcoxon Dataset?
And Why Decision Tree does better on a Gene sampling Dataset?

A probable reason is suggested below:
The Wilcoxon Dataset selection criterion was based on " Which genes show a significant difference between *Relapse and Non-Relapse* samples?"
And the Gene Sampling Dataset selection criterion was based on "Which genes are highly expressed-irrespective of *Relapse and Non-Relapse* samples*?*"
A decision tree classifies the highly expressed genes well because these genes carry highest information gain. So the gene sampling dataset was very well classified. The Wilcoxon on the other hand may or may not have properly expressed genes and so its classification by the decision tree was not good.

The SVM however uses a separation plane for its classification. So, demarcation between samples becomes important. So the Wilcoxon was very well classified by SVM than the Gene Sampling Dataset.

Further studies of *Results 3,4,5,6* are required to determine the functional annotation of the other genes involved in the classification.

10 References:

[1] **http://www.cse.ucsc.edu/research/compbio/genex/svm.html**

[2] Michael P.S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S. Furey, Manuel Ares, Jr., and David Haussler, **Knowledge-based analysis off microarray gene expression data by using support vector machines**.

[3] Cortes, C. & Vapnik, V. (1995). **Support-Vector Networks, Machine Learning**, 20(3): 273-297.

[4] Van't Veer LJ, et.al. **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415:**530-536

[5] Gruvberger SK, Ringner M, Eden P, Borg A, Ferno M, Peterson C, Meltzer P: **Expression profiling to predict outcome in breast cancer: the influence of sample selection.** *Breast Cancer Res* 2002, **5:**23-26

[6] Quinlan, J.R., **C4.5**: Programs for Machine Learning, Morgan Kaufmann, 1993.