Dedicated to my family

Acknowledgements

I would like to send my faithfull and deepest gratitude to my supervisor, Asso. Prof. Ha Quang Thuy who is always behind me and give me valuable encouragement, advices not only in my research activities but also in daily life. This thesis must have been imcomplete if without enthusiastical help and encouragement of Prof. Arndt von Haeseler from Center for Integrative Bioinformatics Vienna-CIBIV, Austria. It's very kind of you to offer me an opportunity to do the research on Bioinformatics field of study.

Thanks to all members of the Data Mining research group for the seminar topics held periodically from which I've gotten lot of meaningfull knowledge. Anyway, thanks to the Information Systems Department, COLTECH, VNUH for it's friendly and suitable to doing the scientific research environment. This work was supported in part by the National Project "Developing content filter systems to support management and implementation public security - ensure policy" and the MoST-203906 Project "Information Extraction Models for discovering entities and semantic relations from Vietnamese Web pages".

Finally, I would like to thank Mr. Le Si Vinh and Mr. Bui Quang Minh for their continued help during the time of implementing this thesis.

Foreword

cDNA microarray data analysis has become an attracted field of study recent years. Nowadays the capability of simultaneously measuring the activity and interactions of thousands of genes using cDNA microarry experiments provides a new and deep insight into the mechanisms of living systems. The direct applications of microarrays include gene discovery, disease diagnosis and prognosis, drug discovery (pharmacogenomics), and toxicological research. These have achieved a lot of valuable results.

With microarray data, scientists can address many main scientific tasks. They are the identification of coexpressed genes, discovery of sample or gene groups with similar expression patterns and the study of gene activity patterns under various conditions (e.g., chemical treatment). The identification of genes whose expression patterns are highly expressed with respect to a set of discerned biological entities (e.g., tumor types) is also one of these scientific tasks. More recently, more interesting scientific tasks based on microarray have been developed such as the discovery, modeling, and simulation of gene regulatory networks, and the mapping of expression data to metabolic pathways and chromosome locations.

All the above mentioned scientific tasks require one or more different data analytical techniques. The thesis explores the interesting and challenging issues concerned with the microarray data analysis in order to lay out the best foundation for futher research. The content of the thesis is organized as follows.

Chapter 1 introduces main challenges and difficulties on microarray data analysis field of study. The process to design a cDNA microarray experiment is mentioned first. Then we describe all aspects relate to the problem of analysis the cDNA data. Moreover classification issues in cDNA data are mainly focused.

Chapter 2 first introduces two most popular graph based ranking algorithms, HITS (Kleinberg, 1994) and PageRank (Brin and Page, 1998). Second we survey the modeling of gene network including Boolean Network, Bayesian Network, Additive regulation model for inference the gene regulatory networks from gene experiment dataset are also included in this section.

Chapter 3 explains for the thesis' proposed method for gene selection in sample classifying problem as the result of applying graph based ranking algorithms mentioned above. Then the final part shows the results from an analysis using two gene expression datatsets available on the internet. They are from yeast *Saccharomyces cerevisiae* and Leukeima disease. We also discuss in the computational issue and its biological meaning.

Chapter 1

Introduction to Gene Expression Data

1.1. Gene Expression

Deoxyribonucleic acid (DNA) is the central issues when learning to understand the gene expression. Both DNA and RNA are polymers, i.e., the molecules whose structure is in the form of a linear strand or sequence of members of a small set of subunits called nucleotides. Each nucleotide consists of a base, attached to a sugar. The sugar is in turn attached to a phosphate group. In the DNA, the sugar is deoxyribose and the bases are named Guanine (G), Adenine (A), Thymine (T), and cytosine (C); and while in the RNA the sugar is ribose and the bases are Guanine (G), Adenine (A), Uracil (U), and Cytosine (C) (Alberts et al, 1989). DNA sequences are organized as a double-stranded polymer where one base, via hydrogen bonds, will bind with bases on the complementary strands via hydrogen bonds according to the rule: Adenine binds to Thymine and Guanine to Cytosine, respectively [35] (Figure 1.1)

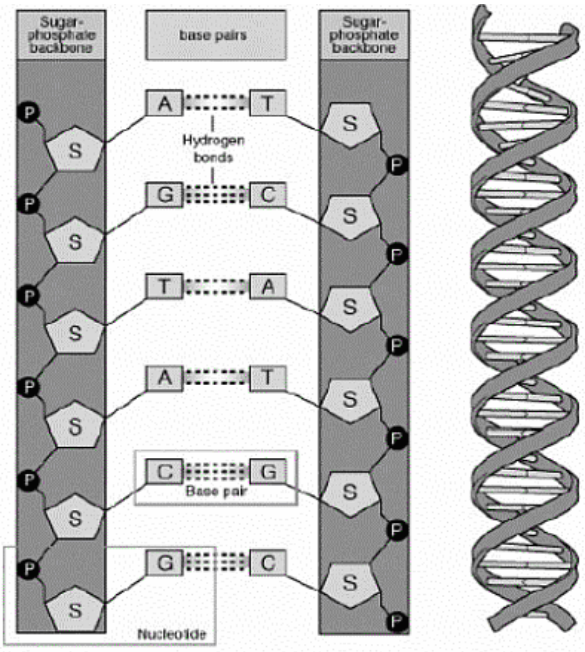

Figure 1.1: Structure of DNA sequence

Due to the complementary characteristic of double-stranded structure, the DNA sequences have the capability of encoding genetic information. They can also replicate themselves by using each strand as a template to generate a new complementary strand.

Genes are unique regions in the DNA sequences and all genes within a cell comprise the *genome*. The information necessary for synthesizing proteins, the material responsible for all functionalities of a cell, are all encoded in the *genome*. Moreover this information also control the expression level of proteins in cells. A variety of important functions of proteins in the cells are ranging from structural (e.g., skin, cytoskeleton) to catalytic (enzymes) proteins, to proteins involved in transport (e.g., haemoglobin), and regulatory processes (e.g., hormones, receptor/signal transduction), and to proteins controlling genetic transcription and the proteins of the immune system .

