SIBIOS AS A FRAMEWORK FOR BIOMARKER DISCOVERY USING MICROARRAY DATA

Bhavna Choudhury

Submitted to the Faculty of the School of Informatics

in partial fulfillment of the requirements

for the degree of Master of Science

in Bioinformatics

Indiana University

August 2006

Accepted by the Faculty of Indiana University, in partial

fulfillment of the requirements for the Degree of Master of Science in Bioinformatics

_____

Malika Mahoui, Ph.D., Chair

Master's Thesis Committee

_____

Mahesh Merchant, Ph.D.

_____

Narayanan Perumal, Ph.D.

ACKNOWLEDGEMENTS

I would like to give my special thanks to Dr. Mahoui, for this research opportunity and her insightful guidance throughout. I thank my thesis committee members, Dr. Merchant and Dr. Perumal, for their continued guidance. I would also like to extend my gratitude to the faculty and staff of the Indiana School of Informatics, Indianapolis and Dr. Miled for their support. It was mainly funded by IUPUI Seed Grants for NIH "Roadmap" Initiative from the Office of Research and Sponsored Programs, at Indiana University Purdue University in Indianapolis. Other supporting funds include CAREER DBI-DBI-0133946 and NSF DBI-0110854.

I would also like to extend my sincere thanks to Dr Susanne Ragg and to Saima Zaidi, for their valuable advises. I thank Bing, Sriram and Xiang, for their input and help.

Very special thanks to my friend Gunjan, for his selflessness and unconditional support through all these years of my life. I sincerely thank him for his constant support and encouragement. I thank my parents for their guidance and assistance in helping me obtain this important milestone in my life.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Bhavna Choudhury

SIBIOS AS A FRAMEWORK FOR BIOMARKER DISCOVERY USING MICROARRAY DATA

Decoding the human genome resulted in generating large amount of data that need to be analyzed and given a biological meaning. The field of Life Sciences is highly information driven. The genomic data are mainly the gene expression data that are obtained from measurement of mRNA levels in an organism. Efficiently processing large amount of gene expression data has been possible with the help of high throughput technology. Research studies working on microarray data has led to the possibility of finding disease biomarkers. Carrying out biomarker discovery experiments has been greatly facilitated with the emergence of various analytical and visualization tools as well as annotation databases. These tools and databases are often termed as *bioinformatics services.*

The main purpose of this research was to develop SIBIOS (System for Integration of Bioinformatics Services) as a platform to carry out microarray experiments for the purpose of biomarker discovery. Such experiments require the understanding of the current procedures adopted by researchers to extract biologically significant genes.

In the course of this study, sample protocols were built for the purpose of biomarker discovery. A case study on the BCR-ABL subtype of ALL was selected to validate the

results. Different approaches for biomarker discovery were explored and both statistical and mining techniques were considered. Biological annotation of the results was also carried out. The final task was to incorporate the new proposed sample protocols into SIBIOS by providing the workflow capabilities and therefore enhancing the system's characteristics to be able to support biomarker discovery workflows.

1. INTRODUCTION

1.1. Importance of Biomarkers

Every human cell consists of a nucleus which is a storehouse of the chromosome, containing the genetic code i.e. DNA (Deoxyribonucleic Acid). The biological human functions are based on the code read from the DNA. DNAs are large doubly stranded structures that contain thousands of genes. DNA maps to mRNAs which eventually code for proteins which have a certain structure and function. It is considered normal if the mappings between the mRNA and proteins are coded properly. The level of mRNA produced, which in turn controls the protein production, corresponds to what is called *gene expression*. Typically, mutations or abnormal coding in the DNA can cause variations in the gene-expression levels making it either up-regulated or down-regulated compared to the normal gene. These changes also affect the protein levels in the body. These global expression patterns that allow us to understand gene-gene interaction networks can be monitored by microarrays.

Differentially expressed genes and proteins in tumor cells can be identified using approaches like genomics, transcriptomics and proteomics. These are the available technologies and resources that empower the process to identify significant genes.

Recently, with the advent of microarray global technology for DNA (1990), it is possible to measure RNA expression levels with extreme ease and precision. DNA microarrays are useful for assessing the expression of mRNAs from a set of biological samples. Gene

expression analysis can contain large amounts of information related to the DNA sequence, state of the cell, biological phenomena like cell growth and development, disease progression etc. Studying DNA microarrays for gene-expression analysis can help researchers and biologists understand the genomic composition of the biological sample better and discover disease causing agents. Genes are involved in the genetic pathway of the disease. Such subsets of genes can correspond to gene signatures that provide substantial information about the specific biological processes. To obtain these genes, expertise in the field of biology and statistics is vital, as well as advanced bioinformatics tools that facilitate such data analysis and research, as shown in figure 1. Mining such information from DNA microarrays requires an extensive analytical approach.

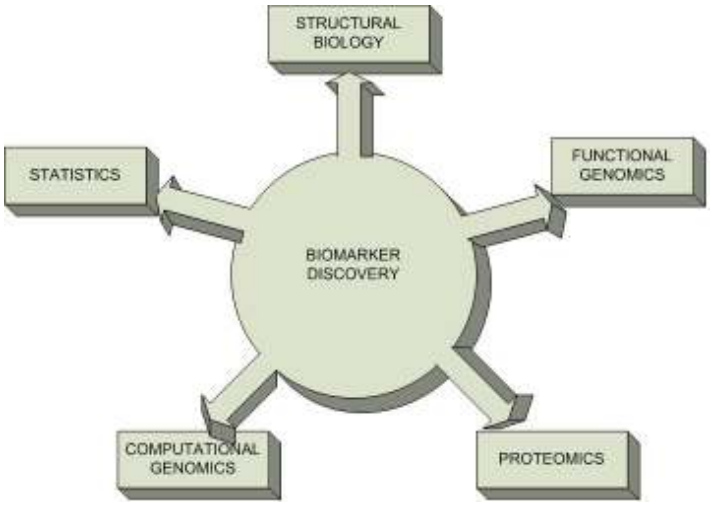


Figure 1: Biomarker discovery in a cross-disciplinary domain

Genomics involves the study of the genome of an organism. The gene intensity level measured is called *gene expression*. Patterns of the gene expressed in a cell reflect the characteristics of the state of the cell. The genes involved in these patterns can be identified

by adopting various methods available. Some of the techniques are cluster analysis such as hierarchical clustering and partition clustering, and machine learning approaches such as support vector machines.

The analysis of genomic data therefore can be done in various ways. Extracting meaningful biological information with the aid of microarray technology in order to uncover the hidden rationale for disease development and progression opens doors to a whole new research area called as *Biomarker Discovery.* Effective and corroborated procedures in gene-expression analysis in the *biomarker discovery* process can be extremely beneficial for disease diagnosis and drug discovery and development. These procedures provide powerful information that can accelerate the identification of causative genes of various diseases.

Since proteins are also very important in the body's functions, proteomics study is also instrumental in the discovery of biomarkers to indicate a specific disease. Methods that are involved in proteomics include 2D-PAGE, MALDI, LC-MS/MS, protein arrays, tissue arrays etc (R Fisler, 2005). Proteomics is essentially the large scale study of proteins, mainly their structure and function (Wilkipedia, 2001) in a global manner. It is often seen as the following step to genomics and is also quite more complicated compared to it. In 2004 a research study was conducted that used serum biomarkers for early detection of diseases. In this study (J Donald, 2004) a new class of biomarkers was obtained from mass spectrometry analysis and did show improved results in early disease detection. Comparative 2d-gel technology combined with mass-spectrometry was used by a cohort of scientists (M Carpenter, 2004) to discover unique proteins and pathways for biological systems. 2-D gel proteomics potentially

presents thousands of proteins which cover a dynamic range of expressions. This group devised data processing methods obtain interesting information. There are various online protein databanks that provide extensive information on the proteins and help in validating results. Some of them are UniProt (Uniprot, 2006), PIR (PIR, 2005), Swiss-Prot (SwissProt, 2006) and PDB (PDB, 2003).

According to the Food and Drug Administration (FDA), a biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biologic or pathogenic processes or pharmaceutical responses to a therapeutic intervention. The detection of this substance can indicate a state of disease. There are many ways in which biomarkers are useful in the medical domain of science:

- Biomarkers are used for drug discovery research, which involve the exploration of appropriate chemicals that target the disease biomarker.
- Biomarkers are used in clinical trials, as they provide knowledge about the genes involved in the pathway of the disease.
- Biomarkers help in the determination of the tumor type, prediction of survival and prediction of response to treatment.
- Biomarkers aid in the early diagnosis and prognosis of a disease.

The main issue related to discovery of significant genes for a particular disease is the large amount of data involved. A microarray chip contains thousands of probe sets with their expression values and their detection calls. The major hurdle is finding the set of genes

primarily responsible for the disease. For this purpose the list of probe sets needs to be brought down to a smaller and more manageable list.

To reduce the number of genes, meaningful analysis is required. The relevant genes are widely distributed and the manual study of each and every gene becomes an infeasible task. An enhanced automated process to is needed to discard irrelevant genes and preserve the important and interesting ones from the initial set of thousands of genes. Simultaneous with the exponential growth of available data, the number of available analysis and visualization tools has increased dramatically as we have many annotation databases (herein after referred to as *bioinformatics services*). These *bioinformatics services* assist in condensing the initial large set of genes to a list of significant ones. To carry out *in-silico* experiments researchers need a platform where they can build their workflow processes. Building *in-silico* experiments, is a cumbersome and error prone process as researchers need to resolve the heterogeneity at the semantic and syntactic level, that is involved in bioinformatics services. *In-silico* refers to the use of computer simulation. SIBIOS, a System for Integration of BIOinformatics Services is developed for biological researchers. The aim of this research was to develop a workflow-based integration system for microarray research, based on SIBIOS. Biomarker discovery used in the environment of workflows is ideally suited for automated workflows since there are thousands of identical repetitive steps for each cancer or disease studied.

1.2. Introduction to SIBIOS

SIBIOS (B. Miled, 2004) is a workflow-based system built with the aim to provide researchers with an environment conducive for biological research. It involves access and execution of various online bioinformatics tools and biological databases that can be combined in an intelligent manner for performing complex queries. It is supported by an ontology that is the description of each individual service which is a part of SIBIOS. This feature is used as a knowledge base to resolve heterogeneity and interoperability between different bioinformatics tools.

The SIBIOS system integrates online web tools and therefore deals with the heterogeneity issues persisting in the field of bioinformatics. The formats of the biological data do not match even though they mean the same thing. SIBIOS makes working on biological data easier so that researchers do not have to worry about the inherent discrepancies in the formats in the data. SIBIOS provides researchers an environment and capabilities that allow them to define their *in-silico* experiments in terms of workflow specification, and to control and automate them. A sample example workflow supported by SIBIOS is illustrated in figure 2.

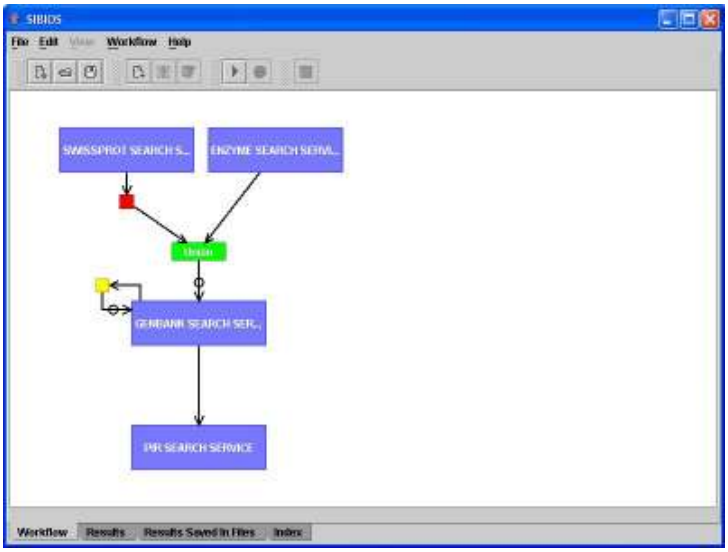


Figure 2: Sample workflow in SIBIOS

The workflow shown in figure 2 consists of 4 bioinformatics services. A protein is queried in the Swissprot and Enzyme search services. The union of the results is fed into Genbank which is recursively executed until a certain condition is attained. Genbank's outputs are then finally queried in PIR for the final results.

The field of biomarker discovery has an increasing interest and involves the use of data analysis and access to various biological databases for gene information and pathway knowledge. Developing such workflows has been the current focus area in the development of SIBIOS. Biomarker discovery is an example of such *in-silico* experiments which have been carried out using this system.

Composition of workflows using the SIBIOS framework can aid in biomarker discovery. They involve a series of repetitive steps which can be automated for the cancer under study. This platform will make SIBIOS a good tool for researchers to work on as it will provide

services that include both life science database search services and http based bioinformatics applications. The discovery of biomarkers is therefore highly suited for workflow based systems like SIBIOS, since a large number of tasks can be automated once the researcher has developed his/her design of the protocol.

Biologists can now use SIBIOS to execute their experiments using its new capabilities.

1.3. Research Challenges

Microarray technology is a recent advancement in the field of bioinformatics. With the advent of the genome project, there has been a tremendous advancement in the knowledge of human genomic sequences as well as other organisms and the genes corresponding to them. A number of techniques have been developed for acquiring gene-expression patterns on microarray chips. Many experiments are devised to study these gene patterns for their cellular responses, biological processes and molecular functions. The simplest way to examine these microarray datasets is to find up-regulated and down-regulated genes in an altered state compared to the normal state.

The field of *Biomarker Discovery* is new and several approaches to identify biomarkers are being proposed and tested. *Biomarker Discovery* can be viewed as a three phase process as shown in figure 3, i.e. The first phase being data preparation which involves various pre-processing techniques. The second phase involves the application of various data mining and statistical tools that shorten the input gene list to a small list consisting of manageable and

significant items. The third phase is the exploratory phase where biomarkers are validated using clinical tests, and wet lab experiments. Phase I is a fairly stable phase. Phase II is less stable and various approaches have been proposed to carry out this phase. Different studies undertaken on the same dataset may lead to a different potential biomarker.

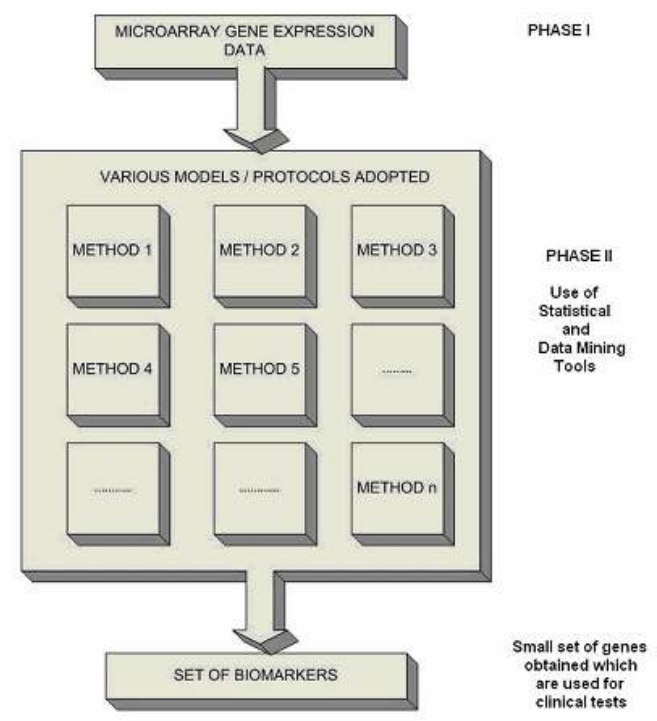


Figure 3: Different approaches adopted by different researchers

To provide researchers with a workflow framework one needs to study a number of approaches, analyze them in detail and then determine a protocol that highlights the number of steps and the tools that can be used at each of the steps. The main challenge that undermines adapting a workflow-based system such as SIBIOS is the lack of standard protocol models that can be used as templates for proposing the basic building blocks for *biomarker discovery*. Therefore the integral part of the solution is to devise models for *biomarker discovery* experiments. As a proof of this concept, a specific type of cancer is

studied here and used to design a protocol that leads to construction of a sample model to be added to SIBIOS and at the same time aim for the discovery of biomarkers.

Understanding the features of the dataset and the experiment's goal is essential to apply the experimental procedures correctly. The design and implementation of a successful software program for correct protocols is a challenge for efficient analysis and data collection. These protocols involve a number of statistical methods and tests. Inappropriate statistics can be misleading hence; statistical results should be interpreted in the context of the experimental study and its purpose. Otherwise issues may arise related to the microarray results as they depend on the validity of the statistics applied. Certain considerations before applying statistical tests generally involve assumptions on the normality of distribution of the data. Such factors become the basis of applying parametric or non-parametric tests.

Various approaches exist to carry out biomarker discovery. The main goal is to identify a list of significant genes that have relation to the disease under study. The approaches can be classified into two techniques:

- Usage of data analysis first and then using annotations to validate biomarkers.
- Utilize annotations first and then use the analytical approaches to obtain a set of potential biomarkers.

The second challenge is that SIBIOS has been designed as a generic open-ended workflow integration system; and as such it focuses on meeting several requirements for building *in-*

silico experiments. Adapting SIBIOS to biomarker discovery creates new research challenges, such as the ability to manipulate analytical tools, maintain sessions while using http-based *bioinformatics services*, facilitate the use of stand-alone tools, etc..

The difficulty in establishing a protocol constitutes a challenge to adopt SIBIOS as a platform for *biomarker discovery*. The issues involved mainly are:

- Understanding the microarray dataset.
- Tackling the complexity of the analytical tools.
- Pre-processing factors affecting the results (i.e. whether the data is normalized or not, or if it has been transformed appropriately)

The results of *biomarker discovery* depend on such factors.

1.4. Contributions

This thesis addresses the main challenges and provides the following contributions to this solution of *biomarker discovery* using SIBIOS.

1. Development of sample models of the workflow that can be used as a template for a *biomarker discovery* pipeline.
2. Adaptation of SIBIOS to allow the building of workflows supporting such workflows.
3. Identification of new interesting biomarkers targeting BCR-ABL, a subtype of Acute Lymphoblastic Leukemia.

To develop sample protocols, a case-study was identified to explore and validate the approach used. A microarray dataset of 132 samples from the St Jude's organization (M E Ross, 2003) was selected as a target dataset for the purpose of this research.

The validation of the new proposed protocol resulted in the identification of potential biomarkers as the direct or indirect "disease causative agents."

2. BACKGROUND

2.1. Related Research

2.1.1. Workflow environment for in-silico experiments

A considerable amount of work has been done in the field of bioinformatics particularly on microarrays. According to the Nucleic Acid Research Database issue published in January 2005 (Galperin, 2005), more than 850 biological databases exist, with around 548 in 2004 and 368 in 2003. Consistent with the increasing growth of available data has been the growth and development of bioinformatics analysis and visualization tools. In spite of the abundance of accessible online bioinformatics tools, designing and performing *in-silico* experiments has been an inefficient and laborious task. The main reasons have been:

- The highly diverse nature of the data stored in biological databases.
- The inherent heterogeneity within the data.
- The immense querying capabilities provided by other resources.
- The incompatibility of data between the different *bioinformatics services*.

Because of these factors, using *bioinformatics services* has been quite difficult as well as time consuming. Typically for the identification of biomarkers in the blood of a diseased patient, a researcher utilizes the gene expression data available from various public online databases. The potential markers most likely lie in the cancerous tissues of the body. Such microarray datasets are usually made available as raw datasets and are need to be normalized or transformed, after which statistical tests are performed to select differentially expressed

genes. These differential genes are then further annotated in public databases for information such as cellular function, biological process and pathways the gene is involved in.

