Identifying Mechanisms Associated with Diseases Using Classification techniques

Jagir R Hussan,
IBM Software Labs,
5th Golden Enclave,
Airport Road, BLR,
INDIA

Krithikaa G Krishnamoorthy,
Dept. of Chemical Engg.,
Birla Institute Of Technology
and Sciences, Pilani
INDIA

Gustavo Stolovitzky,
IBM T.J. Watson Research Center,
1101 Kitchawan road,
Yorktown Heights, NY
USA

Abstract: - Identification of the underlying cancer causation mechanism is extremely important in understanding how to treat the disease. The disregulation of a given cellular pathway may show up in the gene expression profile of the cell. If such is the case, computational techniques that can detect this profile change can be used to detect the pathways that have been disregulated. The present study demonstrates a classification based scheme to detect these pathways. The scheme has been applied on publicly available breast cancer data. Our results show that a mechanism that avoids immune surveillance may be implicated in the more aggressive kind of cancer studied.

Key-Words: - Functional Genomics, Classification, Human Cancer, Microarray data analysis

1 Introduction

In depth understanding of disease causing mechanisms is critical to the successful treatment of the disease. Some diseases like cancer can be caused due to disregulation in one or more mechanisms in the living cell. This is unlike diseases like cystic fibrosis or thalassamia in which a single faulty gene leads to the disease. Treatment for such diseases could be very specific, as the root cause is known. However the situation complicates when the disease (like cancer) is a result of the variation in one or more of several factors and the treatment strategy and results vary significantly based on the factors that have undergone variation. Hence it is critical to understand what factors have undergone a variation in order to provide optimal (best achievable results with minimum side effects) treatment.

The development of technologies like gene expression microarrays [1,2] has allowed us to study the behavior of cells at a genome-wide, molecular level. The data generated by these experiments provides vital information about the cellular environment. However the amount of data that is generated is very large making manual analysis difficult. This necessitates the development of computer tools to mine knowledge from the large volume of data. Currently, a diverse set of data mining techniques has been developed to discover patterns and perform diagnosis based on gene expression data [3-10]. (For a review see [11].) In this paper we propose a machine learning approach that can be used with gene expression information to implicate a specific cellular mechanism as a potential cause of a particular cellular behavior (such as cancer).

The paper has been organized as follows. Section 2 introduces the reader to gene expression data, gives a minimal account of the biology of cancer and formulates the problem being solved. Section 3 provides a brief on related works. Section 4 describes classification using support vector machines. Section 5 provides a description of the proposed solution. Section 6 describes the datasets that were used in the paper. In Section 7 we present the results obtained and in section 8 the conclusions are provided.

2 Problem Formulation

2.1 Gene Expression and Microarray data

The building blocks of the living cell are proteins. Virtually all active functions in a cell can be mapped on to a protein or a group of proteins. The type of proteins present and the interactions that exist among them determine the behavior of the cell. The proteins that are found in a cell are the result of the genes *expressed* by the cell. The expression of a *gene* that contains the code for building a protein is determined by the influence of other proteins in the cell. These interactions, referred to as *pathways or mechanisms,* are well regulated and robust in the normal cell.

When a protein has to be manufactured, a copy of the gene coding for that protein is made. Such molecular copy of the gene is called the mRNA. The mRNA is then used to produce the protein. Thus the

amount of mRNA present in the cell is to a first approximation, proportional to the amount of the corresponding protein that will be produced (the relationship is fuzzy due to other interactions that exist within the cell).

Microarray technology captures the amount of mRNA's that are present in the cell corresponding to a large number of genes in the genome. Each microarray contains thousands of probes each used to measure the quantity (*expression level*) of a specific gene's mRNA. Thus a microarray experiment provides quantitative information regarding the expression level of many genes.

When a series of microarray experiments are conducted on similar cells, one can estimate the *expression profile* of the gene and observe features like the trend in the expression of a gene, how the gene's expression is correlated to the expression of other genes, etc. Such features allow one to compare two or more cells at the molecular level and also to make inferences about an unknown cell based on its gene expression data. Sample gene expression data is shown in Fig. 1, where the logarithm of the gene expression levels for each gene (identified in the first and second columns) normalized by the amount of the same gene in a reference cell population (rows), is given for a number of patients with different types of breast cell malignancies (third and further columns).