DNA self-replication and protein synthesis are two crucial processes of a cell[35]. The protein synthesis consists of two steps. (Figure. 1.2)

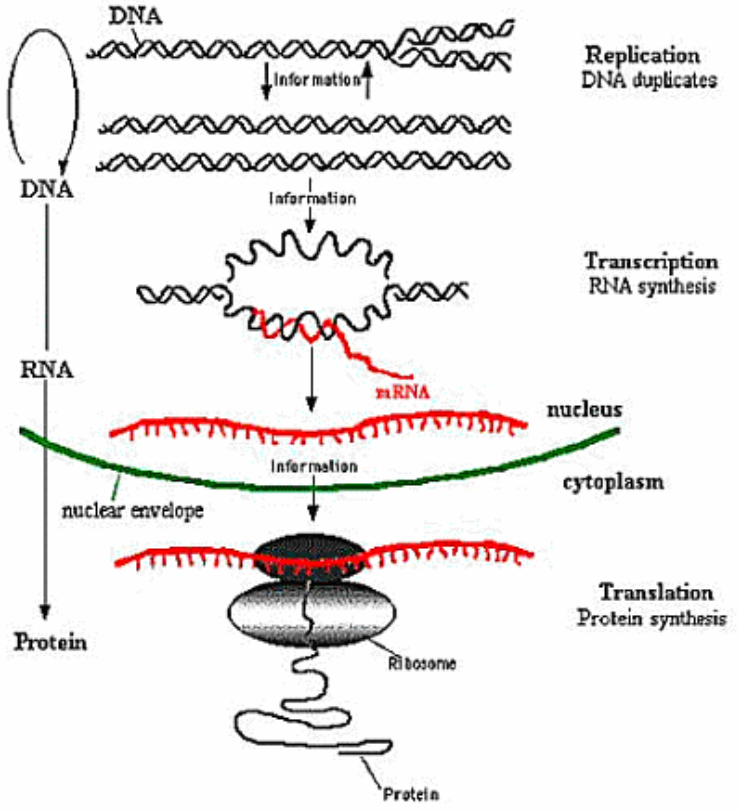

Figure 1.2: Process of gene expression

At the first step, the template strand of the DNA is transcribed into the messenger RNA (mRNA), an intermediate molecular sequence. mRNA is mainly identical to DNA except that all Ts are replaced by Us. At the second stage, the RNA is translated into protein, in which three continuous bases (codon) in the mRNA are replaced by one corresponding amino acid. The overall process consisting of transcription and translation is also known as gene expression. Notice that not all genes in the *genome* are transcribed into RNA and expressed as proteins.

In molecular biology, the term ***proteome*** is used to indicate all the proteins that are synthesized from the gene expression processes of the whole *genome*. Chemically, proteins are polymers composed of 20 amino acids. The protein sequences are themselves the primary structure. Based on this primary structure, the three-demensional conformation of proteins is generated by the so-called "folding" process. It's turn out to be very difficult to capture and describe precisely the processes involved in protein folding. The protein's biological function is determined by three-dimensional arrangement of amino acid sequence. For each amino acid sequence, among all of possible conformation of proteins there are always more than one stable three-dimensional structures. They are called the protein's *native states* and can switch with each others according to their interactions with other molecules.

1.2. DNA microarray experiments

A DNA microarray (also commonly known as gene or genome chip, *DNA chip*, or *gene array*) is a collection of microscopic DNA spots attached to a solid surface, such as glass, plastic or silicon chip forming an array for the purpose of expression profiling, monitoring expression levels for thousands of genes simultaneously [19].

Many biomolecular studies showed that the problem of measuring the real gene expression level is very important. Based on the process of gene expression explained above, one DNA produces only one corresponding mRNA and this mRNA in turn produces only one corresponding protein. That means protein and mRNA abundance are proportional, so the highly accurate information on protein

abundance can be revealed in the DNA microarray experiments which do measure the abundance of mRNA instead of measuring the abundance of proteins. But in practise, the gene expression scenario is much more dynamic and complicated than simplified scenario mentioned above. Proteins are formed and modified in various mechanisms, not simply according to the simplified process of direct one-to-one mapping from DNA to mRNA to protein. Moreover the cell's genome itself is subject to alterations [35]

Despite of not taking into account no information about possible differential translation rates, about post-translational modification and different forms of processed mRNA, but the cDNA microarray experiments still provides us some valuable information quickly and fairly easily in replace. Beside, it is still very expensive to study thoroughly on protein expression and modification because of the involvement the highly specialized and sophisticate techniques. There are still many dificult problems that need to be resolved thoroughly before the high-throughput protein-detecting arrays should be used broadly. This's reason why the scientists must conduct the DNA microarray studies through measurement mRNA.

There are some techniques developed for measuring gene expression levels such as northern/southern blots, spotted cDNA microarrays, spotted oligonucleotide microarrays, and Affymetrix chips [35]. All these techniques exploit the process of hybridization between two strands of the DNA duplex. Hybridization is the process of combining complementary, single-stranded nucleic acids into a single molecule. Nucleotides will bind to their complement under normal conditions, so two perfectly complementary strands will bind to each other readily (Figure 1.3) [19]. The rate and proportion at which the hybridization process happens depend on density of the original single-stranded polymers and on the degree of alignment between these sequences.

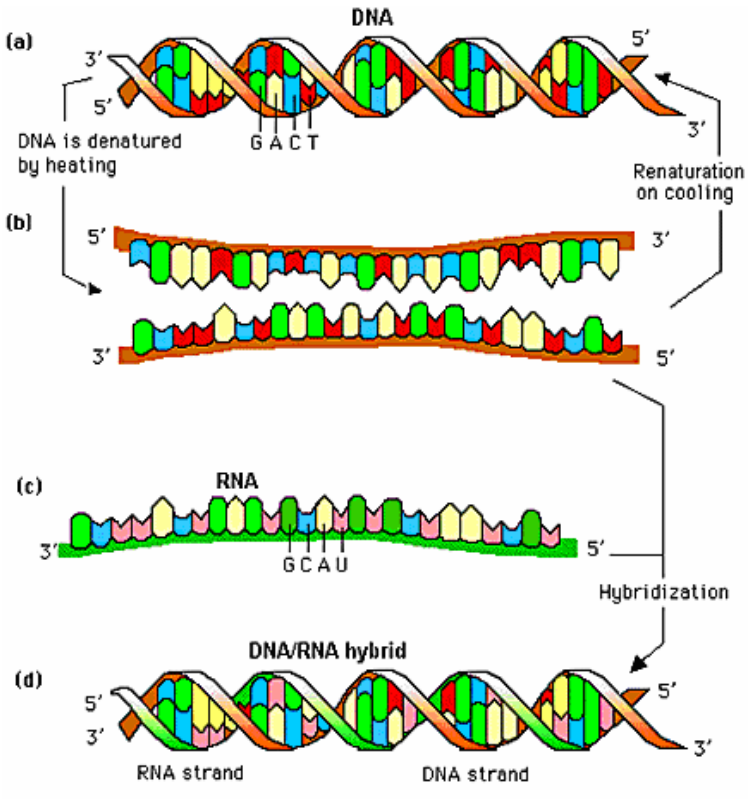

Figure 1.3: Process of hybridization

Before doing the experiment, the mRNA must be labeled with reporter molecules that is the fluorescent dyes (fluors). The cyanine 3 (Cy3) and cyanine 5 (Cy5) are two particular reporter molecules most likely used in microarray experiments [35]. For the purpose of best illustrating the process of deploying a microarray experiment, the DNA microarray experiment is supposed to have two samples of transcribed mRNA from two different sources, *sample 1* and *sample 2*. The mRNA are extracted from multiple copies of many genes contained in both sample sources. The experiment also needs a *probe*, which is a short piece of DNA (on the order of 100-500 bases) that is denatured (by heating) into single strands and then radioactively labeled [19]. The relative abundance of the mRNA complementary to the probe sequence within *sample 1* and *sample 2* are specified through the following process [35] (Figure 1.4):

Step 1. Prepare a mixture consisting of identical probe sequences.
Step 2. Label sample 1 with green-dyed reporter
Step 3. Label sample 2 with red-dyed reporter.