Apart from gene expression data, study of proteins also gives useful information. Proteomic data provides the protein expression levels in the blood of the human body. Proteins in the proteomic datasets generated by methods such as LC/IMS mass spectrometry are identified by searching several sequence databases which involve search algorithms. The results obtained then need to be validated using biological annotations. To find an emerging significant protein pattern in the patient's blood sample, the data has to be pre-processed for normalization, transformation and then analyzed using heuristic methods. Obviously biological researchers need to have a workflow based system which would save a substantial amount of time for them. Therefore the need arises for not only database integration but also the incorporation of analytical tools. Such data integration systems substantially reduce the time involved to perform *in-silico* experiments.

Several systems have been proposed which provide a framework for carrying out workflow based biological experiments. Many ongoing projects that have been developed deal with *bioinformatics service* integration. They include Bio-MOBY (M Wilkinson, 2005) , MyGRID (MyGRID, 2004) , Taverna (T. Oinn, 2004), Transparent Access to Multiple Bioinformatics Information Sources (TAMBIS) (R Stevens, 2000) and Kepler (B Ludascher, 2004). Bio-Moby system is classified into two divisions i.e. Moby-s and s-Moby. MyGRID is a loosely coupled suite of middleware components to support *in-silico* experiments in biology. TAMBIS also is an application that allows query bioinformatics resources. Kepler is

a particular scientific workflow system that involves analytical steps of database access and querying, data analysis and mining, and many steps including computationally intensive jobs. These projects employ different techniques for achieving biological web-services inter-operability. The main aim is to develop a robust integration system to incorporate *bioinformatics services* for biomarker identification in *in-silico* experiments with microarray data.

SIBIOS is a similar system developed to facilitate biological workflows and also has certain distinctive capabilities. SIBIOS successfully allows integration of bioinformatics tools including data sources and analytical applications and also assists researchers to run *in-silico* experiments according to their workflow specifications. SIBIOS has the key features necessary for the platform i.e. to control the execution and also to automate workflows in it.

All workflow-based systems have the capacity to provide workflows for general experiments. SIBIOS however has the additional key features necessary for the platform to assist in building and controlling workflows. The system supports general purpose services like PDB, SwissProt, NCBI Gene, BLAST, FingerPrintScan and many more. These services meet general requirements of *in-silico* experiments and are not targeted towards the automation of microarray analyses like *biomarker discovery.*

2.1.2. Microarray analysis for Biomarker Discovery

Microarrays have emerged as a great technology for unfolding the mysteries and reasons for abnormal gene expression. It is the most promising clinical application of modern genomics. It opens a prospective for more reliable and efficient diagnosis of tumors, prediction of response to treatments, and risk group determination.

A typical oligonucleotide microarray would look like a 2 x 2 array (figure 4) consisting of spots in various shades of red, green, and yellow. The color of each spot corresponds to the intensity level of a gene expressed in a particular sample. These gene expression values are measured with great precision, and they become the basis of the microarray gene expression analysis. There are various instruments used to measure the expression level of each spot from the image of a microarray chip. These instruments are based on image processing concepts.

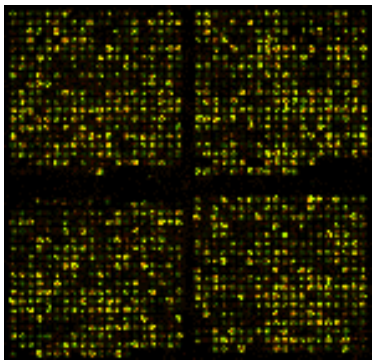


Figure 4: Microarray chip image

An instrument available commercially for gathering microarray data is Agilent 2100 Bioanalyzer (Agilent, 2000) for assessing the RNA integrity. Arrays can be scanned using a

laser confocal scanner and expression values can be calculated using certain software like Affymetrix Microarray software (MAS) (Hardware, 2004). But sometimes even measurement with immense precision can contain certain inherent errors which can give incorrect results (C S Brown, 2001).

While working on microarrays, it is necessary to identify the goal of the study. The hypothesis becomes the basis on which statistical analysis on the experimental research is performed. The analysis of the data depends on the following factors:

1. Type of dataset (cDNA, oligonucleotide microarray etc.)
2. Normalized or raw data
3. Statistics used based on number of independent/dependent variables
4. Purpose of study i.e. what results are need

Microarray analysis is based on the study of gene expression levels. For the purpose of discovery of biomarkers in the blood of diseased samples, gene expression data are utilized and processed. The objective of the experimental design here is to trouble shoot a sample protocol and transform it into a robust process. To develop a biomarker pipeline, a series of steps need to be identified. The main steps include:

1) Data Pre-processing
2) Data Analysis
    a. Analysis-first Annotation-second Model
    b. Annotation-first Analysis-second Model
3) Biological Analysis

2.1.2.1. Data Pre-processing

Pre-processing is the step that enhances the data quality and improves the identification of meaningful characteristics in the dataset. Pre-processing also prepares the data for the application of analysis methods. A popular pre-processing technique is the transformation of the initial gene expression to the logarithm of the raw values. Normalization is another procedure applied to account for the systemic differences in the datasets. It is generally applied to modify the values in order to reduce the noise signals in microarray experiments using different dyes, i.e. Cy3 and Cy5. Data transformations also depend on the type of microarray data under study. Certain logarithmic transformations are equally applicable to cDNA and Affymetrix datasets, while other methods like background correction and probe-level pre-processing are specific to a given technology. Some of the top preprocessing methods that are applicable in the context of microarray experiments are described below (Draghichi, 2003; Speed, 2003).

1) Log Transform

Logarithmic transformations have been used widely for a long time. They provide transformed data that are easily interpreted and are more biologically meaningful. It is effortless to transform gene expression values in order to eliminate the misleading disproportions between two relative changes in two different pairs of values. For example,

Log transformation of base 10 will transform these values into:

$\log_{10}(100) = 2$, $\log_{10}(1000) = 3$, and $\log_{10}(10000) = 4$

This reflects the case that these genes are affected in the same way, only that they are transformed in different proportions. That is, $2 - 3 = -1$ and $4 - 3 = 1$.

These genes are affected by the same magnitude but in different directions. Log transformations also contribute in partially reducing the variance and the mean intensity in the data. Another advantage of the log transformation is that the data becomes symmetrical and almost normally distributed.

Finally, if log of the base 2 is taken then it helps in the data analysis and interpretation. For example, selecting genes with 4-fold variation can be done by cutting a ratio histogram at the value $\log_2$ (ratio) =2.

2) Array Normalization

The main aim of using microarrays extensively is to make random comparisons between gene expression levels in various conditions and tissues possible. For that purpose it is essential for the data to be normalized so that the processed data are independent of the specific experiment and technology used. Currently, no standard way for normalizing the microarray data in a universal manner has been achieved. It is still under question whether the data obtained from different technologies like oligonucleotide and cDNA arrays can be compared directly. Yuen et al (T Yuen, 2002) reported a fairly good correlation between the expression data measured with GeneChip and the cDNA chip, while Kuo et al (W Kuo, 2002) found no correlation. The difficulty may be due to the overall difference in the intensities measured. The goal here is to normalize the arrays in such a way that the values corresponding to the individual genes can be compared directly from one array to another. Some of the methods to achieve that are:

- Divide the intensities by the mean before the log transformation. This method is equivalent to the correction using arithmetic mean.
- Subtract the mean from the intensities after the log transformation. It is similar to the correction using the geometric mean.

3) Normalization issues specific to Affymetrix data

The Affymetrix technology and the data processing techniques are provided by the company itself. As a result most laboratories pre-process Affymetrix data in the same manner.

Some of the ways in which Affymetrix data are preprocessed are:

- Background correction

  Cell intensities are corrected for background using some weighted average of the backgrounds in the neighboring zones. A Perfect Match (PM) probe is a 25-mer oligonucleotide designed to be complementary to a reference sequence. The probe sequence is complementary to the sequence to be hybridized. A Mismatch (MM) probe is a 25-mer oligonucleotide designed to be complementary to a reference sequence except for a single, homomeric (nucleotide mismatch that contains the complementary base to the original) base change at the 13^{th} position (Affymetrix, 2002). An Ideal Mismatch (IM) value is calculated and subtracted from the PM intensity. If the MM is lower than the PM, the IM is taken to be as the difference (PM-MM). The adjusted PM values are then log transformed and the robust mean of the log transformed values are then calculated.

- Signal calculation

  The signal value is calculated as the exponential of this robust mean and scaled using a trimmed mean.

- Detection call

The detection call of a gene signifies if the gene is present in well above the minimum detectable level (P), absent (A), meaning that the gene expression is below the minimum level and Marginal (M) signifying that the gene is present in the near to minimum levels. A discrimination score (R) is calculated for these probe pairs and based on the limit (τ) used, the probe pairs are discarded. By default, $\tau = 0.015$. Increasing τ reduces the number of false positives but also reduces the true detected calls.

2.1.2.2. Data Analysis

In many cases, the aim of microarray experiments is to compare gene expression levels in two different samples. They are generally experimental samples comparing normal data with diseased data. In such experiments biomarkers are the potential genes that are differentially expressed in the two samples compared.

There has been a considerable amount of work done in the field of bioinformatics for significant gene selection. Numerous computational methods have been implemented such as statistical classifiers, significant pattern detection, and bioinformatics service integration systems. In order to formulate a protocol for a microarray experiment, the computational analysis of gene expression as well as biological annotations of the significant gene set needs to be conducted in an organized manner. There are two approaches that can be used in order to carry out the data analysis of the genes. They are namely Analysis First-Annotation Second Model and Annotation First-Analysis Second Model:

a) Analysis First-Annotation Second Model

In order to obtain a list of differentially expressed genes, many computations are required on the gene expression data set. There are many methods that have been developed which can be applied to microarray data. The procedures can be classified into 3 main categories:

- Machine Learning Methods

Some of the types of machine learning methods are decision trees, support vector machines, and principal component analysis.

1) Decision Trees

Decision trees are based on a rule induction algorithm. This method is generally used where systematic selection of a small number of features is used for a decision making process. It increases the comprehensibility of the knowledge patterns. A decision tree is a NP-complete problem. Its construction is based on the fact that the root node of the tree is determined first and then the root nodes of its sub-trees. Every feature can be used to partition the training data. If the partitions contain a pure class of training instances, then this feature is most discriminatory. A decision tree based approach was applied in a research study (D Singh, 2002) to study the expression levels in a prostate cancer dataset. It is applied for the main purpose of classification by grouping genes with similar patterns.

2) Support Vector Machines (SVM)

Support vector machine is a supervised learning algorithm that determines a small number of critical boundary instances from each class and then construct a linear

discriminant function that separates them as widely as possible (H Witten, 1999). It addresses the problem of learning to distinguish between positive and negative members of a given class of n-dimensional vectors. The SVM algorithm operates by mapping a training dataset into a high-dimensional feature space and attempting to locate in that space a plane that separates the positive from the negative as shown in figure 5.

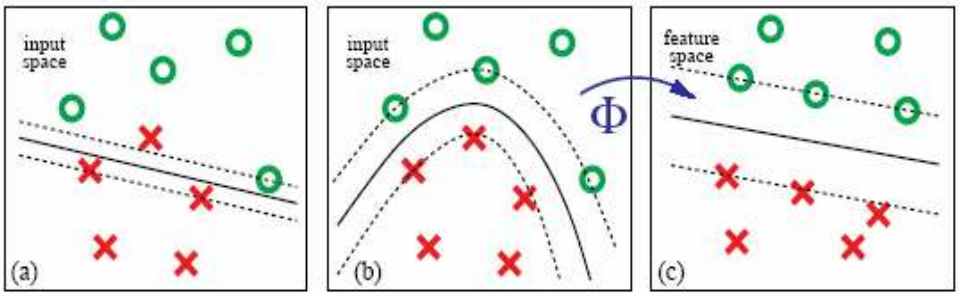


Figure 5: Mapping of input to feature space by "kernal" function (A Zien, 2000)

Using this information, the SVM can then predict the class of an unknown sample by mapping it into the feature space. Support vector machines have been used by Koike *et al* in a study where the prediction of protein-protein interaction sites is done using sequences (A Koike, 2004). In this study the identification of protein-protein interaction sites was important for mutant design and prediction of protein-protein networks. The interaction sites were predicted using SVM and profiles of spatially / sequentially neighboring residues. Another research group used the theory of SVM to classify genes by using gene expression data from microarrays. They used SVMs to predict functional roles for uncharacterized yeast ORFs based on their expression data (M Brown, 2000).

3) Principal Component Analysis (PCA)

A single microarray analysis experiment can generate measurements for thousands of genes. Principal component analysis allows defining a core set of independent features for the experiment state and allows them to be compared directly.

It is a very good statistical technique to identify the significant components in a multi dimensional dataset to identify the differences in the observations. It provides a good visualization outline of the data. When PCA is applied to gene expression data, we can get a summary of the way in which gene responses vary under different conditions. There are many tools like Spotfire (SpotFire, 1996) and Matlab (MATLAB, 1994) that can be used to apply PCA. Figure 6 shows the illustration of a plot of PCA where clusters are made with each describing a particular feature of a data set.

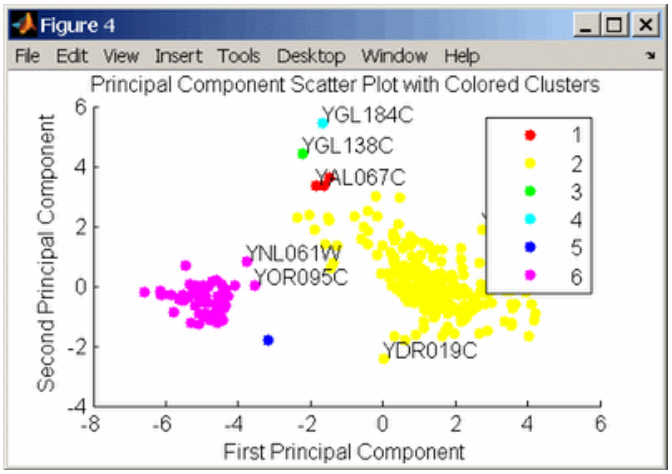


Figure 6: View of dataset using PCA in MATLAB

A research study conducted in 2006 (S Raychaudhari, 2000) is an example of microarray study using PCA.

- Data mining methods

Data mining refers to the process of extracting useful and meaningful information from a large dataset or database. Examples of data mining techniques used include association rules and emerging patterns.

1) Association Rules

Association rules are devised for larger databases with moderate features. But however they have been adopted in the context of microarrays to take into account microarray analysis. Association rules are essentially of the form LHS → Consequent. The LHS is a set of items. It is called the anticipant of the rule and the right hand part consists of the consequent part of the rule. Every rule is associated with a support and confidence. The number of rows in a dataset that match an association rule is called the support of that rule and the probability of that rule being true is called the confidence of the rule. Certain mining algorithms can be applied in the context of microarray analysis in order to obtain association rules. This is a fascinating approach as many biologically interesting rules can be obtained. As compared to random data results, these rules are reliable and are not by chance. In the application of bioinformatics Creighton and Hanash used data mining to find such association rules from expression profile of yeast (C Creighton, 2003). They obtained a number of rules that made sense biologically, and many others suggested new hypotheses that could be investigated further.

2) Emerging Patterns

Emerging patterns are basically related to two classes. The patterns signify the discriminating features between the two datasets. It is a challenge to obtain a manageable set of emerging patterns that are easily understandable to the researcher. Therefore the issue of efficient discovery arises. In order to implement this methodology, there are different algorithms that can be applied. E.g. border based algorithms were applied to discover the most specific and most general emerging patterns from large datasets (J Li, 2001).

b) Annotations First-Analysis Second Model.

This is a novel approach that has been adopted by recent research groups. The idea is to use gene annotations before applying statistical computations in a given microarray dataset. This procedure takes into account the molecular heterogeneity of the different characteristic expression patterns in different patients. In addition to the expression data that is utilized to find significant genes, such approaches utilize the functional annotations from the Gene Ontology (GO, 1999) database. A research group from Germany developed a novel algorithm called Structured Analysis of Microarrays (StAM) (C Lottaz, 2005). They have taken advantage of the functional annotations from the Gene Ontology database to build biologically focused classifiers. They classifiers are then used to discover potential molecular disease sub entities and associate them to biological processes without compromising overall prediction accuracy.

2.1.2.3. Biological Analysis

This is the last step to biologically validate the subset of genes obtained by either of the two approaches. For this purpose, their annotations are studied in detail. Various online biological databases are used such as OMIM (OMIM, 1994), and DAVID (DAVID, 2006). The biological meaning is utilized to relate the gene to the disease under study.

Various groups have been working on biomarkers for various different reasons. Some of them are for disease classification purposes, some for drug discovery, and a few for the identification of genes responsible for the disease in clinical applications. In 2004 a group of researchers from Germany were investigating the possible future application of gene expression profiling for the diagnosis of leukemia (A Kohlmann, 2004). They were working on demonstrating that gene signatures defined for childhood Acute Lymphoblastic Leukemia (ALL) were also capable of stratifying distinct subtypes in a cohort of adult ALL patients. This was a promising finding as global gene signatures were being identified by microarray expression analysis. Microarray technology has been used over the years for various purposes. Another group of collaborators devised an approach to predict new classes of cancer without any prior knowledge of it (T.R. Golub, 1999). The results demonstrated the feasibility of cancer classification by gene expression monitoring and the researchers developed a strategy to discover and predict cancer classes without any previous biological knowledge. Another research group has studied the cell lines of leukemia and identified the chromosomal abnormalities associated with them (B M Fine, 2004). When these cell lines were studied together with the clinical samples, it was found that each chromosomal

abnormality was associated with a characteristic gene expression signature specific to the cell line and the clinical sample.

Even though there are numerous methods to carry out research in biomarker discovery, there are no set standards adopted universally. These different studies on the same microarray dataset lead to different results. In a study conducted in 2001, a technique was utilized by the researchers to compare the results of different statistical approaches (M K Kerr, 2001). This study signifies the absence of set standards because of the lack of agreement with respect to following a universal procedure in the area of biomarker discovery.

2.2. Some results in Biomarker Discoveries

There has been active research for the detection of cancer biomarkers. In the recent years biomarker discovery has been a major focus of cancer research. Many investments have been made for early detection of cancer. Back in 1965, Dr Joseph Gold found out a test for recognizing a common cancer (S K Chatterjee, 2005). A substance from the patient's blood having colon cancer was found and this substance is normally found in fetal tissues. He named it carcinoembryonic antigen (CEA). In the 1980's more biomarkers were found such as CA 19-9 for colorectal and pancreatic cancer, CA 15-3 for breast cancer and CA-125 for ovarian cancer. They have been proven to be early markers for reliable indicators of early disease as they are present in basal levels in normal people and higher concentrations in cancerous individuals. But these markers were not specific for a particular cancer. Lung cancer also often had elevated levels of CEA and CA-125 in women.

Till now, prostate-specific antigen (PSA) has been the best-known cancer biomarker for early detection of prostate cancer. It has been used for screening, diagnostic purposes as well as monitoring disease recurrence. PSA is the only biomarker to be approved by FDA. Some of the biomarkers that have been identified are given in Table 1. (S K Chatterjee, 2005)

Table 1: Cancer biomarkers and their use.

Biomarkers	**Cancer**	**Use**
PSA	Prostate	Screening, diagnostic, predict recurrence
CEA	Colorectal, lung, breast, liver, pancreatic, thyroid, liver	Determine recurrence, monitor treatment, efficacy
CA 125	Ovarian	Diagnostic, monitor treatment, predict recurrence.
BTA	Bladder	Diagnosis, predict recurrence
Calcitonin	Thyroid	Diagnosis, monitor treatment and predict recurrence.
Vimentin	Kidney	Prognosis
Myc and A1B1	Hepatocellular carcinoma	Prognosis
SELDI pattern	Ovarian cancer	Diagnosis, prognosis stage
MMP	Prostate, breast	Prognosis
ICTP	Ovarian	Prognosis, stage.