	A	B	C	D	E	F	G
1	IMAGE_Cl	Descriptior	ADH131	ADH180	DCIS44	DCIS43	DCIS1
2							
3	1473131	TLE2 | tran	-0.68036	-0.2629	-2.18545	-1.83378	-1.3
4	79898	TLE1 | tran	-0.16818	0.130642	0.359383	1.094089	0.46
5	486179	NFAT5 | nu	-1.02709	0.166173	0.413389	-0.09876	-0.7
6	725649	NFATC4 |	0.082842	0.682898	-1.22675	-1.625	1.525
7	432072	NFATC1 |	-0.71894	0.381518	-0.07108	-0.5391	-1.3
8	2043167	BAG3 | BC	-0.11615	0.910833	1.658906	0.367419	-0.1
9	1568561	BCL2L1 | E	0.490202	0.185108	0.352815	0.48523	1.07
10	814899	BNIP3L | E	1.28835	-0.28726	1.309533	-0.61503	1.66
11	342181	BCL2 | B-c	0.410647	0.871013	0.463168	-0.83002	1.69
12	1456701	BCL9 | B-c	-0.25126	0.522915	1.667815	-0.07479	0.61
13	1568561	BCL2L1 | E	0.490202	0.185108	0.352815	0.48523	1.07
14	301976	PPP3CC |	0.580724	0.452371	0.466903	0.184942	0.43
15	840944	FCR1 Lac	-0.83225	-0.73223	-0.70458	-0.286957	-1.8

Figure 1: Sample Gene Expression data

2.2 Brief discussion of Cancer Biology

Cancer is a cellular state in which the cell begins to divide uncontrollably, a situation that leads to the creation of large masses of cells that begin to interfere with the normal physiology. In some cases these cells obtain the capability to travel through the body and start growing in other areas too (metastasis) [12]. In the normal cell the ability of the cell to divide is carefully controlled by a large number of cellular processes; also the immune system attacks those cells that act abnormally hence removing them from the system and preventing cancer. Thus for cancer to occur and proliferate, failure in the cell division regulatory processes and the ability of the cell to evade the immune system is necessary. The cells must acquire the ability to evade the immune system if it has to travel through the body; such cells are called invasive [13-22].

Cancer is a disease that results from alterations in multiple mechanisms. These alterations could be due to any one (or more than one) of the genes participating in those mechanisms. The cancerous behavior can be mapped to some of the several mechanisms that are responsible to maintain normal behavior, which in turn can be mapped to the genes that participate in these mechanisms. Hence the characterization of cancer in a cell can likely be done by observation and discovery of consistent patterns across a group of genes.

2.3 Problem Formulation

Given a particular pathway/mechanism, we wish to identify whether such pathway/mechanism behaves differentially between two cells with different phenotypes.

A generic statement would be "Given the details of the genes that participate in a regulatory mechanism and their expression profiles: identify whether a failure in the mechanism is associated with the disease state".

In this study we consider human cancer and detect whether the disruption of a particular cellular pathway, the FAS-FASL pathway [13-22], leads to invasive cancer. We choose this pathway because of its relevance in evading the immune surveillance.

We shall develop a generic machine learning approach that would in principle determine whether the given pathway operates differentially in a given disease compared to the healthy state.

3 Related Work

Currently classification techniques like Artificial Neural Networks (ANN), Support Vector Machines, K-Nearest Neighbor methods, etc. have been used to classify among different cell types based on gene expression data. Golub *et al.* [23] developed methods to classify among two cancer sub-types AML (Acute Myeloid Leukemia) and ALL (Acute Lymphoblastic leukemia). Khan *et al.* used ANN's in [24] to classify among SRBCT (Small Round Blue Cell Tumors) which are difficult to distinguish using conventional techniques. These studies and others that followed, clearly demonstrate how classification techniques can be used to differentiate among different tissues types using gene expression data. Ramaswamy *et al.* [4] studied various

classification methods to differentiate among various cancer tissue types. They also used feature selection schemes to prune the large number of variables (genes) present in the analysis.

In spite of their success at classification between cancer types, these and other similar methods do not provide information regarding the role that each selected gene plays in the context of the disease under study. Indeed, these methods do not seek to factor in existing biological knowledge into the problem. Rather, they are designed to infer rules to determine tissue type, a task essential for diagnostic applications. Used in that modality, however, no clear biological hypothesis is formulated, making it difficult for the domain experts (typically biologists) to extract novel mechanistic information with these methods.

The study by Nigam Shah *et al.* [3] attempts to integrate the biological knowledge in the feature selection process and subsequently derives inferences about the biological processes that are active in cancer cells being studied.