Step 4. Sample 1 and sample 2 are mixtured with each other and completely hybridized with the probe mixture.

Step 5. Gently stir for five minutes.

Step 6. Filter the mixture to obtain only those probe sequences that have hybridized.

Step 7. Measure the amount or intensity of green and red in the filtered mixture, and the relative abundance of the probe sequence may be output.

Because the RNA is inherent instable in chemical characteristic, so instead of using with mRNA at intermediate steps, the DNA microarray experiments use a more stable complementary DNA (cDNA) obtained by reverse transcription from mRNA at intermediate steps.

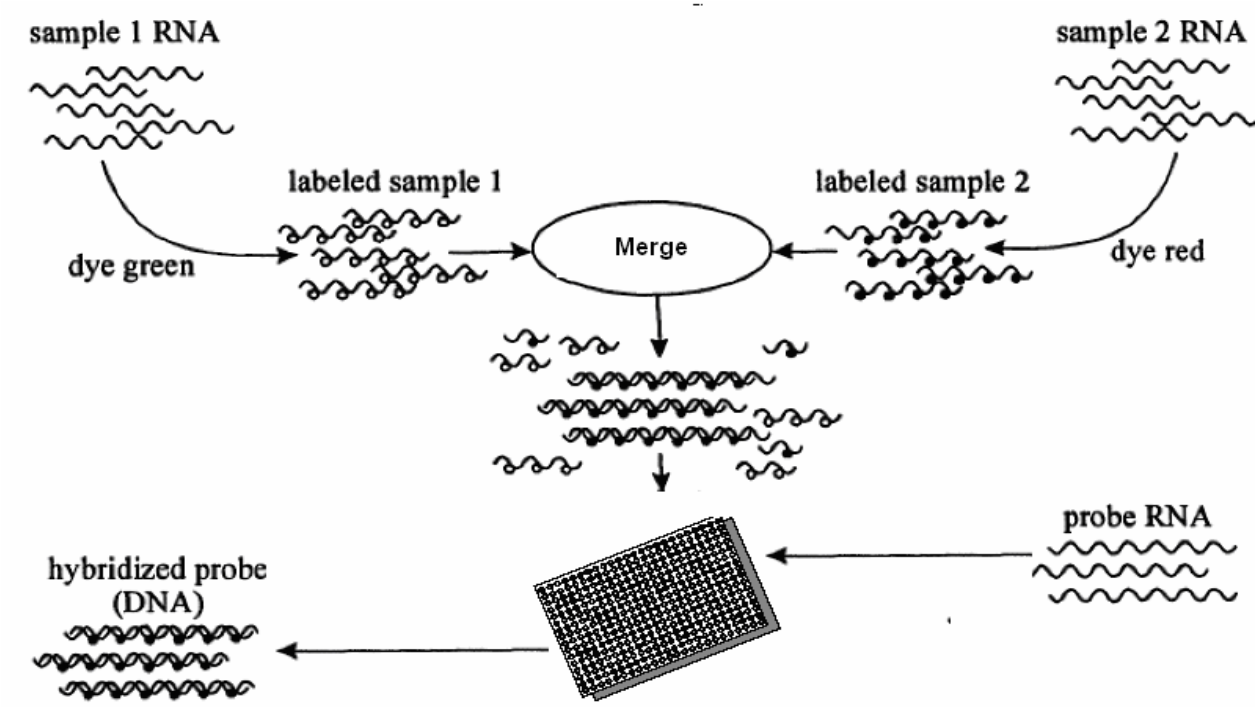

Figure 1.4: Competitive hybridization

1.3. High-throughput Microarray Technology

Genes are expressed at different levels within different kinds of cells, and even within the same cells on different conditions, for example, physical, chemical, and biological conditions. The purpose of a cDNA microarray experiment is to simultaneously measure the expression level of all genes needed to be studied in

different cells within different conditions. As the result of the transcription differences between normal and diseased cells or different patterns of abnormal transcription will be revealed and learned thoroughly.

Let consider a simple scenario in which we want to study the roles of four different genes ***a***, ***b***, ***c*** and ***d*** in two different forms A and B of the same type of cancer. The experiment is deployed on ten patients, six of them suffer from A and the rest four from B. The following are seven steps for completing the experiment (Figure 1.5) [35].

Step 1. Probe preparation.

One DNA microarray is prepared for each patient. A sufficient number of the probes, cDNA sequences with 500 to 2500 nucleotides in length, are created. These cDNA sequence mixtures are then affixed to the array (a glass slide) in a grid-like fashion form. For large microarray experiments with thousands of genes, we need to know where a particular gene is located on the array to trace back the corresponding information later.

Step 2. Target sample preparation.

The target is the mRNA extracted from the cells of one patient, then purified and labeled with reporter molecules. The color red is chosen since it can be easily recognized by human eyes.

Step 3. Reference sample preparation.

Reference is a mRNA sequence that must be prepared and labelled in a color different from that of target samples. The abundance of target mRNA is measured on the comparison to the reference sample refered to as a baseline. The reference samples are divided into two types, standard and control reference. Standard references are mRNAs unrelated to the target samples of the experiment. Whereas , the control references are related to the experiment. For example, in a disease study, the control references may be the mRNAs from normal tissues.

Step 4. Competitive hybridization.

The target and reference mRNAs will both hybridize competitively with probes on array.

Step 5. Wash up the dishes.

This phase is done right after the hybridization process to eliminate any reference and target materials that were not hybridized. The color intensity of each spot is recorded into the microarray.

Step 6. Detect red-green intensities.

Scan the array to determine how many target and reference mRNAs are bound to each spot using a device equipped with a laser and a microscope. This produces a high-resolution, false-color digital image.

Step 7. Determine and record relative mRNA abundances.

At this stage, we need an image processing tool to derive the actual level of expressions.

The seven steps mentioned above are carried out on the ten patients to produce ten arrays. Once finished, a so-called gene expression data matrix is created for later analysis. At the end, the following table is obtained (Figure 1.6).