All these biomarkers have not been approved by FDA. Some reasons for the lack of FDA approval are that most of these biomarkers are of limited clinical use. Many have lacked epidemiological validity or statistical power so they are deficient in universal application. Approval of PSA has motivated researchers to identify suitable markers for other cancers and improve the present predictive capability of cancers.

2.3. Biology of Leukemia

The evolution of a normal cell into cancer involves disruption and deregulation of a number of basic cellular processes. Multi cellular organisms, in their evolution have developed redundant controls through which the homeostasis between different cell types is maintained.

One of the safeguards that prevent excess cell accumulation is a cell-intrinsic program that can induce cell death through apoptosis. The growing understanding that transforming mutations can activate this intrinsic apoptotic response has emphasized the importance of this process in preventing cancer cell development. Apoptosis control mechanisms appear to be impaired in virtually all tumors, suggesting that a required step in carcinogenesis is to disengage the apoptotic machinery and hence it will be beneficial to understand the mechanisms by which normal cells become malignant but also to prevent and treat cancer in humans.

The term leukemia refers to cancer of the white blood cells. The word "leukemia" means "white blood" in Greek. Under normal circumstances, the blood-forming, or hematopoietic, cells of the bone marrow make leukocytes to defend the body against infectious organisms such as viruses and bacteria. But if some leukocytes are damaged and remain in an immature form, they become poor infection fighters that multiply excessively and do not die off as they should (MamasHealth, 2000).

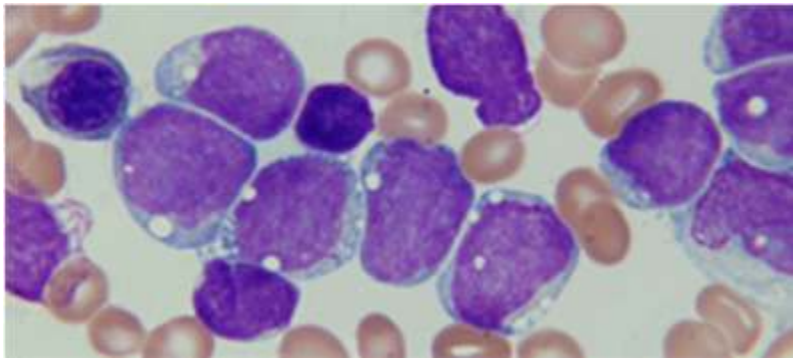


Figure 7: Leukemic cells under a microscope

The leukemic cells accumulate and lessen the production of oxygen-carrying red blood cells (erythrocytes), blood-clotting cells (platelets), and normal leukocytes (figure 7). If untreated, the surplus leukemic cells overwhelm the bone marrow, enter the bloodstream, and eventually invade other parts of the body, such as the lymph nodes, spleen, liver, and central nervous system (brain, spinal cord). In this way, the behavior of leukemia is different than that of other cancers, which usually begin in major organs and ultimately spread to the bone marrow.

As leukemia progresses, the cancer interferes with the body's production of other types of blood cells, including red blood cells and platelets. This results in anemia (low numbers of red cells) and bleeding problems, in addition to the increased risk of infection caused by white cell abnormalities. As a group, leukemia accounts for about 25% of all childhood cancers and affect about 2,200 American young people each year. With the advance in medical science and ongoing research the chances for a cure are very good with leukemia. With treatment, most children with leukemia are free of the disease without it coming back.

Normal human body has controlled levels of white blood cells in the blood. ALL is a quickly progressing disease in which a lot of immature lymphoblasts are found in the bone marrow and blood. Bone marrow normally produces stem cells (which are immature blood cells). These later develop into mature blood cells (figure 8). Mature blood cells are of three types:

a) Red blood corpuscles (RBC)

b) White blood corpuscles (WBC)

c) Platelets

During the progression of leukemia, another kind of white blood cells called as lymphocytes are produced in increased amounts. In ALL lymphocytes do not fight infection well. Lymphocytes increase in the blood and bone marrow thus making less space for healthy WBC, RBC and platelets. These inefficient lymphocytes occur in three kinds in the blood which are:

a) B – Lymphocytes

b) T – Lymphocytes

c) Natural killer cells (they attack viruses)

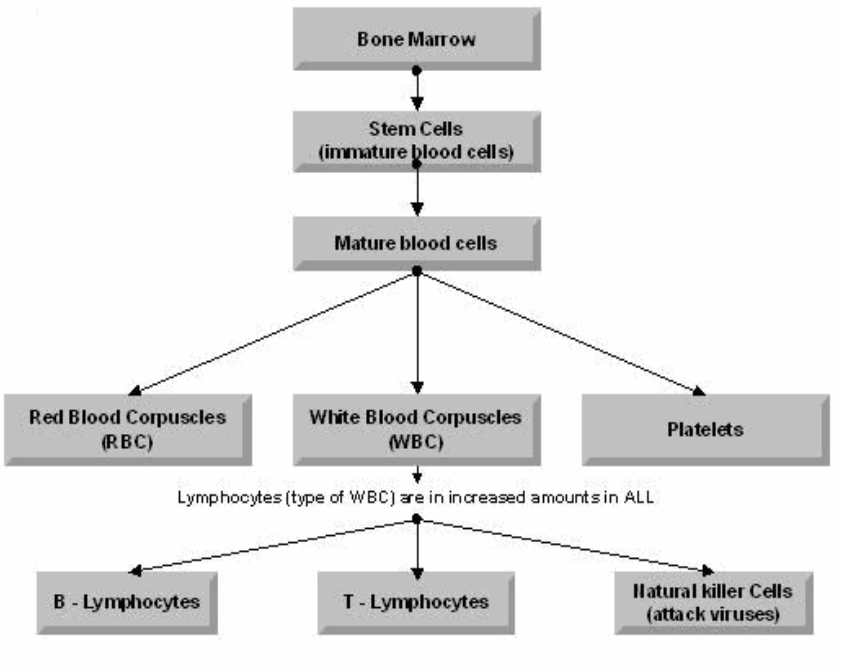


Figure 8: Development stages of Lymphocytes

Pediatric ALL is a heterogeneous disease consisting of various subtypes. One of our objectives is to identify significant relevant genes specific to the various subtypes of ALL. The purpose is to appropriately recognize the subtype as they all vary markedly in their course of treatment. Given the right treatment, recovery rates are good and chances of relapse of the disease are less. The 6 major subtypes of leukemia are:

1. BCR-ABL t(9;22)
2. E2A-PBX1 t(1;19)
3. TEL-AML1 t(12;21)
4. Rearrangements in MLL gene on chromosome 11
5. Hyperdiploid karyotype (>50 chromosomes)
6. T lineage leukemia (T-ALL)

The same protocol developed is applied for each of the six subtypes. We will analyze BCR-ABL in the next section to identify significant genes corresponding to it.

Occurrence of BCR-ABL gene

- BCR (Breakpoint cluster region gene) encodes for BCR protein with locus at 22q11. It functions as a GTPase activating protein.
- ABL (Abelson's murine leukemia viral oncogene) gene has a locus at 9q34 and functions as a non receptor tyrosine kinase.

Leukemic (BCR-ABL subtype) cells have an abnormal feature called the Philadelphia chromosome. The Philadelphia chromosome results from a mutation called a translocation (two chromosomes break, then parts from each chromosome switch places). In BCR-ABL, the translocation occurs between chromosomes 9 and 22 (human DNA is packaged in 23 pairs of chromosomes) and produces a new, abnormal gene called BCR-ABL. This abnormal gene produces BCR-ABL tyrosine kinase, an abnormal protein that causes the excess WBCs. BCR-ABL is associated with a more heterogeneous pattern of gene expression and poor clinical outcome.

The Philadelphia chromosome is an acquired mutation i.e., a person is not born with it and does not pass it on to his/her children. Exactly why the Philadelphia chromosome forms is unknown in most cases, although exposure to ionizing radiations has been shown to be one of the causes for such abnormality.

Based on the biological processes that BCR-ABL is involved in, the results are analyzed. The purpose is to study how the set of genes obtained as significant, are involved in the abnormal process that is created due to the translocation of the BCR and ABL genes.

In the following sections we will identify the main statistical and biological approaches that will be used in the identification of the potential biomarkers. Initially, the concentration has been on the subtype BCR-ABL, for its significant genes. The same protocol can then be applied for the rest of the 5 Leukemia subtypes.

Integral part of this research is to use the existing SIBIOS system for conducting biomarker experiments. The next section describes the SIBIOS workflow integration system.

2.4. Description of SIBIOS

SIBIOS (B. Miled, 2004) is a system for executing biological workflows in the bioinformatics domain. By making rich set of biological databases and annotation tools available, researchers can transparently utilize the output of one bioinformatics service as an input for another bioinformatics service. Several challenges have been encountered during

the design of SIBIOS to enable the seamless integration of bioinformatics services. These challenges are mainly twofold.

- The heterogeneities that characterize bioinformatics services at the syntactic and semantic level.
- The nature of biological workflows that are characterized by:
    - An experimental phase where different services and their composition within a workflow undergo a discovery phase.
    - A production phase where automation of the workflow execution is necessary.

2.4.1. SIBIOS Architecture

SIBIOS is designed as a multi-tier client server architecture. The client module builds the workflows as specified by the user with the help of service composition. This workflow specification is then sent by the client to the workflow enactment module. SIBIOS's knowledge base contains the semantic as well as ground level description of service. The description of each service is in the form of five properties available as a part of SIBIOS's ontology. These properties are used to describe a set of features by which a service can be searched in SIBIOS. These properties for a particular bioinformatics service relates to the input accepted (*has_input*), output obtained (*has_output*), task performed (*perform_task*), algorithm used (*is_function_of*) and the resources accessed by the tool which are mainly databases (*use_resource*). Not all properties are utilized by a bioinformatics service. The low level description of the bioinformatics service is also given in the form of an XML schema also called as *service schema*. It is accessed by the server to ensure service execution during

the enactment. The knowledge base admin/Service Publishing module is used for the maintenance of the knowledge base.

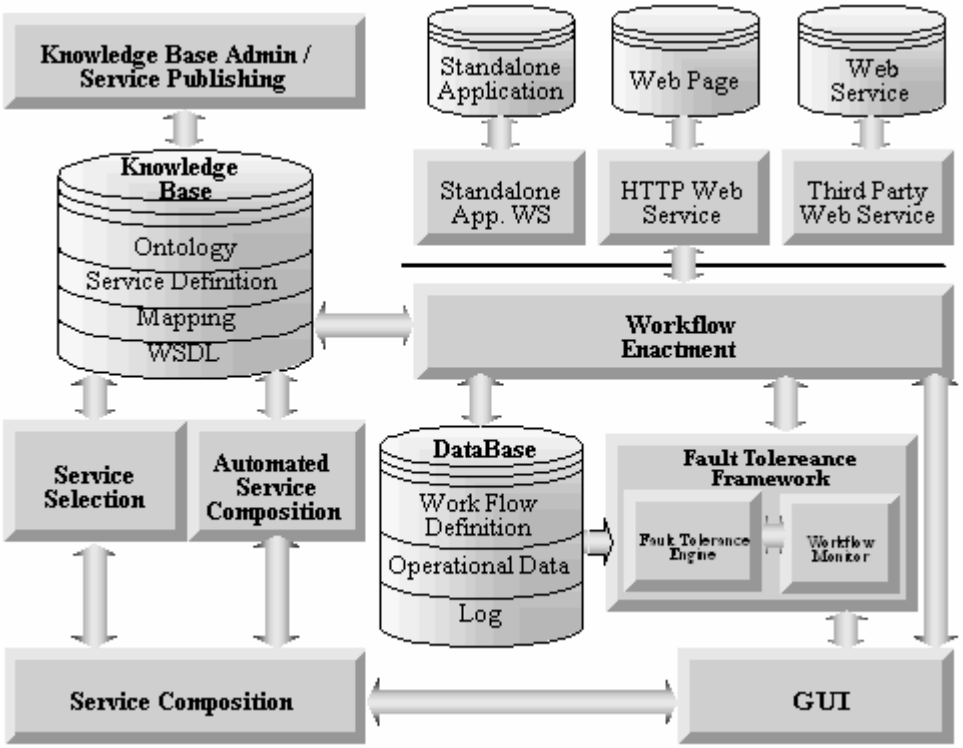


Figure 9: SIBIOS architecture

The overall architecture of the SIBIOS system is shown in figure 9. The user interacts with SIBIOS through the GUI module to construct the workflows according to his/her requirements. Workflow building is aided by the service selection module or the automated service composition module. The automated service composition allows the user to submit a high level description of the workflow, and later generate a list of potential workflows that meet the user's requirements. The user can then select the most appropriate one for execution. This module is still under development. The workflow enactment module interacts with many other modules to achieve the specifications of the scientific workflow. Sometimes

there maybe a failure in any of the components of SIBIOS and may halt the execution of the workflow. The fault tolerance framework addresses these failures in a manner to fulfill the objective of the system even on failure of these components.

2.4.2. Workflow Design

2.4.2.1. Service Selection

The user constructs the workflow he/she wants to execute using the graphical user interface module of SIBIOS. It is done by sequentially adding and linking the services that one desires to execute. SIBIOS presents two approaches to the user for workflow construction. One approach is through the service discovery dialog box (B. Miled, 2004) and the second approach is through the service browsing facility. For the former case, SIBIOS provides a service browsing facility that supports classification of the services by using properties such as *has_input, perform_task*, etc.. Knowing which task is needed to be performed, the user can select the application required to include in the workflow. For the latter case, the user can search for the service desired through the service discovery facility by combining more than one property and then add it to the workflow (L. L. M Mahoui, J Chen, N Gao, 2005). This increases the precision by which the services are discovered as more search elements are considered and this option is generally used by more sophisticated users. A service discovery process can help the user identify the services that are needed in the workflow based on the requirements (e.g. *has_input, perform_task,* etc.) of the service to be executed. It is an intelligent discovery system that utilizes the properties of the services rather than the name of the database or the bioinformatics tool. This allows researchers unfamiliar with large amount

of available bioinformatics services to select the appropriate application for a workflow. The service browsing approach allows the user to progressively drill down through the list of available services using either one of the service properties. Figure 10 and 11 shows how these two service selection facilities are deployed in SIBIOS. To select the next service to be added to the workflow two main cases are considered. Either the user knows the name of the next service to be added. Or the user does not have the information, but he is capable of describing the service he/she wants to add using some features such as service input.

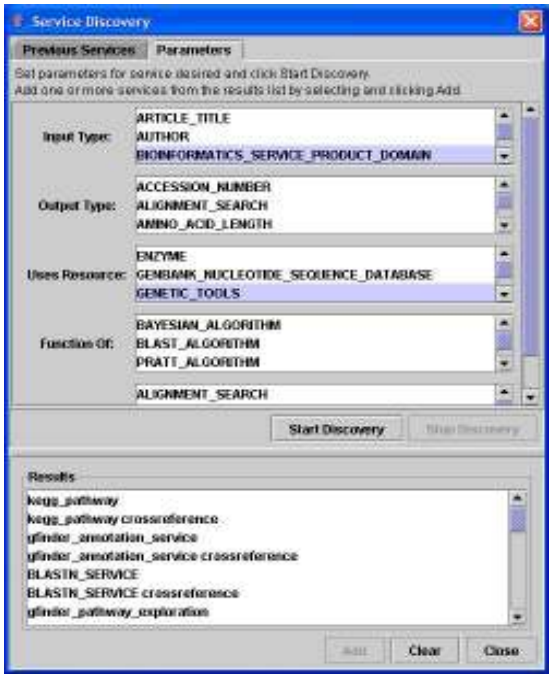


Figure 10: Service Discovery facility on SIBIOS

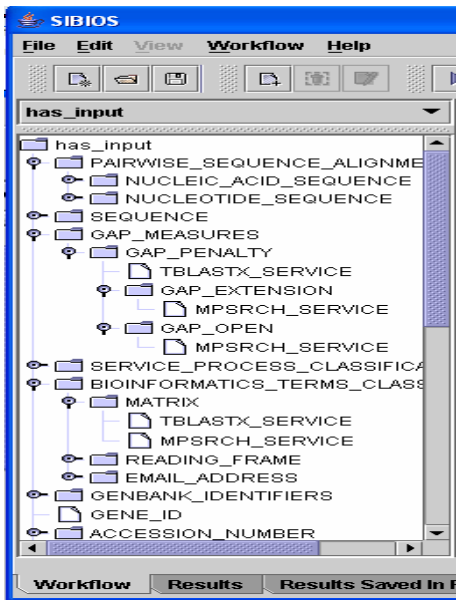


Figure 11: Service Browsing Facility in SIBIOS

As discussed before each module in the workflow corresponds to a web application that provides a particular service. For the execution of the workflow the information needed to contact those services successfully are provided in the schemas that reside in the knowledge base of the system.

2.4.2.2. Service Composition

Service composition is the process of SIBIOS that connects the various services into a meaningful workflow. Two services can be connected in a workflow only if the input parameters of the latter service match the output parameters of the targeted previous service. There are various service connectors (join, union, iteration, pause) available in SIBIOS to join different services (Z. M. M Mahoui, S Srinivasan, M Dippold, B Yang, N Li, 2006). The usage of these operators is illustrated in the sample workflow in figure 2. The red box in the workflow is a filter operator used to pass user-specified information to the subsequent

service. The green box corresponds to the join operator that combines the results from 'SWISSPROT_SEARCH_SERVICE' (SwissProt, 2006) and 'ENZYME_SEARCH_SERVICE' (Enzyme, 2000) before it is sent to 'GENBANK_SEARCH_SERVICE' (Gene, 2004). The yellow box is used to execute repeatedly on a particular service until a certain condition is achieved. And the hollow yellow circle over an arrow connecting two services is used to pause the workflow execution for user intervention in defining service parameters. A pausable service allows the users to take decisions after viewing the results from the previous result in a better way. He/she can define service parameters and select inputs for the next service based on the results he/she views from the previous service.

2.4.3. Workflow Enactment

Workflow enactment module is a complex engine that interacts with various components in the SIBIOS system to achieve the goals of the scientific workflow. The workflow composed by the user is defined in an XML file. This description of the workflow is used by the workflow enactment module. The XML file is then sent to the server of SIBIOS and is validated by a DTD file supplied at the server side. The remote connection with the bioinformatics service is done using the service schemas stored in the knowledge base. This service schema contains the detailed information of the service to be executed. A sample service schema is illustrated in table 2. There are four main parts in the schema that contain detailed information of the bioinformatics service. They are SERVICE_APPLICATION, URL_STRING, EXTRACTION_RULES and INTERFACE_PARAMETERS.