4 Classification and Support Vector Machines (SVM)

The task of classification occurs in a wide range of human activity. At its broadest, the term could cover any context in which some decision or forecast is made on the basis of currently available information. Autonomous systems to perform classification can be built using statistical learning techniques. Such techniques are usually applied when the amount of information to be churned to make decisions is voluminous and/or complex. These systems can then be used to make decisions on new or unseen data. Many known techniques to build classifiers perform empirical data modeling. Empirical data modeling uses induction to build up a model of the system.

SVM is an empirical data modeling technique to build classifiers [25,26]. SVM's provide attractive features and promising empirical performance. SVM's minimize the structural risk [27] and have a greater ability to generalize than methods that work by minimizing empirical risks. SVM's also addresses the curse of dimensionality problem [28].

Let us define a *n*-dimensional space, in which each point **x** represents a cancer patient. The *j*-th coordinate of this point, x_j, represents the expression level of *j*-th gene as measured in a microarray probing the cancer cells of that patient. Suppose we have patients with cancer of class 1 and patients with cancer of class -1. To find a rule that separates class 1 from class -1 points (i.e., patients), it is reasonable to try to find a hyperplane, with equation $\mathbf{w}.\mathbf{x}+b = 0$ (see Fig. 2), that neatly leaves each class at opposite sides of the plane. If that separation is possible we say that the problem is linearly separable. This is in general not feasible, but for the sake of simplicity we will restrict our discussion to linearly separable cases. A support vector machine (SVM) is based on the computation of a vector **w** and a scalar *b* that maximize the distance (margin) between the separating hyperplane and the available examples of each one of the two classes. It turns out that the hyperplane resulting from that maximization depends only on the points that are most proximal to it, called the support vectors (see Fig. 2).

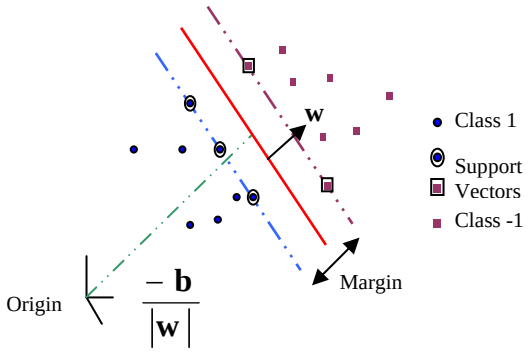


Figure 2: Schematic of hyperplane and support vectors in a sample feature space

The hyperplane, or decision boundary, can be used to compute a decision rule $f(\mathbf{x})=\text{sign}(\mathbf{w}.\mathbf{x}+b)$, which is a bi-valued function that takes the input vector **x** and returns either +1 or –1, depending on whether **x** is deemed to belong to class 1 or class -1. The practical optimization procedure that yields **w** and *b* is an interesting exercise in convex quadratic programming whose description goes beyond the scope of this paper. The interested readers are referred to Ref. [25]. It turns out that the optimal **w** can be written in terms of a set of Lagrange multipliers $\alpha_i \geq 0$ that are determined in the optimization process, in such a way that the decision function can be written as

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{m} y_i \alpha_i \mathbf{x}.\mathbf{x}_i + b\right) \qquad (1)$$

where y_i is the class descriptor (+1 or -1) of the examples $\mathbf{x}_i$, and the sum is over all the examples. (Interestingly, however, the α_i are non-zero only for the support vector examples.) Eq. (1) is valid both in the separable and the non-separable cases. However these cases differ in the constrains imposed on the α_i in the optimization process.

The discussion so far focused on the natural *n*-dimensional space given by the expression level of *n*-genes considered as our features. However given two sets of points corresponding to two cancer classes, the problem will in general be linearly non-separable in its natural dimensionality. As illustrated in Fig. 3, however, it is possible that a mapping Φ that brings the original space onto a higher dimensional space renders the two classes linearly separable in the transformed space.