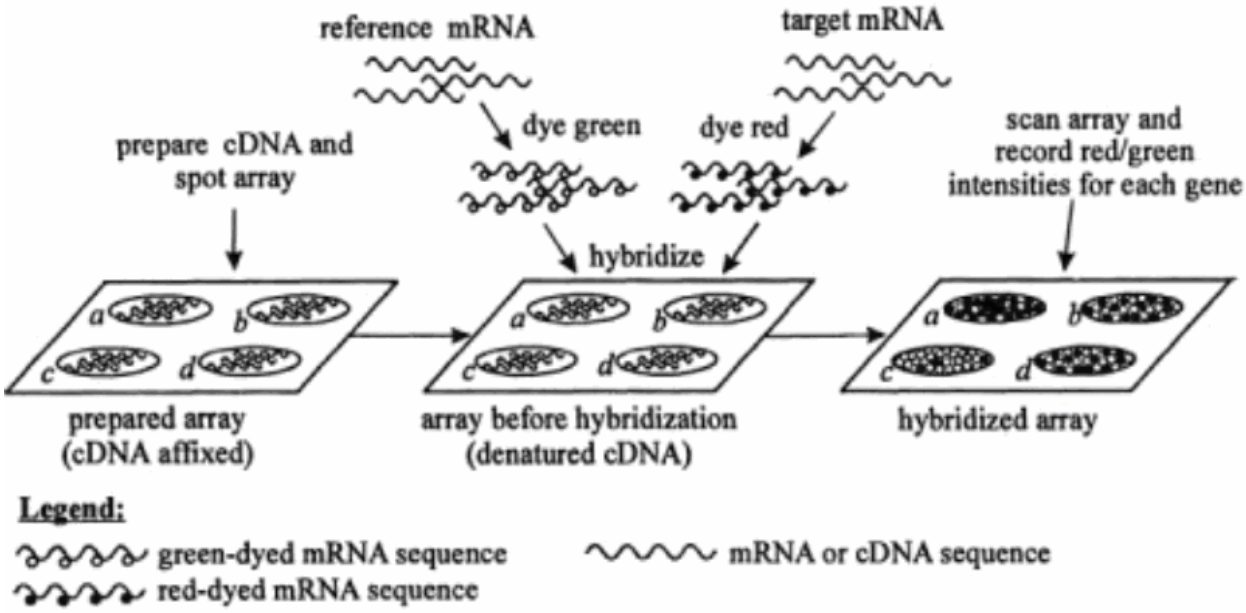

Figure 1.5: A 4-Gene Microarray Experiment

Patient#	1	2	3	4	5	6	7	8	9	10
Tumor	A	A	A	A	A	A	B	B	B	B
r(a)	5.00	1.23	4.05	3.44	2.85	3.04	0.44	0.43	0.11	0.14
r(b)	0.98	0.78	1.05	1.20	?	3.12	1.34	1.25	4.00	0.67
r(c)	0.33	1.50	0.43	0.52	0.42	0.60	2.56	?	5.00	2.44
r(d)	0.85	0.99	1.00	0.98	0.62	1.02	1.43	1.04	0.94	0.84

Figure 1.6: A matrix as the result of microarray experiment

Carefully look at the above table, we can derive several conclusions relating to the tendency in the expression level of genes within each form of cancer type as following [35]:

Conclusion 1:

For patients of tumor A there is likely a tendency that the expression levels of gene ***a*** seem to be two times or more higher than the reference level 1.0. While the tendency to be twice or more lower than 1.0 level is true to ***a***'s expression levels within patients of tumor B. This observation suggests that the gene ***a*** may be involved in deciding into which form A or B the tumor cells will develope.

Conclusion 2

Gene ***b*** and ***d*** have the expression values almost around 1.0, and thus said to be not differentially expressed across the studied tumors. This suggests that these genes are not involved in the cancer type.

Conclusion 3

Within all ten patients, the expression levels of gene ***a*** and ***c*** are in reverse relationship. If the expression levels of gene ***a*** are high, then those of gene ***c*** will be low in the same patient and vice versa. This suggests us a negatively coregulatory relationship between these two genes.

The gene expression data, that the above table is one example, can be generally represented in the form of an n x m expression matrix E as followed:

$$E = (x_{ij}) = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}$$

where x_{ij} denotes the expression level of sample j for gene i, for j=1,…m, and i=1,…n [14].

The column or row vectors in this matrix E can be optionally interpreted as variables or observations respectively. With this notion, the i^{th} gene profile G_i can be defined as the row vector and the array profile A_j. can be defined as the column vector j of the matrix E:

$$G_i = (x_{i1}, x_{i2}, \dots, x_{im})$$
$$A_j = (x_{1j}, x_{2j}, \dots, xn_j)$$

1.4. Microarray data analysis

Microarray data analysis is an interdisciplinary study of the cell behavior with the help of statistical and computational methods. Moreover these methods also need adaptation to the special characteristics of cDNA microarray data. The following picture describes all processes involving in microarray data analysis. The scope of this thesis only focuses on step 4, pre-process matrix, and partially on some tasks in step 5, i.e., classification and gene regulatory network problems.