Table 2: Sample XML Schema

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes" ?>
<TOP>
<SERVICE_APPLICATION>
 <SERVICE_NAME>swisspfam_service</SERVICE_NAME>
      <PATH_PAGE>
       <PAGE_CHARACTER>atv</PAGE_CHARACTER>
       <EXTRACT_LINK>sh='{title}pfam' t='{/title}' l='for ' r=''</EXTRACT_LINK>
      </PATH_PAGE>
</SERVICE_APPLICATION>
<URL_STRING>
      <SINGLE_QUERY>
       <BASIC_URL>http://www.sanger.ac.uk/cgi-bin/Pfam/swisspfamget.pl</BASIC_URL>
       <METHOD>P</METHOD>
       <DEFAULT_PARAMETERS />
       <SWISSPROT_ACCESSION_NUMBER>name=XXXX</SWISSPROT_ACCESSION_NUMBER>
       <SWISSPROT_ENTRY_NAME>name=XXXX</SWISSPROT_ENTRY_NAME>
       </SINGLE_QUERY>
</URL_STRING>
<EXTRACTION_RULES>
       <DOMAIN_INFORMATION>
        gh='{tr class=normaltext }' t='--form stuff--}' s='{tr class=normaltext }'
       <DATA_STORE>sh='{td ' t='{/tr}' l='align center' r='{A HREF'</DATA_STORE>
       <DOMAIN_NAME>sh='{td ' t='{/tr}' l='{b}' r='{/b}'</DOMAIN_NAME>
       <DOMAIN_ACCESSION_NO>sh='href=/cgi-bin/pfam' t='}' l='getacc?'
 r='}'</DOMAIN_ACCESSION_NO>
       <SEQUENCE_START>sh='{td ' t='{/tr} l='{td''{td''{td''}' r='{/td}'</SEQUENCE_START>
       <SEQUENCE_END>sh='{td ' t='{/tr} l='{td''{td''{td''{td''}' r='{/td}'</SEQUENCE_END>
       </DOMAIN_INFORMATION>
</EXTRACTION_RULES>
<INTERFACE_PARAMETERS>
       <MAN_OR>
       <PARAMETER>
       <NAME>SWISSPROT_ACCESSION_NUMBER</NAME>
       <DESCRIPTION>Swiss-Prot Accession Number</DESCRIPTION>
       <TYPE>
       <TEXTFIELD />
       </TYPE>
       <DEFAULT_VALUE />
       </PARAMETER>
       <PARAMETER>
       <NAME>SWISSPROT_ENTRY_NAME</NAME>
       <DESCRIPTION>Swiss-Prot Entry Name</DESCRIPTION>
       <TYPE>
       <TEXTFIELD />
       </TYPE>
      <DEFAULT_VALUE />
      </PARAMETER>
      </MAN_OR>
 </INTERFACE_PARAMETERS>
 </TOP>
```

The URL_STRING contains the information required to connect to a service, SERVICE_APPLICATION has information to identify the final result page from which the

results are extracted. The extraction rules consist of instructions read by the SIBIOS server in order to perform result extraction. The INTERFACE_PARAMETERS contain the input field description of the bioinformatics service. They can also contain optional parameters which may be used for additional specification of the input query to the service.

2.5. Current Understanding

Biomarker discovery is now becoming an essential and significant area of research. It has been proving vital in early cancer detection as well as diagnosis of tumors. Biomarkers are essentially a set of few genes (1, 2 or about 10 maximum), which serve as molecular signatures, that are in unwarranted amounts as compared to normal individuals. These genes (biomarkers) change the routine functioning of the human body by affecting the biological pathways. These genes if identified properly can be used in finding the root cause of cancers and therefore to develop drugs to target these genes to control their abnormal activity. Biomarkers are used to measure the progression of a disease, early warning signs of a disease, detection of disease recurrence, etc. It can also be used to determine which tumors respond to which treatments and predict the likelihood of drug resistance.

This research provides an analysis of pediatric Acute Lymphoblastic Leukemia (ALL) whose gene expression data is available from the St Jude's Website (M E Ross, 2003) and the procedure involved in discovering the biomarkers responsible for this disease. The protocol followed for this process is aided by SIBIOS and thus enhances the functionality of the system.

2.6. Problem statement and motivation

The main problem today lies in the fact that there have been no set procedures to identify biomarkers for a given disease. This makes the process of adapting the SIBIOS workflow system a very difficult task especially when it comes to demonstrating its potential role in discovering biomarkers.

Cancerous biomarkers can be effectively used for accurate evaluation and diagnosis of the disease in different stages as they serve as useful clinical indicators. In future they will guide physicians in every step of disease management. Thus it has been useful to work in this area and detect more and more biomarkers for different diseases.

It is a challenge to develop sample protocols for the biomarker detection. Such a pipeline if proposed can be then utilized to automate the process of biomarker discovery through workflow based systems.

So far, SIBIOS had been useful for carrying out various biological operations for different purposes such as gathering protein information, sequence matching, enzyme search etc. But recently with the onset of the genome project, it becomes important to integrate workflows for biomarker discoveries using microarray data. For this purpose it was needed to add new functionality to SIBIOS in order to support construction of complex workflows for cancer research. Such experiments involve analytical analyses and gene annotations.

2.7. Proposed solution

There has been an extensive study on the existing microarray analysis approaches. All data analyses are quite powerful and robust in themselves. They highly depend on the data under study and its characteristics. The goal is to synthesize these techniques and propose sample protocols for biomarker discovery. These workflows need to be then tested and validated by analyzing the results. For this purpose a dataset is selected and used to validate the proposed protocol. The results are corroborated by performing biological annotations. Finally it is proposed to integrate the sample protocols which are converted into corresponding workflow which will make it feasible to incorporate it into SIBIOS.

While each biomarker discovery analysis is context specific (e.g. for leukemia the gene expression level was studied for each subtype and the top distinguishing genes were found for each type (M E Ross, 2003)), the methodology that is proposed can be applied for other types of diseases in general.

We propose to perform two experiments on the same dataset and study the results obtained by both approaches. One process involves the use of statistics and biological annotations as steps of the protocol developed for this biomarker discovery process. The other approach is a computational procedure which comprises of a series of recursive calculations to mine a set of interesting association rules specific to the type of cancer under study. It is the implementation of an algorithm called FARMER (G Cong, 2004). FARMER is basically a data mining procedure which generates interesting rules that are significant to a class of

samples. The results from both these approaches are then compared and analyzed for their biological relevance and significance.

Every experiment is context specific so methodologies have been used corresponding to the leukemia cancer i.e. its form of occurrence, cells that it affects, clusters in which the subtypes occur, etc.

The main goal of the experiment is to propose sample protocols and then use SIBIOS as a tool to support the construction of this workflow. To make SIBIOS more robust, bioinformatics services that perform analyses and gene annotations on input gene lists were added. These bioinformatics applications involved certain considerations that were needed to be facilitated by the system. In addition, SIBIOS capabilities need to be improved to take into account the characteristics of bioinformatics services used for biomarker discovery.

The goal is not to propose one particular protocol for biomarker discovery but to characterize the main steps involved in this process, their objectives, and their importance and also involve the corresponding tools and databases.

The outcome of this protocol is used

- As a result by itself in proposing a list of potential biomarkers, supplied by good evidences from the available annotations of those genes.
- As baseline developing templates of microarray workflows that will be used by researchers to build their own version of the workflow tailored to their data and disease in focus.

3. METHODS

3.1. Gene expression and microarrays

Gene expression can be used to understand the phenomenon related to the disease studied and microarrays allow the interrogation of thousands of genes at the same time. Two methodologies for biomarker discovery were adopted. These approaches are explained in detail in the next section. The initial dataset that was used to work on was a leukemia dataset (M E Ross, 2003).

3.2. Data Selection

Here the objective of the research is to identify biomarkers specific to the subtype of the tumor, Acute Lymphoblastic Leukemia (ALL) in children. The dataset selected was a set of 132 samples of children suffering from Acute Lymphoblastic Leukemia (ALL). This data set was selected based on the advantages that:

- The microarray data was normalized.
- The samples had signal intensities, detection call as well as p-values provided.
- The data set was an Affymetrix data set on the chip HgU133A, that had more probe sets on it (around 22,000) compared to the previous Affymetrix chip, HgU95Av2 having fewer probe sets in it (figure 12).

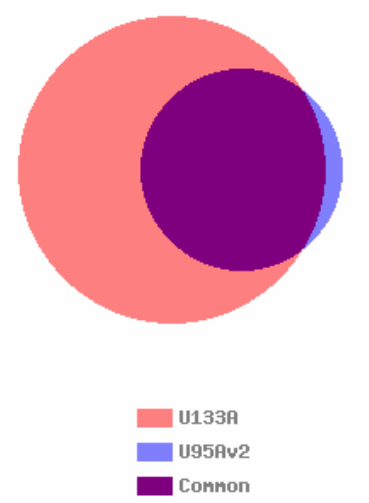


Figure 12: Comparison of the gene chips HgU133A and HgU95AV2

This dataset has been very popular among various researchers also. HgU133A chip is used, because it can identify diagnostic discriminating genes from a larger portion of the human genome. It provides an almost threefold increase in the number of genes as compared to the previous chip, HgU95Av2. The data provides a large amount of information for the support of microarray-based expression profiling on a single platform for the detection of biomarkers of the known subtypes of pediatric ALL.

3.3. Microarray Data Analysis

Significant gene selection involves the process of mining relevant and meaningful genes from the initial dataset. Data mining can be defined as the extraction of previously unknown and potentially useful information from data. It is the science of extracting useful information from large datasets. The microarray dataset under study is on a HG-U133A chip. It contains expression values of specific mRNA levels. The data has been normalized and is for 132

patient samples from each of the leukemia subtypes. We devised two different approaches that give a good overview of the results. The purpose of analyzing the same dataset with two different approaches is to study how different methods affect the results obtained. The next step was to decide meaningful statistical computations to be applied on the dataset. The types of statistical tests applied, plays a major role in the results obtained. Studies applied on the same data having different results often used statistical techniques that are different in their approach. It is not always easy to choose statistical procedures to be applied. Each method has its own peculiar sensitivities and blind spots.

There are two approaches that have been used here to accomplish the task of biomarker discovery analysis. The first approach is based on using analytical processes first and then applying annotations on the subset of gene lists. Two different methodologies have been applied here. One is based heavily on statistics for filtering data until a subset of genes are left for manual biological analysis. The second methodology is on the basis of a data mining algorithm that has been developed and implemented for mining significant genes from the microarray dataset. The aim of the algorithm was to generate interesting rule groups which help in identifying test samples as diseased or normal. These are association rules which are generated essentially from a set of genes that occur in a particular pattern specific to a group i.e. class or non-class. This algorithm is designed to mine the occurrence of specific genes whose occurrence indicates the presence or absence of the disease. Both these approaches are discussed in detail in section 3.3.1 and section 3.3.2.

3.3.1. Analysis First-Annotation Second model

This section describes in detail the two approaches that have been adopted to implement the process of significant gene selection. The purpose is to find an efficient way to the analysis in order to propose a small set of genes from the initial 22,000 probe sets. Section 3.3.1.1 describes the statistical analysis procedure and section 3.3.1.2 describes the data mining algorithm. This method can also be called as a top-down approach. The basic flow diagram involved in this approach is shown in figure 13.

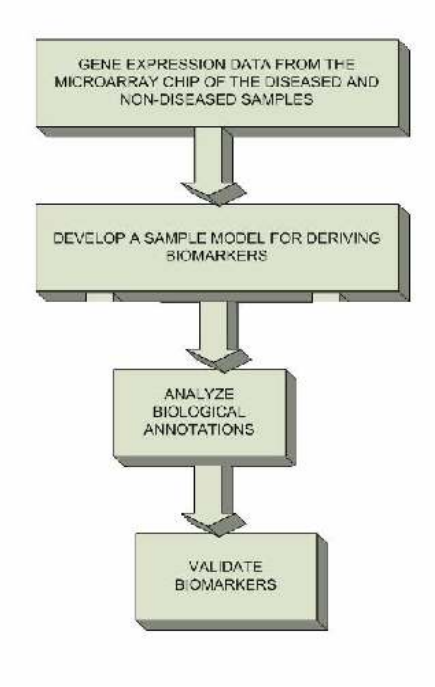


Figure 13: Basic workflow of Analysis First Annotation Second Model

This approach mainly uses gene analysis procedures first before using annotations on the subset of genes. The two methods implemented here were the statistical analysis procedure and the FARMER data mining algorithm.

3.3.1.1. Statistical Analysis Procedure

Initially we have the gene data on the HG-U133A chips with its values of expression levels, detection call, and p-value. The goal is to identify biomarkers specific to each subtype of pediatric ALL. Initially the experiment was performed on one subtype i.e. BCR-ABL and later duplicated to each of the other subtypes to find significant genes specific to them. The protocol followed is the same for each subtype.

From the analysis performed on several microarray studies, the following parameters have been identified to be used as a part of the process of filtering the most significant genes of interest. They are:

- Detection Call
- P-Value
- Fold Change

The procedures involved in obtaining these features and how they have been used in the filtering process have been described in detail below.

Detection Call

The detection call is the qualitative measurement indicating if a given transcript is detected (Present), not detected (Absent) or marginally detected (Marginal) in the sample. So our first step is to exclude the set of genes that are absent across all the patients. Extensive experiments were performed using Microsoft Excel. The other statistical analysis tools that were also incorporated for the microarray study were SpotFire (SpotFire, 1996), PAM (PAM,

2005), SAM (SAM, 2005), and Bioconductor (Bioconductor, 2001). There were a number of samples for each of the subtypes for ALL. The initial step was to label each gene of every sample within a subtype a representative detection call. This call was computed based on the occurrence of the call values across the samples of a particular probe set. The probe sets were given a call of 'A' if the number of presents across the samples were less than half the number of samples comprising the leukemia subgroup under analysis. (e.g. if in 12 samples a probe set X had less than 6 samples having a detection call of 'P', then probe set X was labeled 'A'). If the samples had more than half the samples as 'P' then it was labeled a 'P' as the representative call for that probe set.

The initial dataset had Affymetrix control sets and its signals. They were all removed prior to the analysis. For the study we had a group of diseased samples (E.g. BCR-ABL) called as class and non-diseased samples (E.g. non-BCR-ABL) called as non-class. Based on the detection calls for the class and the non-class probe sets we can see which of them are biologically interesting and which are not. A summary of the results is given in table 3.

Table 3: Detection calls and interestingness measure

Detection Calls		**Magnitude of class/non-class ratio**	**Biological meaning**	**Interesting gene**
Class	**Non-class**			
M/A	M/A	Meaningless, can be anything	The gene is probably not expressed in either condition	No
M/A	P	Meaningless, close to zero	Gene expressed on the baseline array, not expressed in the experiment	Yes
P	M/A	Meaningless, very large	Gene expressed on the experiment, not expressed in the baseline array	Yes
P	P	Meaningful, can be anything	Gene expressed on both channels, useful ratio	Depends on ratio

Fold Change

A fold change is basically a measurement of the amount by which a probe set has been up-regulated or down-regulated in a class compared to the probe set in the non-class. To compute the fold change, initially all expression values are converted to $\log_2$ based values (L_{nm}), as a log transformation generates an appropriate data range.

G_{nm} = expression level of gene n in sample m

n = (1 to 22,215) Number of probe ID

m = (1 to 132) Number of samples

$$L_{nm} = \log_2 (G_{nm})$$

Based on these log transformed values, the fold change for every gene is calculated. While carrying out biomarker discovery, the two data sets that are studied are from normal samples (control) and from diseased (test) patients. The data sets are in pre processed states i.e. they have been scaled and normalized.

For the fold change calculation, the average of the log transformed data from the control group is subtracted from the log transformed values of the test data. This difference is powered to 2, giving the fold change value result. Fold change signifies the amount of up or down regulation of a gene. E.g. fold change of 2 means an up-regulation of 2 fold and 0.5 means a down-regulation of 2 fold.

L_{class} = Set of geometric average of log values of samples (class) across all probe IDs

$L_{non\text{-}class}$ = Set of geometric average of log values of samples (non-class) across all probe IDs

$$D_n = L_{class} - L_{non\text{-}class}$$

$$FC_n = 2 \wedge D_n$$

FC_n = Fold change of gene n in class compared to non-class

P-Value

P-value is the probability that a certain statistic is equal or more extreme to the observed value when the null hypothesis is true. The null hypothesis is that the two classes are the same. The p-values for the genes are calculated to quantify the significance of the differentially expressed genes. P-value helps in detecting significant genes present in the list, which are used further in the study for biological interpretation. A T-Test is carried out for this calculation. In order to apply a T-test it is necessary that the data has the samples as normally distributed as it is a parametric test. In order to see the distribution of the samples across each subtype, we plot their distributions to see the distribution of the microarray intensities. The $\log_2$ transformed expression values of the microarray data are plotted. It is commonly observed in biology that noise increases with the level. This problem is resolved by using transformations on the data. The most common transform is the logarithmic transform and it also has the attractive feature that fold changes of any given size appear as shifts of constant amounts for all genes. Figure 14 and 15 are the distributions of the subtypes BCR-ABL as well as T-ALL. All of the other subtypes showed a similar kind of distribution. The distributions of the intensities are roughly bell-shaped which indicate a nearly normal distribution of the samples. Therefore we can apply the T-test statistic to our data.

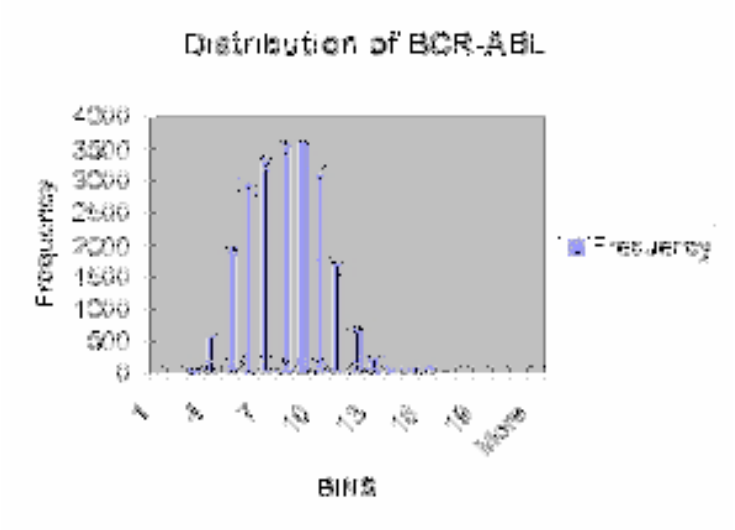


Figure 14: Roughly bell shaped distribution of BCR-ABL

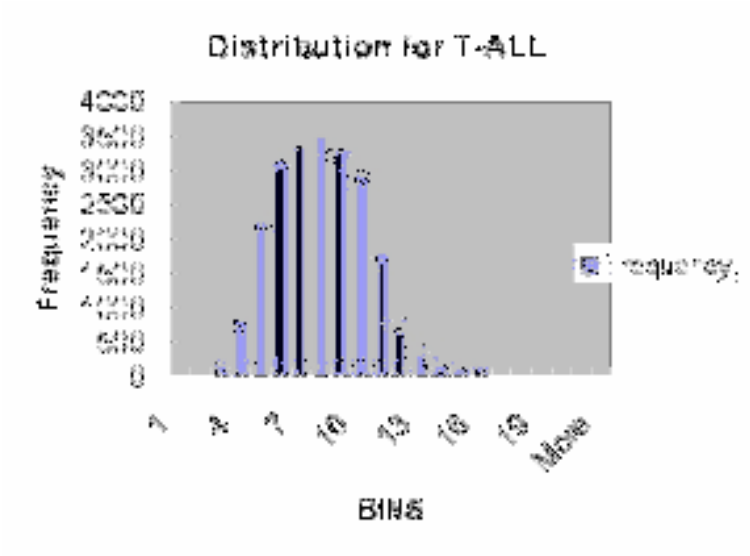
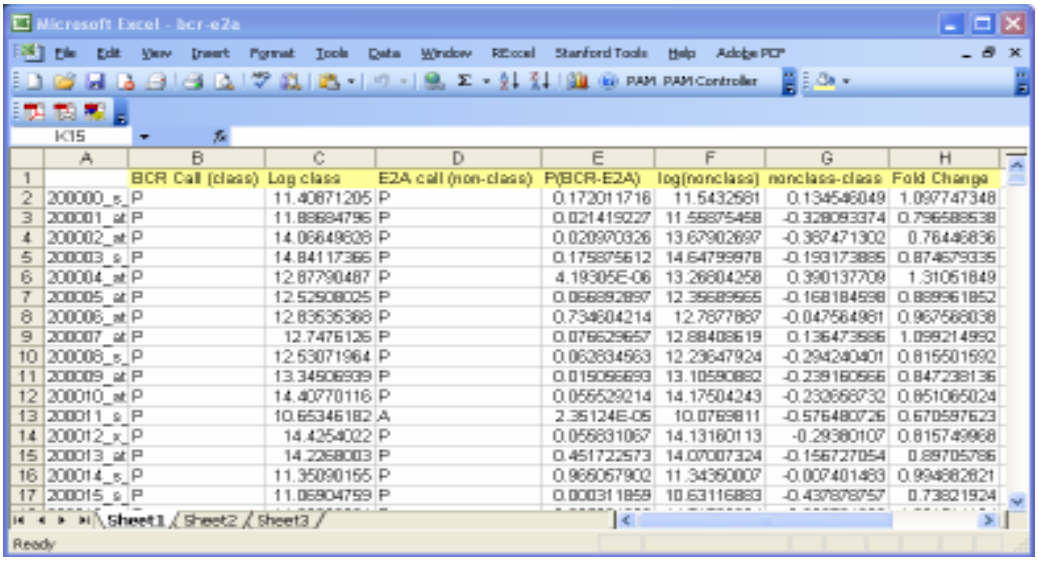


Figure 15: Nearly normal distribution of T-ALL subtype of Leukemia

We want to consider up-regulated as well as down-regulated genes, so a 2-tailed T-test is performed. But as T-tests are used, the problem of false positives also arises. False positive is a Type I error and indicates that the gene has changed, but it actually has not. There needs to be a way to limit the number of false positives. If false positives are not tolerated then an approach involving the *Bonferroni* correction is employed. It means that if 100 T-tests are

performed then each p-value is multiplied by 100 to obtain the correct adjusted value. If the false positives can be tolerated then the use of False Discovery Rate (FDR) is incorporated. E.g. if a false positive rate of 15% can be tolerated, then the FDR is set to 0.15 and the p-value is then calculated. FDR method is less conservative than Bonferroni and is usually more appropriate. This FDR method was implemented using a publicly available software and the p-value obtained was then used to set the threshold for filtering genes.

Based on the calculations performed for detection call, P-value and fold change the probes are filtered based on certain specifications. Initially an excel file is created which contains the data for all probe IDs. There are columns containing log values of class group (BCR-ABL samples), log values of non-class group (non-BCR-ABL samples), detection call for class, detection call for non-class, difference of log values of class and non-class, fold change, and p-values. For instance, while working on significant genes specific to BCR-ABL class, and considering E2A-PBX1 as non-class, the corresponding excel file is shown in figure 16.



	A	B	C	D	E	F	G	H
1		BCR Call (class)	Log class	E2A call (non-class)	P(BCR-E2A)	log(nonclass)	nonclass-class	Fold Change
2	200000_s_	P	11.40871205	P	0.172011716	11.5432681	0.134546049	1.097747348
3	200001_at	P	11.88884796	P	0.021419227	11.55875458	-0.328093374	0.796588538
4	200002_at	P	14.06649628	P	0.020970326	13.67902697	-0.387471302	0.76446836
5	200003_s_	P	14.84117366	P	0.175875612	14.64799978	-0.193173885	0.874679335
6	200004_at	P	12.87790487	P	4.19305E-06	13.26804258	0.390137709	1.31051849
7	200005_at	P	12.52908025	P	0.066892897	12.39689565	-0.168184598	0.889961852
8	200006_at	P	12.83535368	P	0.734604214	12.7877887	-0.047564981	0.967568038
9	200007_at	P	12.7476126	P	0.078629657	12.88408619	0.136473586	1.099214992
10	200008_s_	P	12.53071964	P	0.062634563	12.23647924	-0.294240401	0.815501692
11	200009_at	P	13.34506939	P	0.015096693	13.10590882	-0.239180566	0.847238136
12	200010_at	P	14.40770116	P	0.056529214	14.17504243	-0.232668732	0.851065024
13	200011_s_	P	10.65346182	A	2.35124E-05	10.0769811	-0.576480726	0.670597623
14	200012_x_	P	14.4254022	P	0.056831067	14.13160113	-0.29380107	0.815749968
15	200013_at	P	14.2268003	P	0.451722573	14.07007324	-0.1567270554	0.89705786
16	200014_s_	P	11.35090155	P	0.966057902	11.34360007	-0.007401483	0.994882821
17	200015_s_	P	11.06904759	P	0.00031859	10.63116883	-0.437878757	0.73821924

Figure 16: Sample excel file for filtering

After construction of this dataset, the next step is to filter data based on certain thresholds. The data was filtered using the following criteria:

a) Detection call P for class

b) Detection call P for non-class

c) Fold Change > 2.0

d) P-value < 0.05

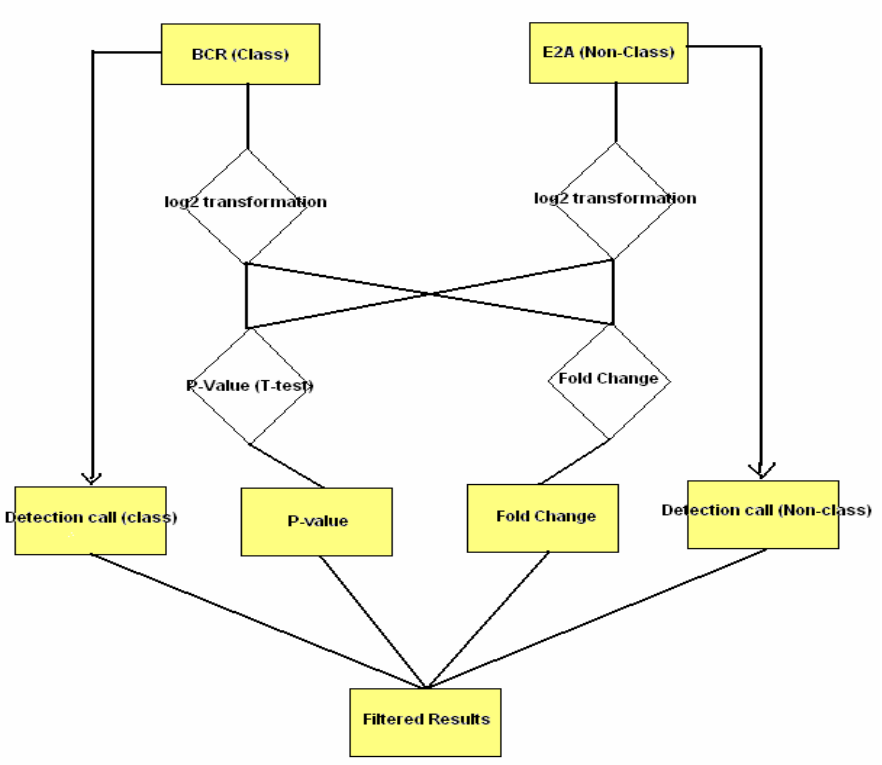


Figure 17: Protocol followed for filtering significant genes

After filtering process the data is substantially reduced and we are left with the set of genes that are significantly relevant towards the BCR-ABL subtype of Acute Lymphoblastic Leukemia. The entire protocol can be summarized as shown in figure 17. The same pipelined protocol was followed for each of the subtypes in order to obtain a set of genes specific to the

leukemia subtypes. The biological analyses were then done on this subset of genes using the SIBIOS system which will be explained later in chapter 4.

After following this sample protocol a list of 108 genes were obtained which were used for biological analysis. After annotation those only 53 genes were uniquely identified and the rest were unclassified.

3.3.1.2. Data mining using FARMER Algorithm

This is the other approach that has been carried out for the purpose of obtaining a list of significant genes specific to a leukemic subtype. It is a data mining approach using FARMER algorithm. FARMER (G Cong, 2004) is an algorithm for Finding Interesting Rule Groups in Microarray Datasets. Interesting rules are the set of association rules generated from the same set of rows. Association rules are basically associations between different items in a dataset, corresponding to a particular class. In the context of microarray data, association rules can reveal biologically relevant associations between different genes. The items in the gene expression data can include genes that are highly expressed or represented, as well as relevant facts describing the biological process of the gene.

Some of the advantageous features of the FARMER algorithm are:

- It is adapted to microarray data where there are small number of rows and a large number of columns.
- The generation of interesting groups is essentially a set of rules that are generated from the same set of rows.

- The process identifies a group of genes instead of individual genes that seem to have more biological support, in the sense that a group of genes are usually more responsible for an abnormality in the body.

The algorithm is designed based on the condition that the dataset is oriented such that the rows are the samples and the columns contain the expression values of every gene ID. The advantage of working on such a dataset is that, interesting rules are generated from the same set of rows. These association rules are searched for row wise and as the size of the row enumeration space is much smaller than the size of the column enumeration space, the search time will be considerably lowered.

Before applying the algorithm, a preprocessing phase is required to be carried out on the microarray data. This involves a process called as Entropy Discretization method (Entropy-based-Discretization, 2002). Basically data mining algorithms work well on categorical data rather than on ordinal data. This preprocessing is done as a step to serve the purpose of data conversion. In addition FARMER, uses pruning strategies used to prevent unnecessary traversals of the enumeration tree that store the data. There are three strategies used here, which are described in detail in the later section.

Adapting FARMER for biomarker discovery