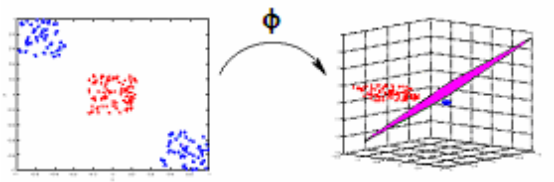


Figure 3: Schematic of a 'Φ' transformation of a 2-D non-separable sample space to a 3-D separable feature space

One of the beauties of the theory of SVMs resides in the fact that one can work on the higher dimensional space without explicitly computing Φ. This can be done because both in the optimization equations, as well as in the decision function given by Eq. (1), the problem can be formulated totally in terms of inner products of the form $\mathbf{x}.\mathbf{x}_j$. In the transformed space, these inner products would be computed as $\Phi(\mathbf{x}).\Phi(\mathbf{x}_j)$. Therefore, all the formulation of the SVM in the transformed space depends on a "kernel function" $K(\mathbf{x}, \mathbf{x}_j)= \Phi(\mathbf{x}). \Phi(\mathbf{x}_j)$. Usually, one chooses the kernel without knowing the transformation Φ. This is justified by Mercer's theorem, which states that for any *positive definite* kernel $K(\mathbf{x}, \mathbf{x}_j)$, there exists a transformation Φ to a higher dimensional space, such that $K(\mathbf{x}, \mathbf{x}_j)= \Phi(\mathbf{x}). \Phi(\mathbf{x}_j)$. Therefore, if one replaces $\mathbf{x}.\mathbf{x}_j$ by $K(\mathbf{x}, \mathbf{x}_j)$ everywhere in the algorithm, the result will be a linear SVM that lives in the higher dimensional space, but that results in a non-linear decision boundary in the original space. The decision boundary in the case of a nonlinear SVM is given by:

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{m} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b\right) \qquad (2)$$

The Kernel formulation confers considerable flexibility to the SVM approach, in that by sufficient experimentation one can find good problem-specific kernels that achieve good levels of performance in the classification problem. However, when data is scant and not too much test data is available, it is hard to determine whether a chosen Kernel will generalize well even when its performance in the limited data set was perfect. In those cases, it is better to take a parsimonious approach and use the simplest kernel. As we shall see, the amount of data to be used in the present paper is certainly limited. Therefore we used the linear SVM, given by the polynomial (degree 1) Kernel, given by

$$K(\mathbf{x}, \mathbf{x}_i) = \mathbf{x}.\mathbf{x}_i \qquad (3)$$

throughout the rest of the paper. This kernel is the simplest possible kernel available for use, as it corresponds to $\Phi(\mathbf{x})=\mathbf{x}$.

Even with the simplest kernel, the good characteristics offered by SVMs, namely its ability to generalize, the fact that its training always finds a global minimum (in contrast with other methods such as neural networks), and the simplicity and ease of interpretation of the results, made SVM the method of choice for the problem to be addressed.

5 Problem Solution

In this section the solution to the problem posed in section 2.3. is discussed by considering cancer as a model disease. However our techniques can be easily extended to other diseases. From the discussion in section 2.1 it is often the case that cancer cells differ from the normal cells very significantly at the molecular level.

With the help of microarray technology we can measure expression levels of many genes and compare them across the two types. Cancer is a multigenic disease, i.e., change in more than one gene leads to causation of cancer. Thus univariate methods may not be successful across all cases. Therefore we require a multivariate approach that takes into consideration many factors simultaneously and hence can be better adapted to the problem at hand.

Our proposed solution consists of the following two steps.

Step 1: Build a classifier that can differentiate between cancer and normal cells (or two types of related cancers) based on the expression values of the genes that are known to participate in the given pathway.

Step 2: Check classifier performance in terms of its accuracy. If the classifier shows good performance (decided on the classification error-rate and corresponding *p*-value) then we conclude that ***the pathway/ mechanism that was used to build the classifier is playing a role***, implicating the

pathway/mechanism as containing valuable information about the observed behavior.

Our solution significantly differs from existing *in silico* gene expression analysis techniques in that it allows the biologist to map his/her biological hypothesis on a set of genes and verify it using the proposed solution. This is unlike many of the feature selection techniques, where genes that are significant from an information theoretic/ statistical point of view are chosen and these selected genes do not provide direct actionable knowledge using which the underlying cellular mechanism can be identified. Thus our method significantly differs from these existing techniques and allows one to investigate at the cellular pathway/ mechanism level.

6 Analyses

6.1 Description of used data

The dataset used and reported by Xiao-Jun Ma *et al*. [8] contained gene expression samples from 61 Breast Cancer tissues (measured using cDNA technology). This dataset contained samples across various stages of the disease ADH (8), DCIS (30), IDC (23). ADH (Atypical Ductal Hyperplasia) is the stage where cells have begun to divide uncontrollably, yet do not have all the characteristics of a cancer cell. In some cases these cells later turn to be cancerous. DCIS (Ductal carcinoma *in situ*) a stage in which the cancer cells haven't yet started to invade the body. IDC (Invasive Ductal Carcinoma) a stage in which the cancer cells have become autonomous and have the capabilities to evade the immune system and travel through the body.