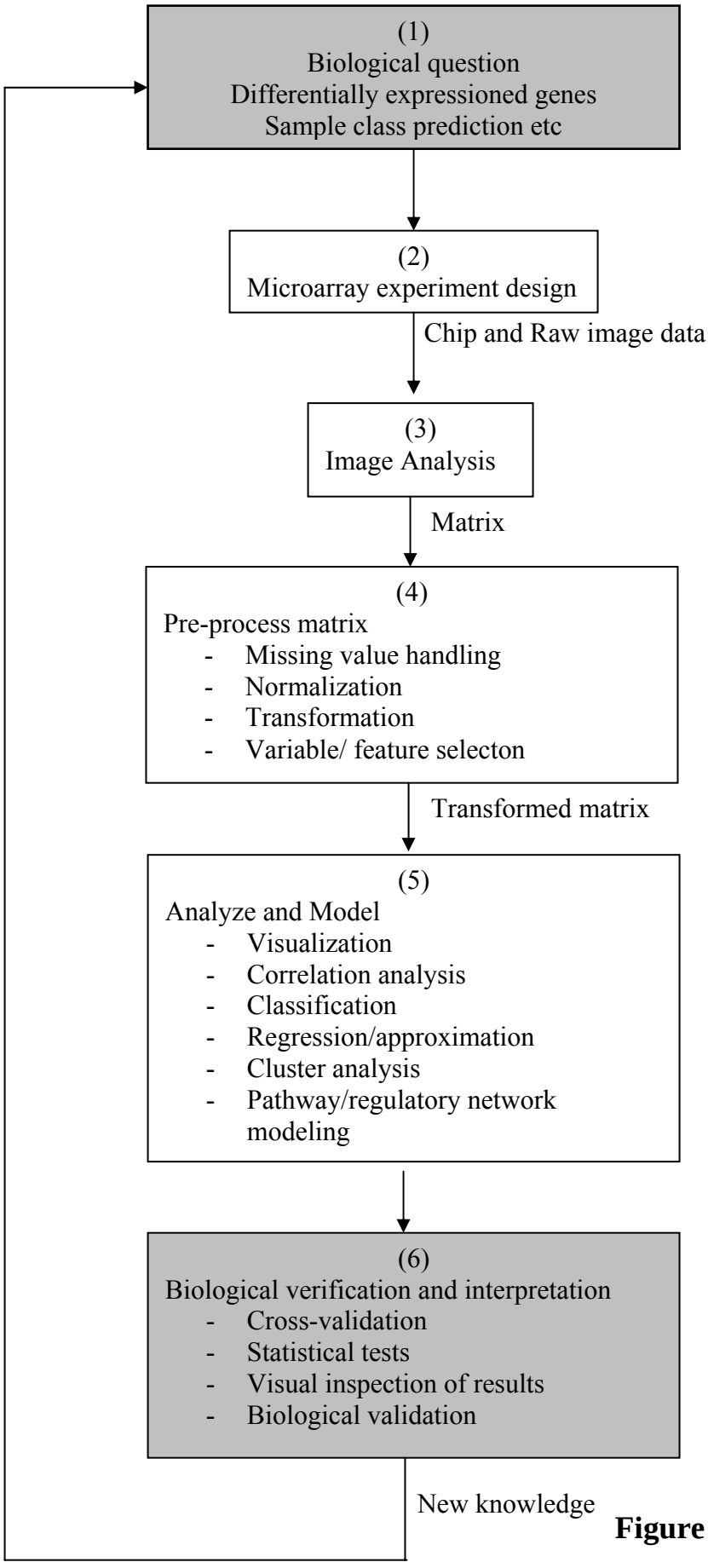

Figure 1.8: Microarray Technology

1.4.1. Pre-processing step on raw data

Arising from Step 3 of the overall analysis process is the gene expression data. The quality of gene expression data strongly depends on the equiments used, the biological variation and the measurement condition. Therefore, the gene expression data must be pre-processed with several techniques such as normalization, standardization and transformation.

For example, the single data matrix is resulted by integration all sets of measurements from each microarray. There of course exists measurement variation between arrays. A standardization procedure must be applied for this matrix to eliminate this variation and to facilitate comparison between different hybridization experiments,.

Moreover, the data matrix is highly complex for further effective and efficient performance of latter data analysis tasks. It is sometimes necessary to employ a useful step called transformation. As the result of this, the complexity of data matrix is reduced and the information is represented in more useful format.

1.4.1.1. Processing missing values

For a variety of reasons the matrix of gene expression levels are not allways filled up. Such reasons include image corruption, insufficient resolution, simply dust or scratches on the slide. In the following are several strategies dealing with missing values.

The first simple and obvious way is to remove the gene or array profiles containing the missing values. This method has a main drawback, that is, it can also remove other valuable data. In the worst case, this approach may remove all valid expression values while actually only min(n,m) missing values distributed equally in rows or columns. And of course the data left for us to analyze become little.

The second approach is to retain the missing values in the data matrix but using a special code for them. This special code is chosen so that it can be distinguished with all possible valid expression values in the data matrix. Clearly,

this approach makes sense only if the proportion of missing values does not exceed an acceptable threshold.

The third way is to replace the missing values with reasonable values. In practice, this substitution values are often chosen as a constant, the expected or standard deviation value of particular gene across all samples. For example, the missing value of gene ***b*** for patient 5 can be replaced by the the expected value of the expression levels of gene ***b*** across condition tumor A of patient 5.

Apart from three above basic approaches, there exist many other methods for processing missing values such as principal components analysis, hierarchical clustering and k-means clustering [26]. Despite being suitable to the problem of processing the missing values but they all require a complete matric computation [26]. Recently three methods: Singular Value Decomposition (SVDimpute), weighted K-nearest neighbors (KNNimpute), and row average are implemented and evaluated using a variety of parameter settings and over different real data sets. The result showed that KNNimpute appears to provide a more robust and sensitive estimator for missing value estimation than the other.[31]

1.4.1.2. Data transformation and Discretization

For data transformation step, each value in the gene expression matrix is converted to its logarithm in base two. As the result of that, we obtain a new gene expression matrix with the bell shape like distribution, a preferred and usefull one in the literature of statistical analysis.

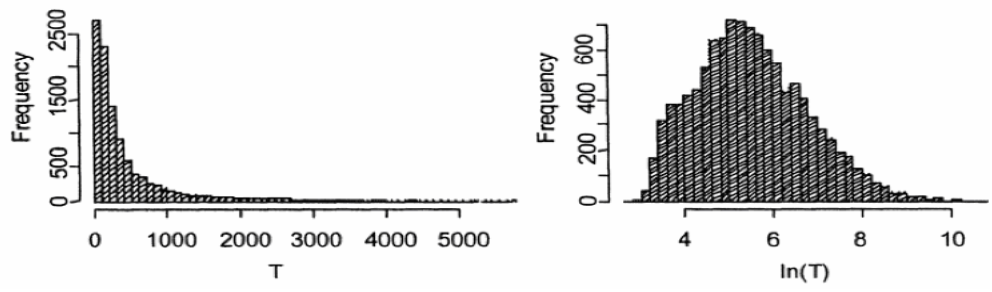

Figure 1.9: Bell shape like distribution after transformation using base-2 logarithm

Besides logarithm transformation, discretization is also a commonly used transformation method where expression level are on a continuous scale meanwhile

many analytical methods require discrete-scaled values. Such methods are Bayesian networks, association analysis, decision trees and rule-based approaches. Three labels, i.e., ***under-expressed***, ***balanced*** and ***over-expressed*** are usually used as the results of discretization for the expression values less than 1, equal to 1 and greater than 1, respectively.

1.4.1.3. Data Reduction

In most of analysis tasks later, it is often required to reduce the matrix size to improve performance of subsequent analysis. In the context of microarray data, the term variable is the one whose values are a particular gene's expression levels over all samples. And the term observation is the one whose values are expression levels of one sample across all studied genes. The following are three common data reduction strategies:

i. Variable selection select a good subset of all variables and only retain them to further analysis.

ii. Observation selection Similar to variable selection, except that observation are in role here.

iii. Variable combination find the suitable combination of existing variables into a kind of "super" or composite variable. The composite variables will be in used for further analysis while the variables used to create them not.

Variable selection is one of the most important issues in microarray analysis, because microarray analysis encounters the so-called n-large and p-small problem. That means the number of studied genes is usually much bigger than the number of samples. Moreover most of genes (variables) are uninformative. One idea is to exhaustively consider and evaluate all possible subsets and then chose the best one. However, it is infeasible in practice since there are 2^n-1 possible unique subsets of the given n genes.

Combining the relevant biological knowledge and heuristics is a simple consideration to select a subset of suitable variables. Besides consideration all subsets, one gene can be considered one by one and then be eliminated or not out of final subset based on whether it sastifies some predefined criteria such as information gain and entropy-based measure, statistical tests or interdependence analyses. In most situations, as the result of selection methods, the good set of

variables obtained may contain the correlating genes. Moreover there are some genes filtered out that only expose their meaningfullness in conjunction with other genes (variables).

Taking into account more than one genes (variables) at once, the multivariate feature selection methods such as cluster analysis techniques, and multivariate decision trees compute a correlation matrix or covariance matrix to detect redundant and correlated variables. In the covariance matrix, the variables with large values tendency tend to have large covariance scores. The correlation matrix is calculated in the same fashion but the value of elements are normalized into the interval of [-1, 1] to eleminate the above effect of large values of variables [35].

The original set of genes (variables) can be reduced by the procedure that merges the subset of highly correlated genes (variables) into one variable so that the derived set contains the mutually largely uncorrelated variables but still reserve the original information content. For example, we can replace a set of gene or array profiles highly correlated by some average profile that conveys most of the profiles' information.

Besides, the Principal Component Analysis (PCA) methods summarizing patterns of correlation, and providing the basis for predictive models is a feature-merging method commonly used to reduce microarray data [26].