Recent studies have shown that association rules are very useful in detecting gene patterns of significant interest in microarray datasets (C Creighton, 2003). They can be potentially very useful in biological research. They are simple to interpret by biologists and can be easily

applied in scenarios like sample classification and significant gene pattern identification. These set of genes can be identified as potential biomarkers for a disease. Another advantage is that the process identifies a group of genes instead of individual genes which seems to have more biological support, in the sense that a group of genes might be biologically related to the disease and are collectively involved in the genetic pathway.

The algorithm implemented here generates association rules of the form LHS → C, where LHS is the set of items and C is the class label. For our dataset we have used two class labels namely BCR-ABL class and non-BCR-ABL class. In order to obtain results, our methodology is based on certain thresholds of support and confidence.

- 'Support of X' refers to the number of rows containing X in the dataset and is referred to as sup(X).
- The probability of an association rule being true is called as the 'confidence of the rule'. It is calculated as sup (LHS U C) / sup (LHS). So, 'the support of the rule' is the number of rows in the dataset that match the rule defined.

In the domain of bioinformatics, there have been various algorithms that have been developed for mining association rules from microarray data. But they are generally computed over a column wise search space. This makes running such algorithms quite computationally expensive. If n is the maximum length of a row in the dataset, then the search space based on column enumeration could be as large as 2^n. Column enumeration methods perform better when $n<100$, but when they are of the order of thousands, it makes such computations almost impractical. On the other hand, the number of rows in such datasets is generally of the order of hundreds to thousands. So if m is the number of rows,

then the size of the row enumeration search space is 2^m. In the domain of microarray data sets, the size of row enumeration space is much less than the size of the column enumeration space. As a result it is more effective to devise an algorithm based on row enumeration.

While mining for association rules, user-specified constraints such as minimum support (a statement of generality) and minimum confidence (a statement of predictive ability) are often imposed.

In order to adapt FARMER algorithm to biomarker discovery, an implementation of the algorithm was necessary. Details of the implementation will be discussed later in this section. Before this step a data pre-processing step was performed on the initial dataset to be ready for the FARMER algorithm.

Data Pre-processing

Recall that the St Jude's research data set (M E Ross, 2003) contains gene expression values of all 132 patients. We use the same dataset that we manipulated with the statistical based approach described in section 3.3.1.1. Our aim is to obtain interesting rule groups for each of the subtypes. The data pre-processing process involves three steps:

1. Class Labeling
2. Feature Reduction
3. Feature Discretization

These 3 steps are discussed in detail.

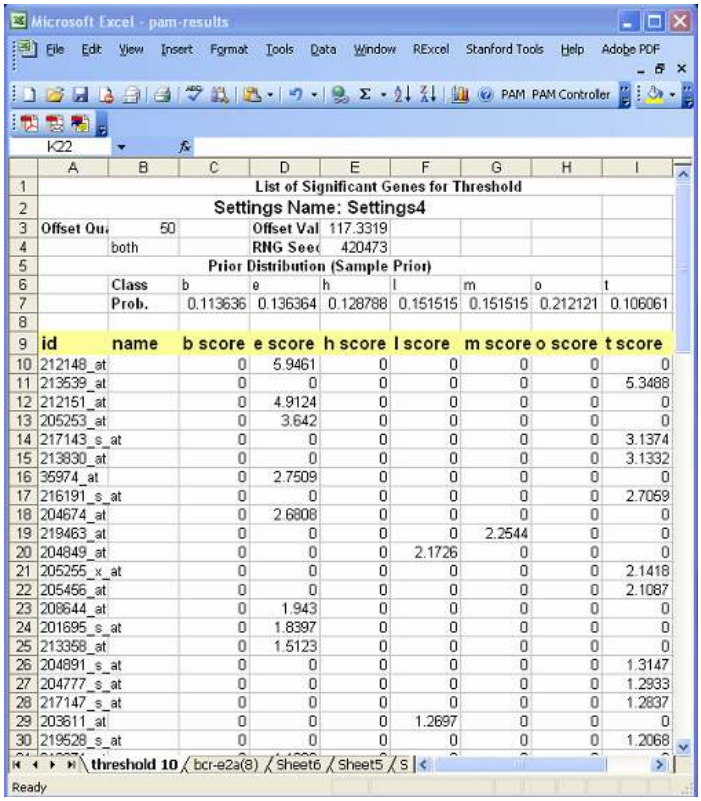
a) Class Labeling

For the purpose of distinguishing each of the classes from each other, we label the samples under study accordingly. A sample is labeled by 1 to represent it as a class (e.g. BCR-ABL) and 0 to represent rest of the data as non class (e.g. non BCR-ABL) in the microarray data.

b) Feature Reduction

In order to run the algorithm on the microarray data, we first obtain a subset of genes from the initial list of 22,000 genes which are comparatively more significant than the rest of the genes. For this process of gene selection we use a software called PAM (Prediction Analysis for Microarrays). It is developed by a research group at the Stanford university (PAM, 2005).

The set of significant genes obtained is based on the threshold that the user specifies. Typically one can try a number of different choices. In selecting the right choice, PAM does a K-fold cross validation for a range of threshold values. PAM was rum on a data set containing expression values of all 132 samples. Each of the six subtypes was labeled separately. As a result, on significant gene selection, a total of 88 genes were obtained with their corresponding scores listed with them at a threshold level of 10. These 88 genes were the genes that were significant across the entire 6 subtypes in the data set. The result of PAM is shown in figure 18.



Microsoft Excel - pam-results

File Edit View Insert Format Tools Data Window REXcel Stanford Tools Help Adobe PDF

PAM PAM Controller

K22

	A	B	C	D	E	F	G	H	I
1	List of Significant Genes for Threshold								
2	Settings Name: Settings4								
3	Offset Qu:	50		Offset Val	117.3319				
4		both		RNG Seed	420473				
5	Prior Distribution (Sample Prior)								
6		Class	b	e	h	l	m	o	t
7		Prob.	0.113636	0.136364	0.128788	0.151515	0.151515	0.212121	0.106061
8									
9	id	name	b score	e score	h score	l score	m score	o score	t score
10	212148_at		0	5.9461	0	0	0	0	0
11	213539_at		0	0	0	0	0	0	5.3488
12	212151_at		0	4.9124	0	0	0	0	0
13	205253_at		0	3.642	0	0	0	0	0
14	217143_s_at		0	0	0	0	0	0	3.1374
15	213830_at		0	0	0	0	0	0	3.1332
16	35974_at		0	2.7509	0	0	0	0	0
17	216191_s_at		0	0	0	0	0	0	2.7059
18	204674_at		0	2.6808	0	0	0	0	0
19	219463_at		0	0	0	0	2.2544	0	0
20	204849_at		0	0	0	2.1726	0	0	0
21	205255_x_at		0	0	0	0	0	0	2.1418
22	205456_at		0	0	0	0	0	0	2.1087
23	208644_at		0	1.943	0	0	0	0	0
24	201895_s_at		0	1.8397	0	0	0	0	0
25	213358_at		0	1.5123	0	0	0	0	0
26	204891_s_at		0	0	0	0	0	0	1.3147
27	204777_s_at		0	0	0	0	0	0	1.2933
28	217147_s_at		0	0	0	0	0	0	1.2837
29	203611_at		0	0	0	1.2697	0	0	0
30	219528_s_at		0	0	0	0	0	0	1.2068

threshold 10 | bcr-e2a(8) | Sheet6 | Sheet5 | S

Ready

Figure 18: Result from PAM containing score of each of the six subtypes for a set of 88 significant genes.

The next step was to obtain the expression values for each of the 88 genes across all of the 132 samples. For every class an average expression value was calculated. This average taken was the geometric mean, in order to avoid the deviation of the actual mean across a set of samples due to the presence of any outliers.

c) Feature Discretization

A method was devised to categorize every expression value in every sample within a subtype based on its comparison with the average mean expression value. To carry out this process a method called as the entropy discretization method was incorporated. This method involves

the conversion of numeric value to a multi-interval discretization i.e. the conversion of numeric attributes to nominal (categorical) attributes.

The process of gene expression transformation is based on the nearest shrunken centroid method. This method computes the standardized centroid for each class (R Tibshirani, 2002). This is the average of gene expressions for every gene in each class divided by the within-class standard deviation for that gene. The nearest centroid classification takes the gene expression profile of a new sample and compares it to each of these class centroids. The class having the centroid closest to, in the squared distance, is the predicted class for that new sample. This shrinkage method is a good approach used for feature selection and has two advantages:

1. By reducing the effect of noisy genes it can make the classifier more accurate.
2. It does an automatic process of gene selection.

Of the many ways for a fine-discretization approach, one of them is based on the heuristic of entropy minimization. Entropy based discretization involved the class entropy calculation. Here the geometric mean calculated with each subtype group serves as a cut point to discretize the data. The expression values of each sample are then compared to the geometric mean and based on the comparison an odd or even value is assigned. If the sample expression is lesser than the mean then an odd value is given else even. As a result we obtain a dataset containing values of attributes which are treated as categorical instead of numeric and are then fed into the mining algorithm to be processed.

Datasets for each of the 6 subtypes were prepared and then each of them were processed by the mining machine for computations.

Mining Algorithm for association rule computation

The data mining algorithm when applied on the dataset containing genes which were significant in comparison to the rest of the other genes, gives an output of association rules, based on a specified support and confidence.

FARMER algorithm performs search by enumeration of row sets to find interesting rule groups with a specified consequent C. The entire data can be visualized as a enumeration tree of the row set from which interesting rules groups are searched for. The generation of this search tree is based on the successive comparisons of every row to the rest of the other rows present below it in the dataset. The result of each comparison is a set of items that are common to both the rows. This result becomes a child in the tree, of the first row node. The visual depiction of the FARMER search tree is given in figure 19.

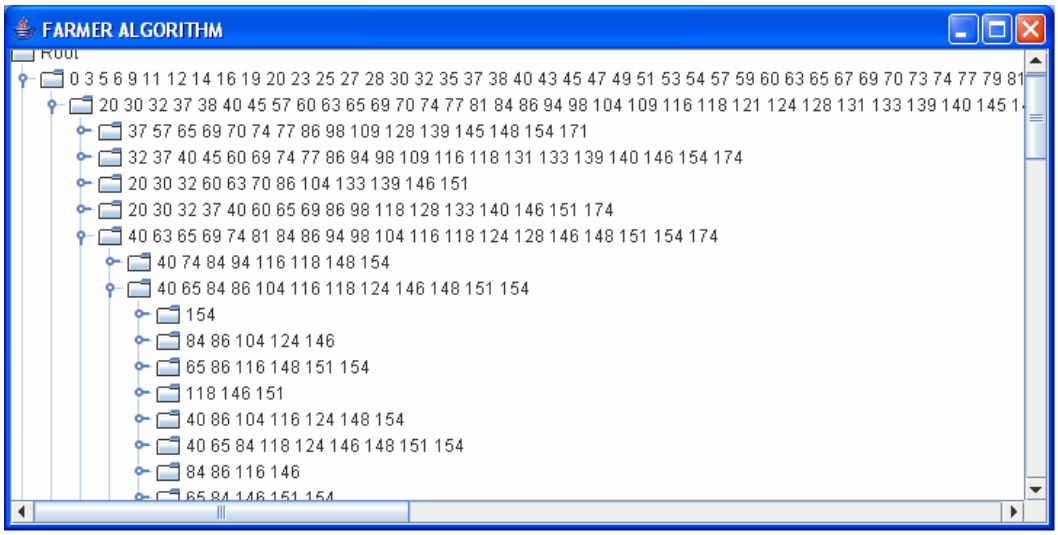


Figure 19: Generation of a row enumeration search tree from FARMER

After the construction of this tree structure, FARMER does a depth-first search on this enumeration tree by moving along its edges. The complete traversal of the row enumeration tree for mining interesting rules is not at all efficient. This is due to the fact that there are a lot of redundancies present in it, and we can avoid them by implementing some pruning strategies. Such pruning techniques snip off many unnecessary searches from the system making it more efficient and computationally economical.

The formal algorithm involves many recursive computations of various nodes by carrying out a depth-first traversal of the row enumeration tree. There are three pruning strategies implemented in this algorithm, which is essential for the efficiency of this system. The emphasis is on the fact that the pruning steps do not clip off any interesting rules and rather prevent unnecessary traversals of the enumeration tree.

Pruning Strategy 1

This strategy is based on the fact that children nodes can deleted from the tree based on certain conditions. It can be visualized as if the child node is compressed to its parent node. For instance if a node has the item set (0, 3, 4) and its parent also has the same item set of (0, 3, 4), then it is obvious that both the nodes will generate the same set of interesting rules. This redundancy can be achieved by pruning off the child node, so the search in the tree below that node is completely discarded. This saves search time considerably in the mining process.

Pruning Strategy 2

This pruning strategy is implemented when an interesting rule has already been identified, and is discovered again later in the enumeration tree. The searching can be pruned below the node which is discovered later because it implies that any rules discovered at the descendants of that node, would have already been previously discovered.

Pruning Strategy 3

In this implementation search space is considerably reduced using the user specified thresholds of support and confidence. We calculate the upper bounds of these measures at every node for a sub tree rooted at a node X. If the estimated upper bound at X is below the user specified threshold, then we stop searching down the node X. A support threshold of for example 6 where the input dataset has 20 samples having leukemia would mean that the rule discovered should match at least 6 samples out of the given 20. These rule groups indicate a significant pattern involving interesting features in the leukemia dataset. Increasing the support makes the system generate better results but also puts more constraints on the system as more number of rows needs to match with the rule discovered. A confidence threshold of for example 97 % indicates the percentage of predictive ability of the system in mining such association rules.

General steps of the FARMER algorithm are described below

Algorithm

1. Input (Microarray data comprising of expression values and class labels, support, confidence)

2. Generate a row wise enumeration tree that is a representation of the row sets to find interesting rule groups with a consequent, C.
3. Traverse the tree for interesting rule groups with pruning strategies to prune of unnecessary searches:
    i. Pruning Strategy 1
    ii. Pruning Strategy 2
    iii. Pruning Strategy 3
4. Result is a set of interesting rule groups.

Interesting Rule Groups

There were a number of interesting rules discovered from this system which indicated interesting patterns in the data set. For each data set of all the 6 subtypes, their respective interesting rule groups were obtained at various levels of support and confidence. The 6 data sets that were used had each two categories. The summary information is given in the table 3.

Table 4: Datasets

Dataset	Class 1	Class 0	# rows
BCR-ABL	BCR	Non-BCR	20
E2A-PBX1	E2A	Non-E2A	20
HD50	HD50	Non-HD50	18
MLL	MLL	Non-MLL	22
TEL	TEL	Non-TEL	21
T-ALL	T-ALL	Non-T-ALL	18

The results obtained consist of a set of significant genes that make interesting rule groups. These set of genes specific to each subtype were analyzed in detail for their biological meaning. This was done by accessing various biological databases for their annotations,

pathways, disease involved, etc. To investigate the set of significant genes, SIBIOS system was adapted to aid in this research. Chapter 4 discusses more about the results obtained and how SIBIOS was used for the analysis of results.

3.3.2 Annotation First-Analysis Second Model

This is a novel approach recently proposed for the discovery of biomarkers. It is suggested that if the biology of the genes are taken into consideration during the filtering of the genes, then one can achieve better results from the analysis. Annotation term is the biological description of the gene function and can serve as useful information in filtering genes that are specific to a disease under study. The purpose of the model is to characterize the diseased patients based on the gene annotations. The method proposed by a research group in Germany (C Lottaz, 2005) falls in this category. They propose a new algorithm called StAM which takes into account the molecular information of the genes during the statistical analysis.

Generally patterns discovered by the available methods of biomarker discovery are not focused from the biological perspective. Therefore the need arises to analyze genes using their underlying molecular disease mechanisms. Our purpose is to use association rules to discover annotation patterns specific to a disease compared to the normal tissue. These annotation patterns allow discovering potential biomarkers for the disease under study. Use of annotations help to filter genes based on their functional relevance to the gene. It is likely therefore to obtain more relevant genes that can serve as biomarkers using this approach.

Our method proposes to use GO (gene annotations) (GO, 1999) with FARMER algorithm to find interesting annotation-based rule groups. FARMER algorithm provides a set of association rules of the form LHS → C, where LHS in the new approach corresponds to a set of annotations. The set of annotations can be mapped to the corresponding genes and those genes can be marked as potential biomarkers.

To apply the Annotation First-Analysis Second Model for extracting biomarkers we used the St Jude's Dataset (M E Ross, 2003) which is same as the data set used for Analysis First-Annotation Second Model. Recall that the dataset contained 132 samples and more than 22,000 genes in it. Before running the FARMER algorithm a set of pre-processing steps need to be applied to the microarray dataset. The initial microarray dataset consists of a set of samples of the diseased and non-diseased patients and their corresponding gene expression values.

3.3.2.1. Data Pre-processing

Before the FARMER algorithm could be applied to the data set a pre-processing phase needs to be performed. The steps can be classified as 5 stages:

1. Class Labeling
2. Feature Selection
3. Feature Reduction
4. Feature Transformation
5. Feature Discretization

a) Class Labeling

Sample datasets were classified as either part of BCR-ABL class (subtype of ALL marked as 1) or of non-BCR-ABL class (marked as 0).

b) Feature Selection

As the purpose of this study is to identify potential biomarkers using annotations, for BCR-ABL the aim of this step was to convert the dataset from gene identifiers to a list of annotation terms. Genes can be described with different biological annotations including gene ontology terms, protein-protein interactions and pathway information. In this study we consider only GO annotations. Different alternatives for annotations exist, varying from general to specific descriptions of the genes. Annotations provided by GO are hierarchically structured into 5 levels, varying from the most general (level 1) to the most specific (level 5). To map a given gene to its annotation we choose to consider the most specific annotation as well as their ancestors in the GO hierarchy. The relationships that exist between the genes and GO annotations is a many to many relationship (figure 20). In addition, many genes do not have any annotations; so they were unclassified.

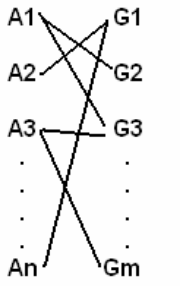


Figure 20: Many to many relationships between annotations (A) and genes (G)

At the end of feature selection phase a set of unique annotations are determined for each gene G. This set is denoted as $\mathcal{A}(G)$ hereafter.

c) Feature Reduction

The purpose of this step is to filter out unimportant annotations. We propose to determine the significance of the annotations of all the unique annotations based on the enrichment score of each annotation. The enrichment score is the number of genes that correspond to a particular annotation term. The annotations are then ranked according to the enrichment score in a descending order. A threshold for the enrichment score is set and then the top n annotations are retained as significant annotations. This set of significant annotations is used in the subsequent data pre-processing steps.

d) Feature Transformation

1. The objective of this step is to determine the gene expressions of each annotation for each data sample. For this purpose we use the gene expression associated to each gene annotation according to equation 1.

$$E(\mathcal{A}_i)_j = \text{geometric mean } (E(G_k)_j) \quad (1)$$

$i \in (1 \text{ to } n)$ (where n is the number of annotations from the feature reduction step)

$k \in (1 \text{ to } m)$ (where m is the initial number of 22,000 genes)

$\mathcal{A}i \in \mathcal{A}(G_k)$

j is the sample number

$E(\mathcal{A}_i)_j$ is called as expression of $E(\mathcal{A}_i)$ and $E(G_k)_j$ is the gene expression of gene k. The equation 1 calculates the geometric mean of the gene expressions over each sample S_j for each annotation $\mathcal{A}_i$.

For example if $\mathcal{A}_i$ maps to three genes G_1, G_2, and G_3 (figure 21), then the value of $E(\mathcal{A}_i)_j$ is the geometric mean of the gene expression levels G_1, G_2, and G_3 for the sample S_j.

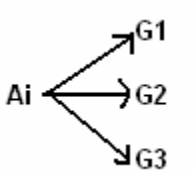


Figure 21: Annotation mapping $\mathcal{A}$i to 3 genes

Table 4 illustrates the sample dataset from the feature transformation phase.

Table 5: Sample data set from Feature Transformation

Samples ↓	Annotations →	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$	. .	. .	. .	. .	$\mathcal{A}_i$
S_1		$E(\mathcal{A}_1)_1$	$E(\mathcal{A}_2)_1$	$E(\mathcal{A}_3)_1$	$E(\mathcal{A}_4)_1$	. .	. .	. .	. .	..
S_2		$E(\mathcal{A}_1)_2$	$E(\mathcal{A}_2)_2$	$E(\mathcal{A}_3)_2$	$E(\mathcal{A}_4)_2$	. .	. .	. .	. .	..
..		…	..	..	..	. .	. .	. .	. .	..
S_j		$E(\mathcal{A}_1)_j$	$E(\mathcal{A}_2)_3$	$E(\mathcal{A}_3)_j$	$E(\mathcal{A}_4)_j$	. .	. .	. .	. .	$E(\mathcal{A}_i)_j$

e) Feature Discretization

The purpose of this step is to categorize annotation expressions into two possible categories.

- One category specifies that the annotation is differentially expressed
- The second category specifies that the annotation is normally expressed

Recall that one of the categories is represented within FARMER by an even number and the other category is represented as an odd number. In order to perform categorization, we need to compare the expression of $\mathcal{A}$i (i.e. $E(\mathcal{A}_i)_j$) to the geometric mean expression of $\mathcal{A}_i$ denoted as $AE(\mathcal{A}_i)$ over all samples. The geometric mean expression of the annotations over all the class and non-class samples ($AE(\mathcal{A}_i)$) is calculated using equation 2.

$$AE(\mathcal{A}_i) = \text{geometric mean } (E(G_k)) \quad (2)$$

$i \in (1 \text{ to } n)$ where n is the number of genes from the feature reduction phase

$k \in (1 \text{ to } m)$ where m is the initial set of 22,000 genes

In case $E(\mathcal{A}_i)_j > AE(\mathcal{A}_i)$ then a odd number is assigned to represent the pxpression of $\mathcal{A}_i$ for sample j otherwise an even number is selected. Table 5 illustrates the sample dataset obtained from the feature discretization pre-processing step.

Table 6: Final Annotation data set

Samples ↓ Annotations→	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$	..	..	..	..	$\mathcal{A}_i$
S_1	0	2	4	6	..	..	..	..	..
S_2	1	2	5	7	..	..	..	..	..
..	…	..	..	..	..	..	..	..	..
S_j	1	3	4	7	..	..	..	..	..

This discretization method is applied in order to distinguish the differentially expressed genes from the normally expressed genes.

Mining association rules using Annotation data set

After this data preprocessing step we obtain a dataset of significant annotations and their categorical values for each sample. FARMER algorithm is then run on the resulting dataset. The algorithm generates a set of annotation patterns more precisely for rules presented as LHS → C. Annotations are the LHS part of the rule. The annotation patterns are mapped back to their corresponding genes. This subset of genes are then said to be the potential biomarkers for the disease that need to be validated for conclusive results.

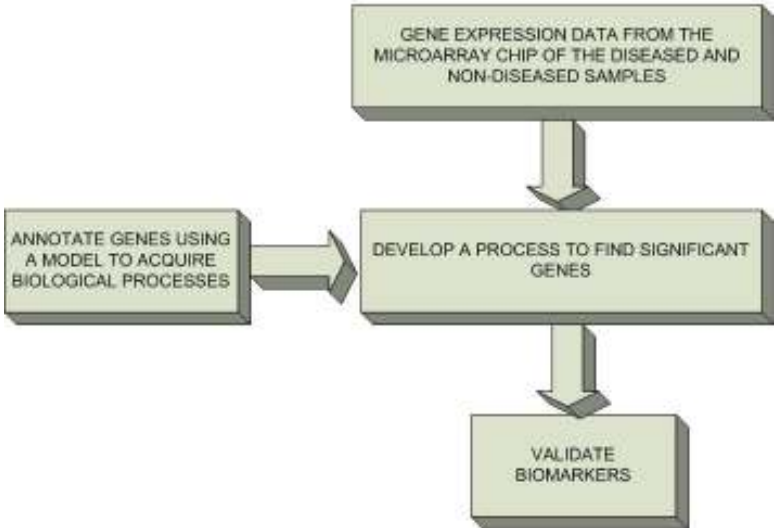


Figure 22: Workflow for Annotation First Annotation Second Model

Figure 22 depicts the high level workflow of the main steps involved in this approach. The implementation of the proposed approach is yet to be carried out. These results obtained will be compared with the results obtained from the previous model i.e. Analysis First-Annotation Second Model. The objective is to find a list of potential biomarkers discovered from more than one approach.

The next step is to then study the issues related to adapting these models into the SIBIOS system.

3.4. Adapting SIBIOS for biomarker discovery workflows

The proposed Biomarker Discovery protocols involve bioinformatics services that perform data analysis as data annotations. The scope of this work is limited to the integration of annotation services.

In order to incorporate the addition of annotation services and database tools that assist in biomarker discoveries, several changes were necessary to be performed on SIBIOS system. Some of the modifications have been made in the structure of the existing XML schemas for SIBIOS in order to accommodate the new services that have been incorporated into the system.

3.4.1. Proposed Enhancements

In this section we discuss the problems faced to adapt SIBIOS for biomarker discovery and the solutions proposed for them.

1. SIBIOS did not include any annotation services. The database tools such as OMIM database are needed for the first model (Analysis First-Annotation Second Model) and annotation tools such as DAVID (DAVID, 2006) are required for the second model

(Annotation First-Analysis Second Model). The newly added biological tools to SIBIOS are:

a. GO (Gene Ontology) (GO, 1999) for obtaining gene annotations such as biological processes, cellular functions and molecular functions.
b. KEGG pathways (Kyoto Encyclopedia of Genes and Genomes) (KEGG, 1995) for obtaining regulatory pathways that the genes are involved in.
c. DIP (Database of Interacting Proteins) (DIP, 1999) for retrieving experimentally determined interactions between proteins.

The bioinformatics tools that were added to the SIBIOS were:

a. GFinder (D Martucci, 2005) for finding gene annotations
b. GOToolBox (GOToolBox, 2004) for computing p-values and enrichment levels of the genes id. It uses certain statistics for these computations.
c. FuncAssociate (FuncAssociate, 2003) for finding annotations of the gene Ids.

2. For the addition of these annotation tools, an issue related to session management was addressed. Integration of such services was possible if the session ID was maintained throughout the querying of the annotation tool using SIBIOS. To tackle this issue a modification was made in the structure of SIBIOS's service schema. An additional tag called <ENTRY_URL> was added in the <URL_STRING> element. This field contained the URL of the initial homepage of the annotation service where the session value is assigned to an application. The purpose of adding this field is to allow SIBIOS to capture the session cookie for the service. The modification of the wrapper that interprets the service schema information was also performed. An example XML schema for a service called GFinder is given in the appendix.