All the gene expression levels are measured relative to normal breast tissue cell's gene expression. In other words, the data to be used for a given gene is the gene expression of that gene in a given patient, divided by the gene expression of that same gene in a group of cells taken from normal breast tissue. From this dataset, three data subsets: ADHdf, DCISdf, IDCdf were created, where ADHdf, DCISdf and IDCdf contained ADH, DCIS, IDC samples respectively and each of these datasets contained only those genes that are related to the FAS/FASL pathway [13]. We also decided to test our procedure with a second well characterized metabolic pathway. Therefore three more data sub sets ADHdg, DCISdg, IDCdg with gene expression information of only those genes that play a role in the Glycolysis pathway were created. The details of the genes that were used in the dataset and known to play a role in the FAS/FASL pathway and the Glycolysis pathway are provided in Table 1 and 2 respectively. All the datasets was generated using the cDNA microarray technology.

6.2 Estimating Statistical Significance of results

In order to determine the statistical significance of the results, the following technique was used. The Receiver Operating Characteristic (ROC) [29] and the error rate (ER) of each classifier were determined by performing leave-one-out analysis.

The error-rate ER is given by

$$ER = \frac{FP}{FP + TN} + \frac{FN}{FN + TP} \qquad (4)$$

where,
TP – True positives identified by the classifier
TN – True negatives identified by the classifier
FP – False positives identified by the classifier
FN – False negatives identified by the classifier

Table 1: List of genes that participate in the FAS/FASL pathway and were observed in the used dataset.

Gene Id	Description
BAG3	BCL2-associated athanogene 3
BCL2	B-cell CLL/lymphoma 2
BCL2L1	BCL2-like 1
BCL9	B-cell CLL/lymphoma 9
BNIP3L	BCL2/adenovirus E1B 19kD-interacting protein 3-like
DAXX	death-associated protein 6
EGR1	early growth response 1
FAP48	FKBP-associated protein
NFATC1	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 1
NFATC4	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 4
NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive
TLE1	transducin-like enhancer of split 1
TLE2	transducin-like enhancer of split 2

Table 2: List of genes that participate in the Glycolysis pathway and were observed in the used dataset.

Gene Id	Description
ALDOC	Aldolase C, fructose-bisphosphate
C4.4A	GPI-anchored metastasis-associated protein homolog
ENO1	Enolase 1, (alpha)
GPI	Glucose phosphate isomerase
PDHA1	pyruvate dehydrogenase alpha 1

To estimate the statistical significance of the classification error rates observed in our data set, randomized datasets were generated by permuting the labels of the experiments and both the ROC and the error rate for each such randomized dataset was determined. The *p*-value for the error rate was estimated by

$$\text{p-value} = \sum_{i=1}^{n} \phi(ER_i)/n \qquad (5)$$

where n=1000 is the number of random datasets considered, ER_i is the error rate in randomized dataset i, and

$$\phi(ER_i) = \begin{cases} 0 \text{ if } ER_i > ER \\ 1 \text{ if } ER_i \le ER \end{cases},$$

where ER is the error rate for the actual dataset.

The ROC is evaluated by means of a plot of the true positive fraction (sensitivity) versus the false positive fraction (1-specificity) using a continuously varying decision threshold. Each observation (training sample) describes a point on a two dimensional plot where the ordinate indicates the fraction of class 1 examples left out in the cross-validation that fell on the correct side of the decision boundary (the class 1 side), and the abscissa indicates fraction of class -1 examples left out in the cross-validation that fell on the wrong side of the decision boundary (the class 1 side). A good classifier will have a point where the true positives rate is high and the false positive rate is low. The ROC curve in this case will lie in the upper left corner (see line **a** in Fig. 4).

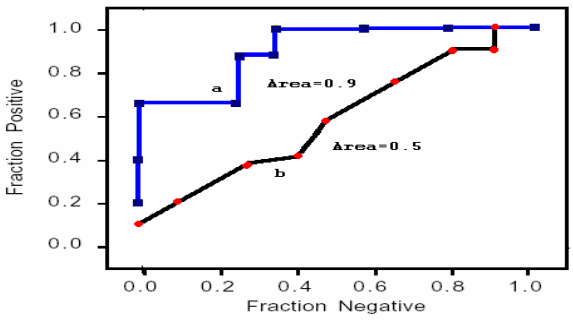


Figure 4: Schematic showing sample ROC curves

A classifier with no discrimination will have the positives and negatives mixed together and will produce a line like **b** in Fig. 4. Thus, the area under the ROC curve, referred to as the ROC score, is a measure of correct classification. An area of 0.9, for instance, indicates a good classification performance. The area under the ROC curve can be used without any transformation to examine the sensitivity and specificity of the classifier.