1.4.1.4. Normalization

Ideally, the expression matrix contains the true level of transcript abundance in the measured gene-sample combination. However, because of naturally biased measurement condition, the measured values usually deviate from the true expression level by some amount. So we have ***measured level = truth level + error***, Where ***error*** comes from systematic tendency of the measurement instrument to detect either too low or too high values [35] and the wrong measurement. The former is called ***bias*** and the latter is called ***variance***. So ***error*** is the sum of ***bias*** and ***variance***. The ***variance*** is often normally distributed, meaning that wrong measurements in both directions are equally frequent, and that small deviations are more frequent than large ones.

Normalization is a numerical method designed to deal with measurement errors and with biological variations as follows. After the raw data is pre-processed with tranformation procedure, e.g., base-2 logarithm, the resulting matri can be normallized by multiplying each element on an array with an array-specific factor such that the mean value is the same for all arrays. Futher requirement, the array-specific factor must sastify that the mean for each array equals to 0 and the standard deviation equals 1.

1.4.2. Data analysis tasks

Right after the data pre-processing step is employed, a numerical analysis method is deployed corresponding to the scientific analysis task. The elementary tasks can be divided into two categories: prediction and pattern-detection (Figure 1.9). Due to the scope of this thesis, only two topics classification and gene regulatory network will be discussed in the following sections.

Prediction	Pattern-detection
Classification	Clustering
Regression or	Correlation analysis
Estimation	Assosiation analysis
Time-series	Deviation detection
Prediction	Visualization

Figure 1.10: Two classes of data analysis tasks for microarry data.

1.4.2.1. Classification on gene expression data

Classification is a prediction or supervised learning problem in which the data objects are assigned into one of the *k* predefined classes $\{c_1, c_2, \ldots, c_k\}$. Each data object is characterized by a set of g measurements which create the feature vector or vector of predictor variables, $X=(x_1,\ldots,x_g)$ and is associated with a dependent variable (class label), $Y=\{1,2,\ldots,k\}$. We call the classification as binary if k=2 otherwise as multi-classification. Informly a classifier *C* can be thought as a partition of the feature space X into k disjoint and exhaustive subsets, $A_1, \ldots, A_k$, containing the subset of data objects whose assigned classes are $c_1, \ldots, c_k$ respectively.

Classifiers are derived from the training set L= $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ in which each data object is known to belong to a certain class. The notation $C(.; L)$ is used to denote a classifier built from a learning set L [24]. For gene expression data, the data object is biological sample needed to be classified, features correspond to the expression measures of different genes over all samples studied and classes correspond to different types of tumors (e.g., nodal positive vs. negative breast tumors, or tumors with good vs. bad prognosis). The process of classifying tumor samples concerns with the gene selection mentioned above, i.e., the identification of marker genes that characterize different tumor classes.

For the classification problem of microarray data, one has to classify the sample profile into predefined tumor types. Each gene corresponds to a feature variable whose value domain contains all possible gene expression levels. The expression levels might be either absolute (e.g., Affymetrix oligonucleotide arrays) or relative to the expression levels of a well defined common reference sample (e.g., 2-color cDNA microarrays). The main obstade encountered during the classification of microarry data is a very large number of genes (variables) w.r.t the number of tumbor samples or the so-called “large p, small n” problem. Typical expression data contain from 5,000 to 10,000 genes for less than 100 tumor samples.

The problem of classifying the biological samples using gene expression data has becomed the key issue in cancer research. For successfullness in diagnosis and treatment cancer, we need a reliable and precise classification of tumors. Recently, many researchers have published their works on statistical aspects of classification in the context of microarray experiments [14,17]. They mainly focused on existing methods or variants derived from those. Studies to date suggest that simple methods such as K Nearest Neighbor [17] or naive Bayes classification [13,3], perform as well as more complex approaches, such as Support Vector Machines (SVMs) [14]. This section will discuss the native Bayes and *k* Nearest Neighbours methods. Finally we will describe issue of performance assessment.

The naïve Bayes classification

Suppose that the likelyhood $p_k(x)=p(x \mid Y=k)$ and class priors π_k are known for all possible class value *k*. Bayes' Theorem can be used to compute the posterior probability p(k | x) of class k given feature vector x as

$$p(k \mid x) = \frac{\pi_k p_k(x)}{\sum_{l=1}^{K} \pi_l p_l(x)}$$

The native Bayes classification predicts the class $C_B(x)$ of an object x by maximizing the posterior probability

$$C_B(x) = \arg\max_k \; p(k \mid x)$$

Depending on parametric or non-parametric estimations of p(*k*|*x*), there are two general schemes to estimate the class posterior probabilities p(*k*|*x*): density estimation and direct function estimation. In the density estimation approach, class conditional densities $P_k(x) = p(x \mid Y=k)$ (and priors π_k) are estimated separately for each class and Bayes' Theorem is applied to obtain estimates of p(*k* | *x*). The maximum likelihood discriminant rules (Fisher, 1922); learning vector quantization [18]. Bayesian belief networks [8] are examples of the density estimation. In the direct function estimation approach, posteriors p(k | x) are estimated directly based on methods such as regression technique [19]. The examples of this approach are logistic regression [19]; neural networks [19]; classification trees [20] and nearest neighbor classifiers [17].

Nearest Neighbor Classifiers

Nearest neighbor classifiers were developed by Fix and Hodges (1951). Based on a distance measurement function for pairs of samples, such as the Euclidean distance, the basic k-nearest neighbor (kNN) classifier classify a new object on the basis of the learning set. First, it finds the *k* closest samples in the learning set with the new object. Then, it predicts the class by majority vote, e.g. choose the class that is most common among those k nearest neighbors.

In kNN, the number of neighbors *k* should be chosen carefully so as to maximize the performance of the classifier. This is still a challenging problem for most cases. A common approach to overcome this problem is to select some

specific values of k and implementing classifier on the training set w.r.t each value of k. The error rate for each value of *k* is calculated based on the so-called leave-one-out cross validation fashion. The value of *k* with smallest cross-validation error rate is chosen as the best parameter for the classsifier. The nearest neighbor rule can be refined and extended to deal with unequal class priors, differential misclassification costs, and feature selection. Instead of the equal treatment among the neighbors, the refinement version introduces the weighted voting scheme for the neighbors. That is, each neighbor is assigned a weight based on its contribution to the process of deciding the new objects' class.

1.4.2.2. Feature selection

Feature selection as mentioned above is one of the most important issues in classification, especially when applied to microarray data. In the machine learning [24], feature selection can be distinguished into two categories as filter and wrapper methods. In the former, feature selection process is performed priorily to building of classifier. Some univariate test statistics such as: t- or F-statistics; ad hoc signal-to-noise statistics [17]; non-parametric Wilcoxon statistics [19]; p-values are implemented to procedure a subset of genes considered as feature set. The selection process is implemented based on the predefined number of genes or statistical test value cut-off. In wrapper methods, the feature selection is implemented as an inherent part of the classifier building procedure. Different classifiers will use the different approaches to feature selection. For example, in classification trees (CART; Breiman et al., 1984), features are selected at each step by the procedure of pruning the tree using cross-validation. In the nearest neighbor classification, the automatic feature selection can be obtained by suitable modification the distance measuring function.