3. The other issue related to annotation tools is the querying of large gene lists for their annotations. The gene lists can contain thousands of genes that need to be queried. Due to this the data needed to be queried becomes very large and cannot be sent through the conventional POST/GET http methods. The query needs to be written directly to the output stream as a sequence of bytes. For this purpose we added a tag <WRITE> to the XML schema where the query (e.g. gene lists) can be specified in this tag. An example of this tag is given in the appendix for the service GFinder.

In the Results section the workflows designed in the SIBIOS system for gene explorations in the context of biomarker discovery are described.

4. RESULTS

In this section here we shall discuss the genes identified using our first approach for biomarker discovery. Their detailed analysis is done in section 4.1. In section 4.2 we shall discuss our workflow implementations in SIBIOS. In 4.3 the bioinformatics services added to the system are discussed.

4.1. Results

After the identification of the candidate genes by feature selection, it needs to be determined whether any of them are causal genes and if there are any surrogates. For this experiment we performed biological annotations and studied the gene regulatory pathways for each of the genes identified as significant. Genes are connected in a circuit or network. Expression of a gene in a network depends on the expression of some other genes in the network. It is then analyzed to determine if the genes affect other genes positively, negatively or in more complicated ways.

A cancer cell can never express all the genes in a rule group in the same way with the same expression intensities. For this reason we observe each gene filtered out by the mining algorithm and analyze them in extreme detail. From the previous approach we got a list of 53 unique genes and using the FARMER algorithm for BCR a set of 8 genes were identified to be significant. There were 3 genes were common in both the methods but we analyzed all 8 in detail for the determination of biomarkers for ALL.

A general understanding in the process of cancer is that cell becomes incapable of normal differentiation. The search for the causation of leukemia has followed several approaches: infectious, genetic, physical and chemical. Over time it has become evident that all approaches may be correct and that several factors contribute for causing leukemia.

A limited search was conducted to pick keywords related to leukemia such as apoptosis, chromosome abnormality, gene translocation, and poor DNA repair and tyrosine kinase inhibitor like terms were identified.

This study focuses on BCR-ABL leukemia subtype. Studies show that BCR-ABL is associated with a more heterogeneous pattern of gene expression than the others. Also BCR-ABL leukemias show relatively few consistent features in their expression patterns. This confirms the heterogeneous nature of this subtype. In BCR-ABL, enzymes that should regulate the growth and development of white blood cells go awry, resulting in uncontrolled growth of the cells.

A chromosomal translocation creates the BCR-ABL genes and leads to expression of chimeric BCR-ABL protein with enhanced tyrosine kinase activity. This disrupts the basic cellular functions such as the control of proliferation, adherence to stroma and extracellular matrix and apoptosis.

Mining for significant genes for BCR-ABL subtype gave 8 genes which can be potential biomarkers for this type of leukemia. The identified genes are:

1. Major histocompatibility complex, class II, DR alpha
2. A kinase (PRKA) anchor protein 2
3. Lymphoid-restricted membrane protein
4. Protein tyrosine phosphatase type IV/A, member 2
5. CD47 antigen (Rh-related antigen, integrin-associated signal transducer)
6. Melanoma antigen, family D, 1
7. Pre-B-cell leukemia transcription factor 1
8. T cell receptor delta locus

Description for each one of the listed above follows:

Major histocompatibility complex, class II, DR alpha :

The major histocompatibility complex (MHC) is a gene cluster of various loci grouped together at a single location. The MHC of humans, or the HLA complex, is located on the right arm of chromosome six approximately 32 cM from the centromere. The HLA complex in its entirety is more than 3800 kb. The loci of the HLA complex may be divided into three classes: Class I, Class II, and Class III. The products of Class I and Class II genes play an important role in the communication between cells. These MHC molecules are surface proteins which present antigenic peptides to T lymphocytes. The Class I and Class II genes exhibit a high degree of polymorphism. The genes are both multiallelic and multigenic.

Class II genes are often referred to as immune response genes or IR genes (7).

The surface proteins produced by Class II genes are only present on specific types of cells known as antigen presenting cells (APC's). APC's include B lymphocytes, macrophages, and dendritic cells. These genes are shown in figure 23 and figure 24.

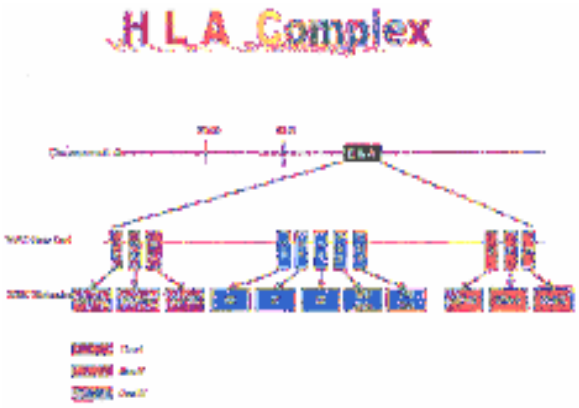


Figure 23: HLA Complex

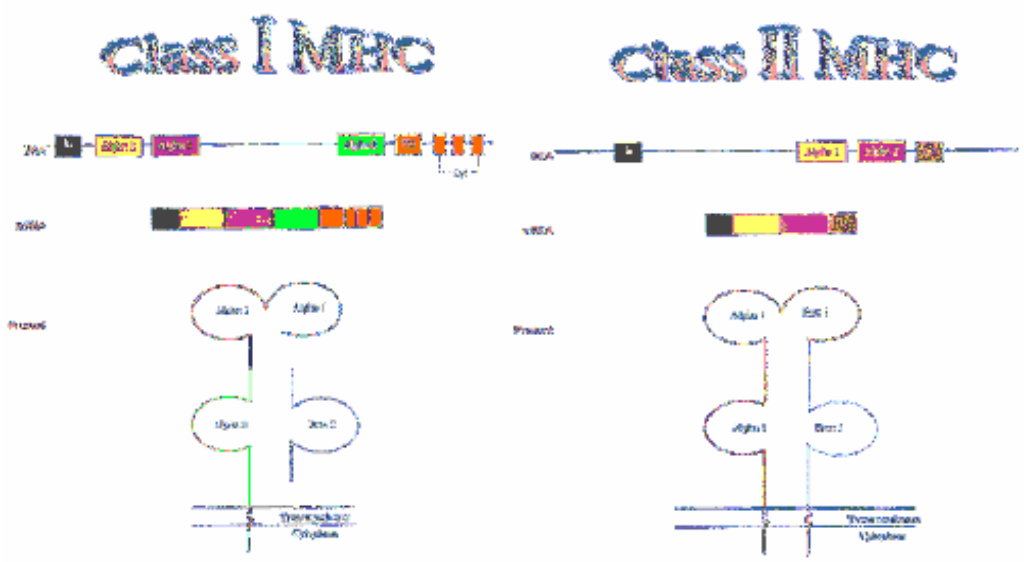


Figure 24: Class I MHC genes, Class II MHC Genes

MHC is said to have been involved in antigen dependent B cell activation and is a B lymphocyte cell surface molecule. Understanding these functions MHC can be further investigated to prove its association with leukemia.

A kinase (PRKA) anchor protein 2

The A-kinase anchor proteins (AKAPs) are a group of structurally diverse proteins, which have the common function of binding to the regulatory subunit of protein kinase A (PKA) and confining the holoenzyme to discrete locations within the cell. This gene encodes a member of the AKAP family but its specific function has not been determined. This gene shares the same 5' exons with a separated gene PALM2 which encodes a protein implicated in plasma membrane dynamics. Alternative splicing of this gene results in several transcript variants encoding different isoforms, but the full-length nature of most variants has not been defined.

The BCR-ABL gene formed by the mutation of the chromosomes 9 and 22 encodes tyrosine kinase which drives abnormal uncontrolled multiplication of leukemic cells.

Lymphoid-restricted membrane protein

The protein encoded by this gene is expressed in a developmentally regulated manner in lymphoid cell lines and tissues. The protein is localized to the cytoplasmic face of the endoplasmic reticulum. OMIM (OMIM, 1994) records have evidence for being one of the genes for cancer.

Protein tyrosine phosphatase type IV/A, member 2

The protein encoded by this gene belongs to a small class of prenylated protein tyrosine phosphatases (PTPs), which contains a PTP domain and a characteristic C-terminal prenylation motif. PTPs are cell signaling molecules that play regulatory roles in a variety of cellular processes. This tyrosine phosphatase is a nuclear protein, but may primarily associate

with plasma membrane. The surface membrane association of this protein depends on its C-terminal prenylation. Overexpression of this gene in mammalian cells conferred a transformed phenotype, which implicated its role in the tumorigenesis. Studies in rat suggested that this gene may be an immediate-early gene in mitogen-stimulated cells. This process is involved in the process of leukemia as it is linked to the over regulation of tyrosine phosphatase.

CD47 antigen (Rh-related antigen, integrin-associated signal transducer)

This gene encodes a membrane protein, which is involved in the increase in intracellular calcium concentration that occurs upon cell adhesion to extracellular matrix. The encoded protein is also a receptor for the C-terminal cell binding domain of thrombospondin, and it may play a role in membrane transport and signal transduction. This gene has broad tissue distribution, and is reduced in expression on Rh erythrocytes. Four alternatively spliced transcript variants encoding distinct isoforms have been found for this gene.

Melanoma antigen, family D, 1

This gene is a member of the melanoma antigen gene (MAGE) family. Most of the genes of this family encode tumor specific antigens that are not expressed in normal adult tissues except testis. Although the protein encoded by this gene shares strong homology with members of the MAGE family, it is expressed in almost all normal adult tissues. This gene has been demonstrated to be involved in the p75 neurotrophin receptor mediated programmed cell death pathway. Three transcript variants encoding two different isoforms have been found for this gene.

Pre-B-cell leukemia transcription factor 1

Human pre-B-cell acute lymphoblastic leukemias are frequently associated with a t(1;19)(q23;p13.3) chromosomal rearrangement. Kamps *et al.* (1990) and Nourse *et al*. (1990) demonstrated that several cell lines carrying this translocation synthesize chimeric mRNAs with 5' sequences encoded by the E2A transcription factor gene (147141) on chromosome 19 and 3' sequences encoded by a homeobox-related sequence, called Prl or PBX1, on chromosome 1. In the chimeric transcription factor, the DNA binding domain of E2A is replaced by a putative DNA binding domain of PBX1. In 3 cell lines reported by Nourse et al. (1990), identical E2A-Prl mRNA junctions were observed, suggesting that the fusion transcripts are a consistent feature of this translocation.

T cell receptor delta locus

Levels of T cell receptor delta gene arrangements serve as suitable markers of commitment of the lymphoid lineage. TCE delta rearrangements – most frequently found in B precursor acute lymphoblastic leukemia have been studied.

With the study of all these genes we can confidently say that they are potential biomarkers of the leukemia disease. They affect the normal regulation of the body thereby causing leukemia. Below is a summary of all the genes investigated in table 7.

Table 7: Summary of 8 significant genes

Affy ID	Gene Name	Cellular	Molecular	Biological	Chromosome	Gene Type	SP Keyword
212148_at	pre-B-cell leukemia transcription factor 1	nucleus	transcription factor activity	regulation of transcription, DNA dependent	19(LL), 1(LL)	Gene with protein product, function known or inferred(ll)	Chromosomal translocation, DNA-binding, nuclear protein, phosphorylation, proto-oncogene, transciption, transcription regulation, Acute lymphoblastic leukemia
208616_s_at	protein tyrosine phosphatase type IVA, member 2	cell membrane, cytoplasm	prenylated protein tyrosine phosphatase activity	dephosphorylation, macromolecule metabolism	1(LL)	Gene with protein product, function known or inferred(ll)	Alternative splicing, hydrollase, lipoprotein, prenylation, protein phosphatase
216988_s_at	protein tyrosine phosphatase type IVA, member 2	cell membrane, cytoplasm	prenylated protein tyrosine phosphatase activity	dephosphorylation, macromolecule metabolism	1(LL)	Gene with protein product, function known or inferred(ll)	Alternative splicing, hydrollase, lipoprotein, prenylation, protein phosphatase
217143_s_at	T cell receptor delta locus				14(LL)	Gene with protein product, demonstrates somatic rearrangement(ll)	
202759_s_at	A kinase (PRKA) anchor protein 2	membrane		regulation of cell shape	9(LL)	Gene with protein product, function known or inferred(ll)	kinase, alternative splicing
35974_at	lymphoid-restricted membrane protein	endoplasmic reticulum membrane		hemocyte development	12(LL)	Gene with protein product, function known or inferred(ll)	phosphorylation, endoplasmic reticulum
211075_s_at	CD47 antigen (Rh-related antigen, integrin-associated signal transducer)	plasma membrane		integrin mediated signalling pathway	3(LL)	Gene with protein product, function known or inferred(ll)	Alternative splicing, cell adhesion, immunoglobin domain, transmembrane, signal

209014_at	melanoma antigen, family D, 1	cytoplasm, cell membrane		involved in apoptotic response	X(LL)	Gene with protein product, function known or inferred(ll)	Alternative splicing, repeat.
208894_at	major histocompatibility complex, class II, DR alpha	plasma membrane, lysosome, plasma membrane	MHC class II receptor activity	Immune response	6(LL)	Gene with protein product, function known or inferred(ll)	glycoprotein, immune response, MHC II, polymorphism, signal, transmembrane

4.2 Gene Exploration using SIBIOS

The advantage of using SIBIOS here is that we can work on multiple workflows and execute them in parallel. These results can be compared with each other to provide a broad picture of the genes under study. In addition, by undertaking several data analyses one might point out certain peculiarities in the genes which might indicate a potential discovery.

There are various online web applications which are very powerful and do diverse computations. It is very advantageous to exploit these applications and use them for biological analyses on a common platform. The issue of heterogeneity can be easily tackled; saving considerable time for the researchers. Many biologists themselves are not aware of the various formats and the conversions required in order to work with such gene lists.

4.3. Analysis of results in SIBIOS

Our first group of genes under study was a set of 9 genes for BCR which were explored using the SIBIOS system.

Shown in figure 25 is the workflow that has been constructed which executes two separate analyses on the same set of genes.

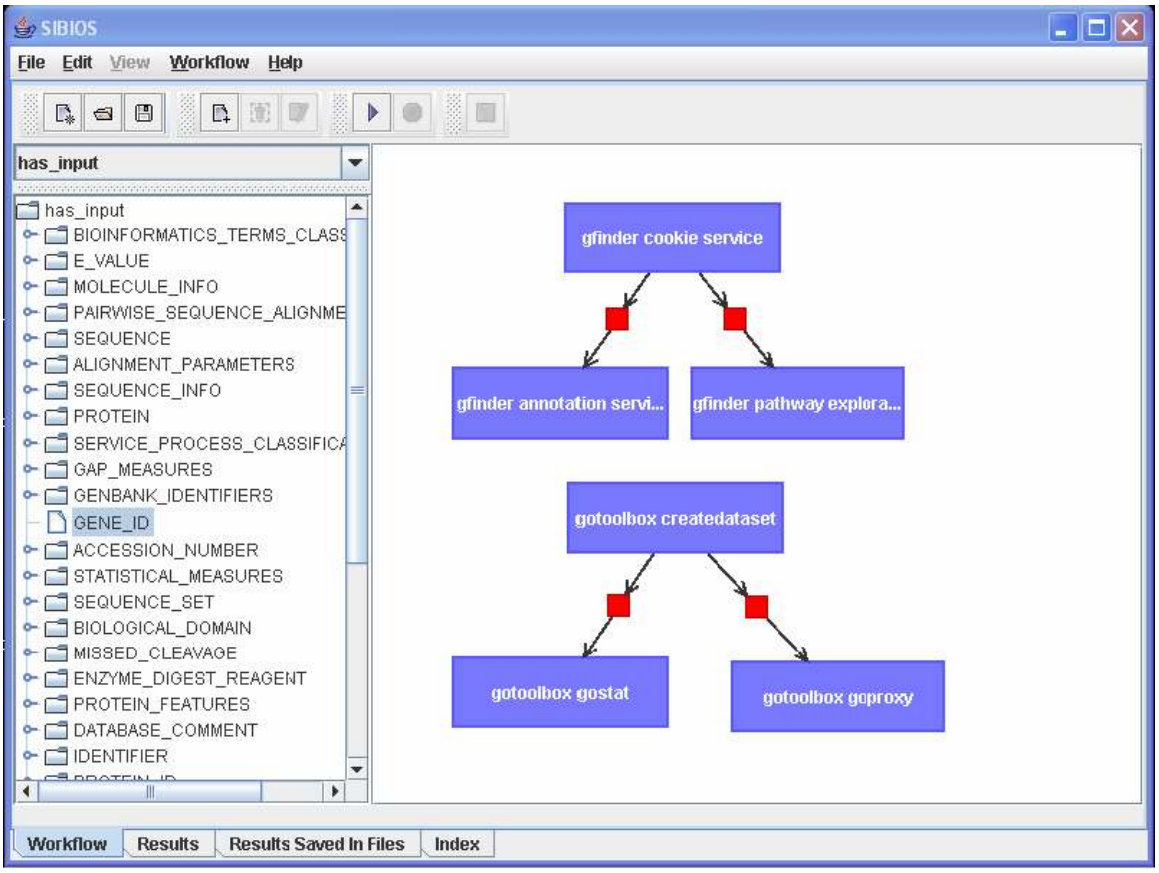


Figure 25: Workflow for analyzing microarray genes for biomarker discovery

The first workflow is of the web application GFinder (D Martucci, 2005). GFinder is a web tool that enables to better understand microarray experiment results and mine knowledge by examining user sequence ID lists or gene lists. It also applies certain clustering and statistical methods on the genomic annotations from several databases to better understand the input list. It has two modules i.e. the gene annotation module and the pathway exploration module. The annotations provide are gene symbol, cytogenetic location, Entrez gene ID, GO category

and disease category. In the pathway exploration module the biochemical pathway information is available to the user from the SIBIOS platform.

The second workflow involves the web application GOToolBox (GOToolBox, 2004). This service works on a data set which is generated by its own server, so it is necessary to generate a format which is acceptable to GOToolBox. This process is carried out by the first module in the workflow. The file information is then sent to the next modules i.e. GO-Stat and GO-Proxy. These two modules analyze the input gene lists using some statistical approaches which can be selected by the user. The GO-Stat module points statistically over or under represented terms within the input dataset. The statistical tests provided are Fisher exact test, Binomial-based probability calculations as well as *Bonferroni* correction for multiple testing. The analysis of enrichment or depletion of the input gene terms can be done with GO-Stat. The p-value correction can be done using the *Bonferroni* correction option. In the GO-Proxy module, clustering of genes is done based on the GO annotations. The distances between pairs of genes are calculated using the Czekanowsky-Dice formula. The distance matrix is then processed using a clustering algorithm, and then the resultant tree is then portioned given user-defined class attributes.

SIBIOS has therefore been adapted to assist the process of biomarker discovery. The models that had been provided above can be used as templates to be incorporated into SIBIOS. The capabilities to provide such protocols have been developed. Currently we have a partial template adapted to SIBIOS. The main challenge faced while adapting microarray experiments completely to SIBIOS is the issue to integrate stand alone tools. There are

various stand alone applications that provide various capabilities which need to be exploited by integrating such tools.

5. CONCLUSION

5.1. Conclusion

In this research, sample protocols for microarray analysis in the context of biomarker discovery has been designed, implemented and deployed. There have been two concepts that were proposed i.e.

1. Analysis First-Annotation Second model
2. Annotation First-Analysis Second model

The first model had two approaches i.e. the statistical analysis and the data mining method that has been studied and implemented. The framework for the second model has been designed and needs to be implemented. The protocols that have been devised for the purpose of biomarker discovery are useful in the following ways:

1. They provide a framework for conducting different experiments related to the identification of biomarkers.
2. The results of applying some of the protocols provide with a list of candidate biomarkers.
3. SIBIOS has been provided with a framework for facilitating a workflow for microarray analysis experiments. This provides a platform for efficient and time-saving research approach to biological researchers.

5.2. Future Work

The future work for this research area can be divided into two parts:

1. There has been a framework developed for Annotation first and Analysis Second module. This approach needs to be implemented and validated for its results.
2. The full implementation of the proposed protocols needs to be carried out within the SIBIOS system. This will necessitate the integration of bioinformatics services that implement statistical computations which form a part of microarray analysis.
3. An enhancement needs to be made in SIBIOS where it can incorporate the addition of standalone services. There are various efficient and resourceful applications for microarray applications like Genesis (Genesis, 2006) and Bioconductor (Bioconductor, 2001). They perform many statistical computations and are very helpful for significant gene selection in the context of biomarker discovery.

5.3. Summary

The aim of this research was to propose SIBIOS workflow integration system as a design environment for building *in-silico* experiments for biomarker discovery. To undertake this task, a deep understanding of the current procedures followed by researchers to extract the most biologically significant genes was paramount to this study.

In the course of validation of the proposed protocols, a case study on the BCR subtype of ALL was selected.

During the course of this research there were various details that were important to understand in the microarray analysis. They were needed to be studied and were not described in the published work of the researchers. They were subjective decisions that were taken during the filtering process; and that by moving to a standard procedure less subjectivity will exist in the process.

This thesis has led to the identification of potential biomarkers for pediatric Acute Lymphoblastic Leukemia. A framework has been provided on SIBIOS that assists in the design and deployment of workflows based on the integration of applications which work on microarray data.

REFERENCES

A Kohlmann, C. S., S Schnittger, M Dugas, W Hiddemann, W Kern, T Haferlach. (2004). Pediatric acute lymphoblastic leukemia (ALL) gene expression signatures classify an independent cohort of adult ALL patients. *Nature*, 63-71.

A Koike, T. T. (2004). Prediction of Protein Interaction Sites and Protein-Protein Interaction Pairs Using Support Vector Machines. *Protein Engineering, DEsign and selection, 17*, 165-173.

Affymetrix. (2002). *GeneChip Expression Analysis, Data Analysis Fundamentals*.

Agilent. (2000). http://www.chem.agilent.com/Scripts/PDS.asp?lPage=51.

B Ludascher, I. A., C Verkley, D Higgins, E Jaeger, M Jones, E A Lee, J Tao, Y Zhao. (2004). Scientific Workflow Management and the KEPLER System.

B M Fine, M. S., M Schrappe, M Ho, S Viehmann, J Harbott, L M Boxer. (2004). Gene expression patterns associated with recurrent chromosomal translocations in acute lymphoblastic leukemia. *Blood, 103*.

B. Miled, N. G., O. Bukhres, L. Lu, Y He, M Mahoui, J Chen. (2004). SIBIOS: A System for thr Integration of Bioinformatics Services. *CLADE*.

Bioconductor. (2001). http://www.bioconductor.org/.

C Creighton, S. H. (2003). Mining gene expression databases for association rules. *Bioinformatics, 19*, 79-86.

C Lottaz, R. S. (2005). stam - a Bioconductor compliant R package for structured analysis of microarray data.

C S Brown, P. C. G., P K Sorger. (2001). Image metrics in the statistical analysis of DNA microarray data.

D Martucci, O. G., M Masseroli. (2005). GFinder: Genome Function INtegrated Discoverer.

D Singh, P. F., K Ross, D Jackson, J Manola, C Ladd, P Tamayo, A Renshaw, A D'Amico, J Richie, E Lander, M Loda, P Kantoff, T Golub, W Sellers (2002). Gene Expression correlates of clinical prostate cancer behaviour. *Cancer Cell*, 203-209.

DAVID. (2006). http://david.niaid.nih.gov/david/version2/index.htm.

DIP. (1999). DIP: Databse of Interracting Proteins, http://dip.doe-mbi.ucla.edu.

Draghichi, S. (2003). *Data Analysis Tools for DNA Microarrays*: Chapman & Hall/CRC Mathematical Biology and Medicine Series.

Entropy-based-Discretization. (2002). Data Pre-processing http://sdmc.lit.org.sg/gedm/MainPage.html.

Enzyme. (2000). Enzyme search : http://www.expasy.org/enzyme/.

FuncAssociate. (2003). http://llama.med.harvard.edu/cgi/func/funcassociate.

G Cong, A. K. H. T., X Xu, F Pan, J Yang. (2004). FARMER: Finding Interesting Rule Groups in Microarray Datasets. *SIGMOD* 13-18.

Galperin, M. Y. (2005). The Molecular Biology Database Collection: 2005 update. *33*.

Gene. (2004). NCBI Gene : http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=search&term=.

Genesis. (2006). http://genome.tugraz.at/ [Electronic Version] from http://genome.tugraz.at/.

GO. (1999). http://www.geneontology.org/.

GOToolBox. (2004). GOToolBox: http://crfb.univ-mrs.fr/GOToolBox/index.php.

H Witten, E. F. (1999). Data mining: Practical Machine learning tools and techniques with java implementation. *Morgan Kaufmann*.

Hardware, M. (2004). http://ihome.cuhk.edu.hk/~b400559/array.html#Hardware [Electronic Version] from http://ihome.cuhk.edu.hk/~b400559/array.html#Hardware.

http://ihome.cuhk.edu.hk/~b400559/array.html#Hardware.

J Donald, J. A. J., D Michael, Mcguigsn, A R Patel, S Tomov, S Ross, T P Conrads, T D Veenstra, D A Fishman, G R Whiteley, E F Petricoin, L A Liotta. (2004). Clinical Proteomics and Biomarker Discovery. *New York Academy of Sciences*, 295-305.

J Li, G. D., K Ramamohanarao. (2001). Making use of the most expressive jumping emerging patterns for classification. *Knowledge and Information systems, 1*, 131-145.

KEGG. (1995). KEGG (Kyoto Encyclopedia of Genes and Genomes).

M Brown, W. N. G., D Lin, N Cristianini, C W Sugnet, T S Furey, M Ares, D Haussler. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS, 97*(1), 262-267.

M Carpenter, C. R., H McCorkle, C Lamartiniere. (2004). 2D-gel Proteomics in Biomarker Discovery.

M E Ross, X. Z., G Song, S A Shurtleff, K Girtman, W K Williams, H C Liu, R Mahfouz, S C Raimondi, N Lenny, J R Downing. (2003). Classification of Pediatric Acute Lymphoblastic Leukemia by gene Expression profile.

M K Kerr, G. A. C. (2001). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *PNAS*.

M Mahoui, L. L., J Chen, N Gao. (2005). A Dynamic Workflow Approach for the Integration of Bioinformatics Services. *Cluster Computing*.

M Mahoui, Z. M., S Srinivasan, M Dippold, B Yang, N Li. (2006). SIBIOS Ontology System: A Robust Package for Biological Data Integration in SIBIOS

DILS.

M Wilkinson, H. S., R Ernst, D Hase. (2005). BioMOBY Successfully Integrates Distributed Heterogeneous Bioinformatics Web Services. The PlaNet Exemplar Case, Plant Physiol., *138(1)*, 5-17.

MamasHealth. (2000). http://www.mamashealth.com/leukemia.asp [Electronic Version].

MATLAB. (1994). http://www.mathworks.com/.

MyGRID. (2004). http://www.myGrid.org.uk/.

OMIM. (1994). Online Mendelian Inheriitance in Man http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM.

PAM. (2005). PAM: Prediction Analysis for Microarrays, Class Prediction and Survival Analysis for Genomic Expression Data Mining.

PDB. (2003). http://www.rcsb.org/pdb/Welcome.do.

PIR. (2005). http://pir.georgetown.edu/.

R Fisler, B. C., O Scaros, PharmD. (2005). Biomarkers in clinical development: Implications for Personalized Medicine and Streamlining R&D. *Life Sciences Reports*.

R Stevens, P. B., S Bechhofer, G Ng, A Jacoby, N.W. Paton, C.A. Goble, A Brass. (2000). TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. *Bioinformatics*, 184-186.

R Tibshirani, T. h., B Narasimhan, G Chu. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS 2002 99:6567-6572 (May 14).*

S K Chatterjee, B. R. Z. (2005). Cancer biomarkers: knowing the present and predicting the future. 37-50.

S Raychaudhari, J. M. S., R B Altman. (2000). Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series. 455-466.

SAM. (2005). SAM: Significance Analysis of Microarrays, Supervised learning software for genomic expression data mining

Speed, T. (2003). *Statistical Analysis of Gene Expression Microarray Data*: Chapman & Hall / CRC.

SpotFire. (1996). http://spotfire.com/.

SwissProt. (2006). http://www.expasy.ch/sprot/.

T Yuen, E. W., R Pfeffer, B Ebersole, S Sealfon. (2002). Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Research*.

T. Oinn, M. A., J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, P. Li. (2004). Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*

T.R. Golub, D. K. S., P. Tamayo, C Huard, M Gaasenbeek, J.P. Mesirov, H Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*.

Uniprot. (2006). www.uniprot.org.

W Kuo, T. j., A butte, L Ohno-Machado, I Kohane. (2002). Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics, 18(3)*, 405-412.

Wilkipedia. (2001). http://en.wikipedia.org/wiki/Main_Page.

APPENDIX

XML Schema for GFinder annotation service demonstrating the <ENTRY_URL> tag.