The *p*-value for ROC scores was determined using

$$p\text{-value} = \sum_{i=1}^{n} \phi(ROC_i)/n \qquad (6)$$

where n=1000 is the total number of random datasets, ROC_i is the ROC score in randomized dataset i, and

$$\phi(ROC_i) = \begin{cases} 0 \text{ if } ROC_i < ROC \\ 1 \text{ if } ROC_i \ge ROC \end{cases},$$

where ROC is the ROC score for the actual dataset.

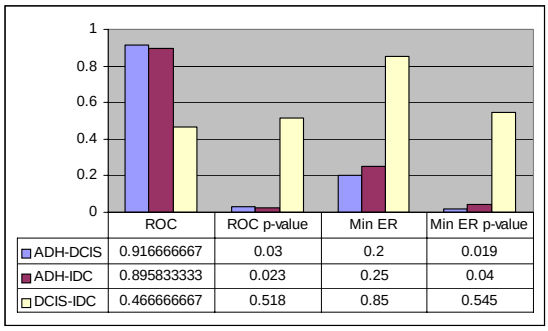
The tool Genes@Work [30] was used to perform the classification procedure and in performing the statistical significance tests. Genes@Work is available for download from www.research.ibm.com/FunGen/

7 Results

Three classifiers where built to differentiate among the classes ADH, DCIS and IDC for both the FAS/FASL and the glycolysis pathways. Three groups were built with the following pairings: group 1{ADH, DCIS}, group 2 {ADH, IDC}, group 3{DCIS, IDC}. We expected to see a very good classification in groups 1 and 2 and a reduced accuracy in group 3. The reason for this is as follows. ADH is a stage where the cells have not yet obtained the capabilities for immune evasion, whereas DCIS is a stage where the cells have obtained some capabilities for immune evasion; similarly cells in IDC stage have immune evasion capabilities [12]. Thus with respect to the FAS-FASL pathway (which, as mentioned earlier, plays a role in enabling immune evasion) samples of type ADH should differ significantly from that of DCIS and IDC. Since cells in either IDC or DCIS stage have obtained the immune evasion capability these samples must have similar expression profile and hence result in a higher error-rate as the FAS/FASL pathway does not differentiate between them.

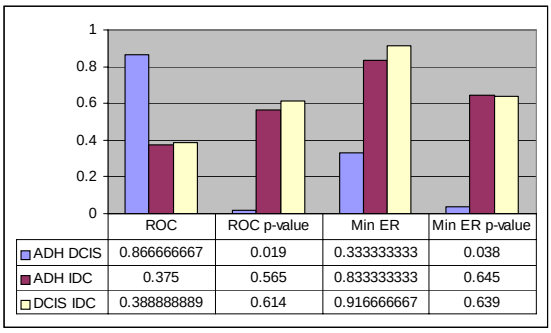
In order to justify that the classification results obtained using the pathway information are significant and are not just an artifact of the classification procedure, we selected a pathway believed not to play a role in the causation of cancer. The Glycolysis pathway is a metabolic pathway and

plays a role in generating energy in the cell. This pathway was chosen with a naïve belief that metabolic activities across the cells will remain undisturbed and will not be associated with the cancer state of the cells. Classifiers were built to perform classification using the expression data of the Glycolysis genes across the groups 1, 2 and 3.
The classifiers were built and the error-rate and ROC scores (with their respective p-values) that each group scored across the two pathways is shown in Fig. 5 and Fig. 6.