1.4.2.3. Performance assessment

Obviously, different classifiers exhibit different accuracy rates. So it is necessary to develop a technique to reliably estimate of the classification error. As a result, it guarantees the best classifier will be chosen for latter implementation. This is very important to the problem of classifying tumor samples, because the misclassification could lead to misdiagnosis and assignment to improper treatment protocol.

For the purpose of performance assessment, we need a test set of labeled objects which are the samples independent from the available learning set. The classifier is applied on the test set and the classification error rate will be calculated as the proportion of test cases with discordant prediction to the true class labels. In practice, the original learning set L is randomly divided into V subsets L_v, v=1, ….,V of nearly equal size. The sets L-L_v is used as training set for building the classifiers and error rates are computed from validation sets L_v. This procedure is repeated over each subset L_v and the average error is obtained. This scheme for performance estimation is called Cross-Validation Estimation. It has a limitation that it reduces effective sample size for training purposes especially in the microarray data since the number of samples is relatively small.

1.5. Research topics on cDNA microarray data

cDNA microarray experiment is concerned with a lot of research fields. Differential Gene Expression is the first field of study, which investigates only a single gene profile to find those genes that expose different expression levels under different experimental conditions. These include different tissue types on different developmental stages of the organism. Its frequent application is in the investigation of gene expression level under normal-versus-diseased state.

The second research field is gene co-regulation study, that mainly focuses on comparision more than one gene profiles to identify genes whose expression levels are correlated in certain experiment conditions. The gene co- regulation has two basic forms: positive and negative. Two genes are called to have a positive co-regulation if their expression levels indicate the same increasing or decreasing tendency. They imply a negative co-regulation if the tendency is on the opposite direction. For example, the genes ***a*** and ***c*** in our four-gene example (Table 1.6) exhibit a negative co-regulation pattern across the ten studied samples.

The third field of study is gene function identification. Here the function of new genes can be revealed through the process of comparizing its expression profile against the profiles of genes with known functions. The prior known

functions of genes with high similarity will be used for inferring the function of the new gene.

The next topic is to study the inside cells to understande gene functions. The fourth research field, Identification of Pathways and Gene Regulatory Networks, will help to answer it. This helps biologists understand the processes by which genes and their products (i.e. proteins) play their roles in cells, tissues, and organisms. Given a pathway, the changes in expression level of investigated genes are monitored to identify pathways. Moreover, the gene expression level is also regulated by the products of other genes. This means that there exists a gene regulatory network by which a gene activities are regulated by others. Therefore, reconstructing gene regulatory network is the main goal of this field of study. These studies require time-course microarray data to be generated.

Fifthly, in clinical diagnosis, the cDNA microarry data is useful for discovering expression patterns that are characteristics of a particular disease and also useful for the inference unknown subtypes of known diseases. This is achieved by revealing characteristically different expression profiles that correlate with clinically distinct subtypes of a disease.

For each different disease already identified, suppose that we know a chemical compounds used to cure this disease. It is necessary to study the different changes in gene expression pattern in response to different dosages of this chemical compound. This is the goal that the sixth research field, dose-response study, must achieve.

The diseases are always correlated with some common variations such as insertions, deletions of a few nucleotides in sequences or in the repeated number of a motif. It is any sequence pattern that decides the molecule's function, structural feature, or family membership. The seventh research field called sequence variantion will takes advantages from cDNA microarry data to discover these variations. These common type of sequence variation are called *single nucleotide polymorphism* (SNP) [35] which occur with a frequency of roughly 1 per 1,000 nucleotides. The microarray experiments can be useful for the good analyzing SNP variation if at least three following categories in designing the microarray

experiment are qualified: (a) arrays including all known SNPs of a human genome sequence, (b) microarrays containing a sample of SNPs located across the entire human genome, and (c) devices for re-sequencing a sample of the human genomic sequence.

Finally, the cDNA microarray data is time series sometimes where the expression level of each gene are repeatedly measured after an interval of time from the same source. The eighth research field of study, time-course study, is given the birth for the purpose of processing this special temporary characteristics of time-series data. This study is also usefull in studing cell cycle phenomena and also in reconstructing the gene network.

Chapter 2

Graph based ranking algorithms with gene networks

2.1. Graph Based Ranking Algorithms

Ranking is the problem of ordering a given set such that more important elements come first in the order list. With X denoting the considered set of object, the ranking function is usually defined as a function f: X → R that assigns a ranking value to each object such that more important objects will receive higher ranking scores.

The problem of ranking has recently gained much attention in machine learning [10,11, 15,34]. Although the main form of data considered so far is vector-valued data, but sometimes there are the intrinsic realationships exist among objects in the set. The Webpages, journals and conferences, publications and scholar authors are best illustrations for this intrinsic relationships, i.e., link/citation relationships [7,21, 30]. Here each object corresponds to a node in the graph and the intrinsic relationships are described by directed edges between the nodes. This characteristic is taken into account by graph based ranking algorithms to improve the accuracy of the ranking result. Being used in the Information Retrieval systems, the graph based ranking algorithm take advantages of anchor links among webpages to generate a of webpages with the highest relevance to the interested topics required by end users from millions of pages. The graph based ranking algorithm is also used to compute the impact factor of the journals and conferences. It represents each journal or conference as a vertice. Each weighted edge between these vertices represent the total number of citations from one journal or conference to another. Apart from this, the honour ranking of paper's authors can also be calculated.

PageRank by Brin and Page [7] and Hypertext Induced Topic Selection (HITS) by Kleinberg [21,22] are the most successful graph-based ranking algorithms. For the rest of the thesis, we will use the following notations.

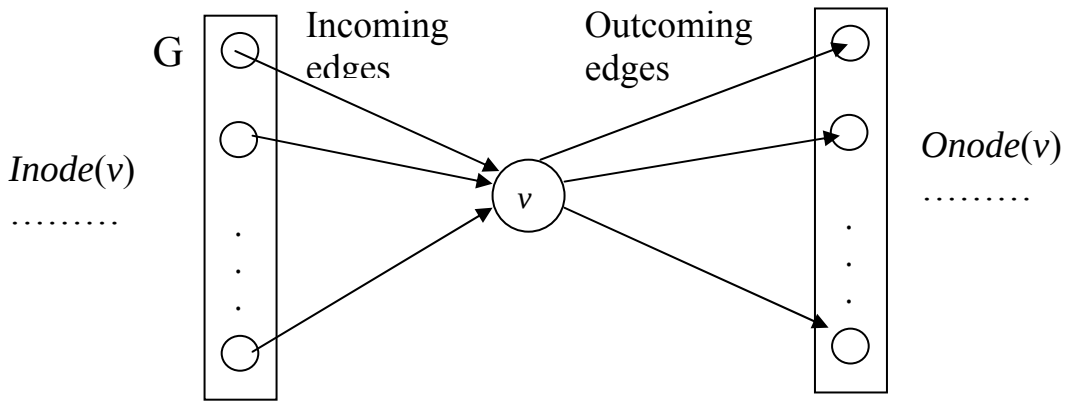

Figure 2.1. Graph, incoming and outcoming nodes

G=(V,E) denotes a directed graph with the set of vertices V and set of edges E. The *Inode*(*v*) represents the set of all nodes pointing to the node *v* and *Onode*(*v*) are the set of all nodes that the node *v* point to. (Figure 2.1)