```
<?xml version="1.0" encoding="ISO88591" standalone="yes" ?>
<TOP>
<SERVICE_APPLICATION>
<SERVICE_NAME>gfinder_cookie_service</SERVICE_NAME>
<PATH_PAGE>
<PAGE_CHARACTER>{title}uploaded gene list</PAGE_CHARACTER>
<EXTRACT_LINK>sh='{title' t='title}' l='}' r='{'</EXTRACT_LINK>
</PATH_PAGE>
</SERVICE_APPLICATION>
<URL_STRING>
        <SINGLE_QUERY>
        <BASIC_URL>http://promoter.bioing.polimi.it/gfinder/eng/abc_process.asp</BASIC_URL>
        <METHOD>P</METHOD>
        <UPLOAD>filename=XXXX</UPLOAD>
        <WRITE>null</WRITE>
        <ENTRY_URL>http://promoter.bioing.polimi.it/gfinder/default.asp</ENTRY_URL>
        <DEFAULT_PARAMETERS />
        </SINGLE_QUERY>
</URL_STRING>
 <EXTRACTION_RULES>
        <FILECODE>sh='input name="file"' t='input name="estensione"' l='value="'
        r='"}'</FILECODE>
        <EXP_NO>sh='{input name="esperimento" type="hidden" id="esperimento"' t='}' l='value="'
        r='"'</EXP_NO>
        <CEXP_NO>sh='{input name="esperimento" type="hidden" id="esperimento"' t='}'
        l='value="' r='"'</CEXP_NO>
        <COOKIE1>null</COOKIE1>
</EXTRACTION_RULES>
<INTERFACE_PARAMETERS>
<MAN_OR>
<PARAMETER>
<NAME>UPLOAD</NAME>
<DESCRIPTION>File Name Location</DESCRIPTION>
 <TYPE>
<UPLOAD_FILE />
</TYPE>
<DEFAULT_VALUE />
</PARAMETER>
</MAN_OR>
</INTERFACE_PARAMETERS>
</TOP>
```

XML Schema for GFinder annotation service demonstrating the <WRITE> tag.