	ROC	ROC p-value	Min ER	Min ER p-value
ADH-DCIS	0.916666667	0.03	0.2	0.019
ADH-IDC	0.895833333	0.023	0.25	0.04
DCIS-IDC	0.466666667	0.518	0.85	0.545

Figure 5: Results using FAS/FASL pathway



	ROC	ROC p-value	Min ER	Min ER p-value
ADH DCIS	0.866666667	0.019	0.333333333	0.038
ADH IDC	0.375	0.565	0.833333333	0.645
DCIS IDC	0.388888889	0.614	0.916666667	0.639

Figure 6: Results using Glycolysis Pathway

From the results we can observe that ADH is well distinguished at the molecular level from DCIS and IDC. The ROC and the Minimum error-rate (p-value < 0.05) for ADH vs DCIS and ADH vs IDC is very good compared to DCIS vs IDC, confirming the biological expectation. In the case of the Glycolysis pathway, the ROC and Minimum error-rate for ADH vs DCIS indicates that the pathway is significantly different between ADH and DCIS. This was further investigated and the reason for the difference can be interpreted as follows.
The cells in the ADH stage undergo dramatic phenotypic changes and genes involved in the glycolytic function are upregulated. There seems to be a significant alteration in the metabolic functions of the cell that is getting transformed to the DCIS [10].

The above results show that the method also helps in correcting incorrect or misunderstood assumptions and the method should be used in conjunction with a body of knowledge that would enable to differentiate between causative and non-causative mechanisms.

8 Conclusions

Identification of mechanisms associated with a disease remains a challenging task. Challenges are mainly due to the lack of knowledge and effort/cost involved in verifying mechanistic hypotheses. Microarray technology has allowed scientists to capture instances of cells at various stages of behavior. A microarray experiment usually captures the expression level of thousands of genes. Direct discovery of mechanisms playing a role in a disease has so far been unsuccessful, due to high error rate in the microarray experiments, the nuances associated with normalization of the expression information, the presence of large number of variables (genes) and very low number of samples (microarray experiments).
In this study we have taken two pathways one of which is known to play a role in the causation of invasive cancer and shown proof of principle that the pathway can be detected and implicated *in silico*.

This attempt is significantly different from the existing efforts (see section 3 for details) wherein classification techniques have applied to differentiate among different tissue types. Also the existing methods do not have provision to include the existing body of biological knowledge into the classification process. Such methods do not directly provide any information regarding the underlying cellular mechanisms that result in the difference and hence cannot be used for studies where the knowledge of the underlying mechanism is critical.

The proposed machine learning approach allows one to identify putative mechanisms that contain information about the class of tissue under classification, and therefore might play a role in the causation of the disease. In doing so, we have come up with a methodology that can potentially find probable causative mechanisms from the entire pool of known/hypothesized mechanisms.
The proposed method is very sensitive and detects even low discriminative features. This is seen as a result of the analysis that was performed on the glycolysis pathway. According to our initial understanding the Glycolysis pathway was not expected to play a role. However our method showed that the pathway acts as a significant

biomarker presumably indicating the setting in of cancer during the transition period from ADH to DCIS.

The proposed technique may be helpful in choosing and targeting specific treatments which have better therapeutic effects and minimal side effects. For example antiestrogen treatment is effective only in patients with estrogen receptor positive breast cancer [31]. Hence it important to detect whether the estrogen associated pathways/ mechanisms have been disregulated before prescription of the antiestrogen based treatments.

Although the proposed method has been proven to detect pathways with sufficient correlation to the target disease causation, the process of choosing the ROC/Error-rate cutoff's and the p-value score still remains non-standardized and hence these decisions have to be done on a case by case basis. Also the method needs to be used in conjunction with a body of domain knowledge to differentiate between diagnostic and causative mechanisms.

The method developed can be easily extended to verify other mechanisms and it is possible to build a group of classifiers that would co-operatively identify the possible causes for the observed phenotype.

Acknowledgements

We are grateful to Jorge Lepre for interesting discussions, and for his help in the utilization of Genes@Work, the algorithm used to do the SVM runs. We also thank Albee Jhoney and S Venkatakrishnan for the careful reviewing of the manuscript and their useful suggestions. KGK acknowledges support from an IBM internship.

References:

[1] Brown, P. O. and D. Botstein. "Exploring the new world of the genome with DNA microarrays." *Nat Genet 21(1 Suppl)*: 33-7 1999.

[2] Lockhart, D. J., H. Dong, et al. "Expression monitoring by hybridization to high-density oligonucleotide arrays." *Nat Biotechnol 14(13)*: 1675-80,1996.

[3] Nigam Shah et al, "Can We Identify Cellular Pathways Implicated in Cancer Using Gene Expression Data?", *Proceedings of the Computational Systems Bioinformatics (CSB'03)*

[4] Sridhar Ramaswamy et. al, "Multi-class Cancer Diagnosis Using Tumor Gene Expression Signatures", *Proc. Natl. Acad. Sci. USA*. 98 15149-15154

[5] David G. Beer et al, "Gene-expression profiles predict survival of patients with lung adenocarcinoma", *Nature Medicine*, Vol. 8, No. 8, August 2002

[6] Alon U et al, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", *Proc. Natl. Acad. Sci. USA*. 96 6745-6750

[7] Califano A. et al, "Analysis of gene expression microarrays for phenotype classification". *Proc Int Conf Intell Syst Mol Biol*, 2000. Vol 8 p75-85.