```
<?xml version="1.0" encoding="ISO88591" standalone="yes" ?>
 <TOP>
 <SERVICE_APPLICATION>
  <SERVICE_NAME>gfinder_annotation_service</SERVICE_NAME>
 <PATH_PAGE>
  <PAGE_CHARACTER>{title}annotation</PAGE_CHARACTER>
  <EXTRACT_LINK>sh='{title' t='/title}' l='}' r='{'</EXTRACT_LINK>
  </PATH_PAGE>
  </SERVICE_APPLICATION>
 <URL_STRING>
 <SINGLE_QUERY>
  <BASIC_URL>http://promoter.bioing.polimi.it/gfinder/eng/gesesperimenti2.asp</BASIC_URL>
  <METHOD>P</METHOD>
```

```
 <ENTRY_URL>null</ENTRY_URL>
        <WRITE>vistempo=1!livello=!soglia=!Submit=Execute!nodo=!nomenodo=!ontologia=!classe=!t
        empo=!numrighe=!numrighetot=!ordina=!maggmin=!reload=!phenotype=!bg=undefined!nuov
        afin=undefined!risultati=undefined!Y1=undefined!Y0=undefined!lev=undefined!conta=undefin
        ed!colonna=undefined!test=undefined!calcola=undefined!sottoclassi=undefined!distance=und
        efined!nome=undefined!tabella=undefined!numcolonne=undefined!tabtemp=undefined!nomee
        sperimento=undefined!colonneregolaz=undefined!ID=undefined!chip=undefined!llid=undefine
        d!accno=undefined!pfam=undefined!nomepfam=undefined!cod=undefined!kegg=undefined!pa
        thway=undefined!versione=undefined!numcategorie=undefined!cominciada=undefined!radiob
        utton=Visualizza!cutoff=undefined!tuttioannotati=undefined!nolistapredef=undefined!listapre
        def=undefined!maxrighe=undefined!evidence=undefined!specie=undefined!speciepresente=u
        ndefined!minIDs=undefined!conversione=undefined!conversionepresente=undefined!poolcate
        g=undefined!finestragrande=undefined!rappresentate=undefined!sogliaR=undefined!correzion
        e=undefined!viscolonna0=undefined!colonna0=1!ordinacol0=ASC!viscolonna1=si!colonna1=2!
        ordinacol1=ASC!viscolonna2=si!colonna2=3!ordinacol2=ASC!viscolonna3=si!colonna3=4!ordin
        acol3=ASC!viscolonna4=undefined!colonna4=5!ordinacol4=ASC!viscolonna5=undefined!colonn
        a5=6!ordinacol5=ASC!viscolonna6=undefined!colonna6=7!ordinacol6=ASC!viscolonna7=undefi
        ned!colonna7=8!ordinacol7=ASC!viscolonna8=undefined!colonna8=9!ordinacol8=ASC!viscolon
        na9=undefined!colonna9=10!ordinacol9=ASC!viscolonna10=undefined!colonna10=11!ordinaco
        l10=ASC!viscolonna11=undefined!colonna11=12!ordinacol11=ASC!viscolonna12=undefined!co
        lonna12=13!ordinacol12=ASC!viscolonna13=undefined!colonna13=14!ordinacol13=ASC!viscol
        onna14=undefined!colonna14=15!ordinacol14=ASC!viscolonna15=undefined!colonna15=16!or
        dinacol15=ASC!viscolonna16=undefined!colonna16=17!ordinacol16=ASC!viscolonna17=undefi
        ned!colonna17=18!ordinacol17=ASC</WRITE>
 <DEFAULT_PARAMETERS />
 <GENE_NAME>viscolonna1=XXXX</GENE_NAME>
 <GENE_SYMBOL>viscolonna2=XXXX</GENE_SYMBOL>
 <EXP_NO>esperimento=XXXX</EXP_NO>
 <CEXP_NO>codesperimento=XXXX!esperimento=XXXX</CEXP_NO>
 </SINGLE_QUERY>
 </URL_STRING>
<EXTRACTION_RULES>
<ANNOTATION>
 gh='{table class="normale" width="1%" border="1" align="center"}' t='{/table}' s='{tr'
 <SEQUENCE_ID>sh='{td' t='!nbsp;{/td}' l='}' r='!nbsp;{/td}'</SEQUENCE_ID>
   <GENE_NAME>sh='(http://ihome.cuhk.edu.hk/~b400559/array.html#Hardware)'   t='!nbsp;{/td}'
        l='}' r='!nbsp;{/td}'</GENE_NAME>
<GENE_SYMBOL>sh='a                target'              t='!nbsp;{/td}'              l='}'
        r='(http://ihome.cuhk.edu.hk/~b400559/array.html#Hardware)'</GENE_SYMBOL>
 </ANNOTATION>
 </EXTRACTION_RULES>

<INTERFACE_PARAMETERS>

<MAN_OR>

<PARAMETER>
 <NAME>GENE_NAME</NAME>
 <DESCRIPTION>Gene Name</DESCRIPTION>

<TYPE>
<TEXTAREA>
<CHOICE>
 <VALUE>si</VALUE>
 </CHOICE>
 </TEXTAREA>
 </TYPE>
 <DEFAULT_VALUE />
</PARAMETER>
<PARAMETER>
 <NAME>EXP_NO</NAME>
 <DESCRIPTION>exp Name</DESCRIPTION>

<TYPE>
<TEXTAREA>
```

```
  <CHOICE>
   <VALUE>si</VALUE>
   </CHOICE>
   </TEXTAREA>
   </TYPE>
   <DEFAULT_VALUE />
   </PARAMETER>
  <PARAMETER>
   <NAME>CEXP_NO</NAME>
   <DESCRIPTION>cexp Name</DESCRIPTION>

  <TYPE>
  <TEXTAREA>
  <CHOICE>
   <VALUE>si</VALUE>
</CHOICE>
</TEXTAREA>
</TYPE>
   <DEFAULT_VALUE />
</PARAMETER>
</MAN_OR>
</INTERFACE_PARAMETERS>
</TOP>
```

CURRICULUM VITAE

Bhavna Choudhury, bchoudhu@iupui.com (317-341-5874),
319 N West St, # 339, Indianapolis, IN 46202

OBJECTIVE

A full time position in the discipline of software development with major responsibilities that will effectively utilize my technical, leadership and organizational skills.

SUMMARY

- Experienced in development of both front-end and back-end applications using Java 2, JSP 1.2, Java Script, JDBC, XML and HTML.
- Expertise in Object Oriented Programming and Relational Databases.
- Experienced in working with Relational Databases like Oracle 9i/8i and MS Access.
- Experienced in coding Stored Procedures, Functions, Packages and Triggers using PL/SQL.
- Experienced in working with Application and Web Servers Apache Tomcat 4.1/5.1.
- Knowledge of writing AJAX (Asynchronous Java Script and XML) pages.
- Extensive knowledge working with IDE tools like Eclipse 3.1.
- Experienced in requirement gathering, prototype designing and documentation.
- Good understanding in Software Life cycle phases like Feasibility, System studies, Design, Coding, Testing, Debugging, implementation and Maintenance.
- Good analytical skills and the determination to solve problems.
- Excellent English verbal and written communication skills.
- Dependable and reliable in supporting and enabling team effort.

EDUCATION

MS : Bioinformatics, Indiana University Purdue University Indianapolis, (Aug 2004 - June 2006)
BS : Computer Science, JSS Academy of Technical Education, Noida, (Sep 2000 - May 2004)

PROFESSIONAL SKILLS

Languages	Java , C, C++, Perl
Technologies	J2SE 1.4, J2EE 1.4, JSP1.2, RMI, JDBC, XML, SAX, JAXP, XQuery, XPath
Web Server	Tomcat 4.1/5.1
Data Bases	Oracle 9i/8i and MS Access, SQL Server, SQL Loader, PL/SQL
Operating Systems	Windows based (2000, 9x, NT) , UNIX, MS-DOS
Scripting Language	HTML, Java Script, MS FrontPage
Tools	Eclipse, ANT, MATLAB 7, SPSS, Spotfire
Office Tools	MS- Word/Excel/PowerPoint/FrontPage
Biological Database	Swiss-Prot, Protein Data Bank, GeneBank, ENBL, GO, DIP, OMIM, NCBI
Course work	Computer Networks, Analysis and Design of Networks, Biostatistics

WORK EXPERIENCE/ PROJECTS UNDERTAKEN

Institution:Indiana University Purdue University Indianapolis, IN **Jan 2005 – June 2006**
Project: System for Integration of Bioinformatics Services (SIBIOS)
Role: Java Developer

SIBIOS is an application designed to facilitate biological workflows for researchers. This application utilizes the web applications which become modules of the workflow. A reasoning system is provided by an ontology. It is based on the protocol that involves making http connections with the web applications and extract results through these remote applications. These results are then provided on the SIBIOS platform

Responsibilities:

- Involved in requirement gathering and designing the flow.

- Engineered SIBIOS to enable researchers with a generic framework to work on in-silico experiments in biomarker discovery
- Used Eclipse 3.1 as the IDE tool for developing different modules.
- Developed the module to maintain sessions of bioinformatics applications through SIBIOS.
- Used ANT tool for creating and deploying the .jar files for the system

Environment: Eclipse 3.1, J2SE1.4, HTML, XML

Institution:Indiana University Purdue University Indianapolis, IN **Jan 2006 – May 2006**
Project: FARMER (Data Mining Algorithm for microarray datasets)
Role: Java programmer

This is an application that mines interesting association rules from the microarray datasets that discovers significant gene patterns indicating the presence of cancer or not in a patient.

Responsibilities:

- Built a java based application that finds interesting rule groups from a microarray dataset based on the FARMER algorithm.
- Did mining of association rules which help in Biomarker detection from Microarray datasets based on the gene expression intensities.

Environment: Eclipse 3.1, J2SE1.5

Institution:Indiana University Purdue University Indianapolis **Jan 2006 – May 2006**
Project: Integrated MicroArray Gene Explorer (IMAGE)
Role: Java/J2EE developer

This tool was developed to maintain microarray experiments and query biological databases to obtain various information related to the significant genes of interest.

Responsibilities:

- Developed a tool in java for microarray data analysis for gene annotation using public databases.
- The tool helps in analyzing data for gene expression levels for information on diseases, associated proteins, etc.
- Used Eclipse Java IDE tool for creating JSPs and XML.
- Involved in JDBC coding for establishing connection with the database and connecting to MS Excel documents.

Environment: J2SE1.2,JSP, JDBC, JavaScript, Oracle 10g, XML, SAX

Institution:Indiana University Purdue University Indianapolis **Aug 2005 – Dec 2005**
Project: Image Search Engine
Role: Java programmer

Developed a system that searches for images from online PDF papers and journals in the biological domain. User inputs a query in the form of gene names or protein IDs and retrieves papers based on such queries and gives the image on the result page.

Responsibilities:

- Designed database tables using Oracle 10g which contained BLOB type of data to store images as a bitmap.
- Involved in JDBC coding for establishing connection with the database
- Developed Perl scripts using the LWP module to download mass PDF documents from a website together.
- Involved in deployment of the application in Tomcat server

Environment: J2SE1.2,JSP, JDBC, Oracle 10g

Institution:Indiana University Purdue University Indianapolis **Jan 2005 – May 2005**
Project: Informatics Faculty Search System
Role: Java programmer
Responsibilities:

- Built a faculty search system for the entire faculty of the Indiana School of Informatics.
- The user could search courses provided by the faculty, his research interests, etc.
- Developed the module to create and edit profile of faculties.

Environment: J2SE1.2,JSP, JDBC, JavaScript, Oracle 9i