[8] Xiao-Jun Ma et al,,"Gene expression profiles of human breast cancer progression", *Proc. Natl. Acad. Sci. USA*, 100 5974-5979

[9] Sorlie T et al, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications", *Proc. Natl. Acad. Sci.* USA, 98 10869-10874

[10] Porter DA et al, "A SAGE (Serial Analysis of Gene Expression) View of Breast Tumor Progression", *Cancer Res*. 2001 Aug 1; 61(15):5697-702

[11] Stolovitzky, G. "Gene selection in microarray data: the elephant, the blind men and our algorithms." *Curr Opin Struct Biol 13(3): 370-6, 2003.*

[12] L.M. Franks et al, "Introduction to the Cellular and Molecular Biology of Cancer", *Oxford Univ. Press*.

[13] Joe O'Connell et al, "The Fas Counterattack: Cancer as a site of immune privilege" *Immunology Today* 1999, 20:46-52

[14] M M Kavurma and L M Khachigian, "Signalling and Transcriptional control of Fas ligand gene expression", *Cell death and differentiation*, Jan 2003, 10: 36 –44.

[15] Fruman DA et al, "Characterization of a mutant calcineurin A alpha gene expressed by EL4 lymphoma cells", *Mol Cell Biol*. 1995 Jul;15(7):3857-63.

[16] Yamamoto K et al, "BCL-2 is phosphorylated and inactivated by an ASK1/Jun N-terminal protein kinase pathway normally activated at G(2)/M", *Mol Cell Biol*. 1999 Dec;19(12):8469.

[17] Paul R. Mittelstadt and Jonathan D. Ashwell, "Cyclosporin A-Sensitive Transcription Factor Egr-3 Regulates Fas Ligand Expression", *Mol Cell Biol*, July 1998, p. 3744-3751, Vol. 18, No 7

[18] Rengarajan J et al, "Sequential involvement of NFAT and Egr transcription factors in FasL regulation", *Immunity* 12: 293-300

[19] Kavurma MM et al, “Sp1 phosphorylation regulates apoptosis via extracellular FasL-Fas engagement”, *J. Biol. Chem*. 276: 4964-4971

[20] Xiao S et al, “FasL promoter activation by IL-2 through SP1 and NFAT but not Egr-2 and Egr-3”, *Eur. J. Immunol*. 29: 3456-3465

[21] Matsui K et al, “Identification of two NF-kappa B sites in mouse CD95 ligand (Fas ligand) promoter: functional analysis in T cell hybridoma”, *J. Immunol*. 161: 3469-3473

[22] Kasibhatla S et al, “Regulation of fas-ligand expression during activation-induced cell death in T lymphocytes via nuclear factor kappaB”, *J. Biol. Chem*. 274: 987-992

[23] Golub, T.R, et al. (1999), Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531-537.

[24] Khan, J., et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med, 2001. 7(6): p. 673-679.

[25] Christopher J.C. Burges, “A tutorial on Support Vector Machines for Pattern Recognition”, *Kluwer Academic Publishers*.

[26] V. Vapnik,”The Nature of Statistical Learning Theory”. *Springer-Verlag*, New York.

[27] S. R. Gunn et al,”Network Performance Assessment for Neurofuzzy Data Modelling”. *Lecture Notes in Computer Science*, 1280:313-323, 1997

[28] R. Bellman, “Dynamic Programming”. *Princeton University Press*, Princeton,NJ, 1957

[29] Kelly H. Zou et al, “Statistical Validation Based on Parametric Receiver Operating Characteristic Analysis of Continuous Classification Data”, *Academic Radiology*, Vol 10, No 12, Dec 2003

[30] Lepre J et al, “Genes@Work: An efficient algorithm for pattern discovery and multivariate feature selection in gene expression data”, *Bioinformatics (2004)*, in press.

[31] Roelof J. Bennink et al, “In Vivo Prediction of Response to Antiestrogen Treatment in Estrogen Receptor-Positive Breast Cancer”, *Journal of Nuclear Medicine*, Vol. 45 No. 1 1-7, Jan 2004