HITS Algorithm

Originally, the HITS algorithm is applied to web documents. Here, the documents represent the vertices in the graph and the links between them represent them represent the edges. The HITS algorithm assigns each vertex v_i a so called authority score $a(v_i)$ and a hub score $h(v_i)$. The algorithm defines a good hub as a vertex with many outgoing edges, and a good authority as a vertext receiving a lot of incoming edges. Hubs and authorities exhibit a *mutual relationship*: a better hub links to many good authorities, and a better authority is pointed to by many good hubs. The authority and hub score of each vertex is recursively calculated according to the global information of all vertices in the graph rather than local one of this node as the following:

$$a(v_i) = \sum_{v_j \in Inode(v_i)} h(v_j)$$

$$h(v_i) = \sum_{v_j \in Onode(v_i)} a(v_j)$$

The authority and hub scores were proven to be converged after a number of iterations. Moreover, each edge connecting two vertices shows the different affect of the pointing node to the pointed one. If each edge is assigned a particular weight, the HITS algorithm is extended as the following equations:

$$a(v_i) = \sum_{v_j \in Inode(v_i)} w_{ji} h(v_j)$$

$$h(v_i) = \sum_{v_j \in Onode(v_i)} w_{ij} a(v_j)$$

PageRank Algorithm

Introduced by Page *et al.* [25] as the solution for ranking the webpages, the PageRank algorithm computes the importance of a page based on the following recommendation: "a page with high PageRank is a page referenced by many pages with high PageRank". Being simple and robust, this algorithm is implemented as the part of many today's commercial search engines such as Google. PageRank score of vertex can be considered as a "vote" for its importance by all the other vertext pointing to it. For the vertex v_j whose number of out links is $|Onode(v_j)|$, each of these out links will convey the vote value $\frac{PR(v_j)}{|Onode(v_j)|}$ for the vertex to which it points (Figure 2.2). The PageRank PR(.) of a page is calculated through the PR(.) of incoming nodes as follows.

$$PR(v_i) = (1-d) + d * \sum_{v_j \in Inode(v_i)} \frac{PR(v_j)}{|Onode(v_j)|}$$

Where d is a predefined parameter called the *damping factor*. The damping factor is usually set to 0.85 as the default value. Recently, some methods such as exponential damping, quadratic hyperbolic damping, general hyperbolic damping, empirical damping have been proposed for efficiently selecting a suitable value of the damping factor [4].

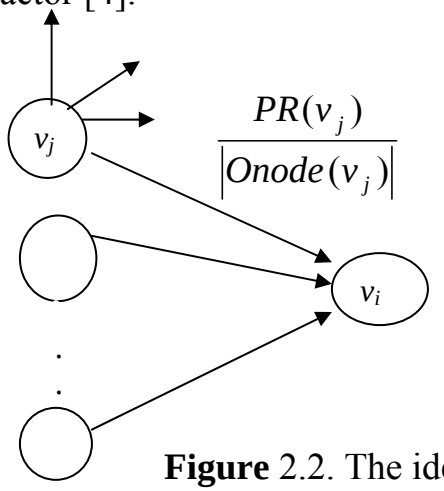

Figure 2.2. The idea behind PageRank algorithm

Starting with arbitrary initial values assigned to each node in the graph, the PR values are recursively calculated until converged. It was proven that the final values are not influenced by the initialized values but the number of iterations to reach the convergence may be changed [7].

By the definition above, the PageRank algorithm encounters some weakness. First of all, the node x will get the bigger score if there are cycles contained in the connected component whose some of its member nodes point to x. An example is illustrated in figure 2.3. In the graph on the left-hand side, node 0 gets 4 citations, wherease nodes 10 and 6 in the other two graphs receive 3 citations. However, the PageRank generates score of nodes 10 and 6 higher than that of node 0. This is because nodes 10 and 6 are parts of a cycle.

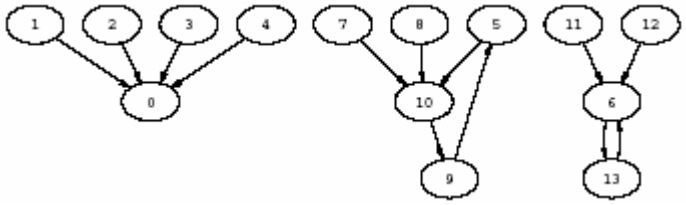

Figure 2.3: Graph toplogies in which PageRank is weak.

The second is that a node score is more influenced by the scores of the incoming nodes rather than the number of the incomingedges. In Figure 2.4, for example, although node 1 gets 6 citations while node 0 gets only one citation, node 0 still has a higher score than that of node 1 as the result of the PageRank algorithm.

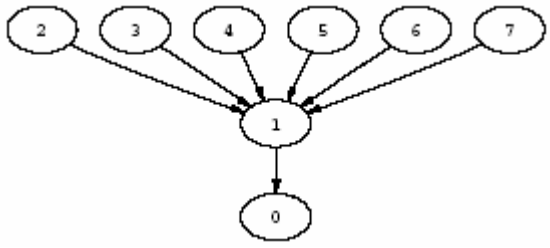

Figure 2.4: The second type of graph topology on which PageRank is weak.

2.2. Introduction to Gene Network

Understanding the regulation during the process of protein synthesis is one of the central issues in molecular biology [2]. The regulation machinery turns out to be determined by proteins that bind to regulatory regions along the DNA. So the genetic information contained on each gene is not sufficient for gaining insight into biological processes. Our understanding in the mechanism of a biological system requires the knowledge about regulatory network between genes, the so-called gene regulatory network.

A gene regulatory network shows the interaction between genes, thereby governing the rates at which genes are transcribed into mRNA. A special protein called regulatory factor binds to a particular gene and then regulates the expression level of this gene (Figure 2.5). This gene regulation is often context sensitive, e.g., gene A upregulates gene C, but only if gene B is present as well. If a large number of measurements of the gene expression levels are available, we are able to model and reverse engineer the gene regulatory network that controls their expression level.

The cDNA microarray experiments provide the data and a “genomic” viewpoint on gene expression. There exist two different types of gene expression data to reconstruct gene regulatory networks: time-series and steady-state. Various modeling techniques have been proposed for time-series and steady-state data, including Probabilistic/Boolean networks, Bayesian networks, Recurrent Neural networks and sets of differential equations. Due to a greate amount of gene expression data required, efficient computational methods are essential. The gene network reconstruction is only at its starting phase of study since 6 years [2]. It is now one of hot topics in systems biology and opens new challenges to reconstructing a large scale gene network.

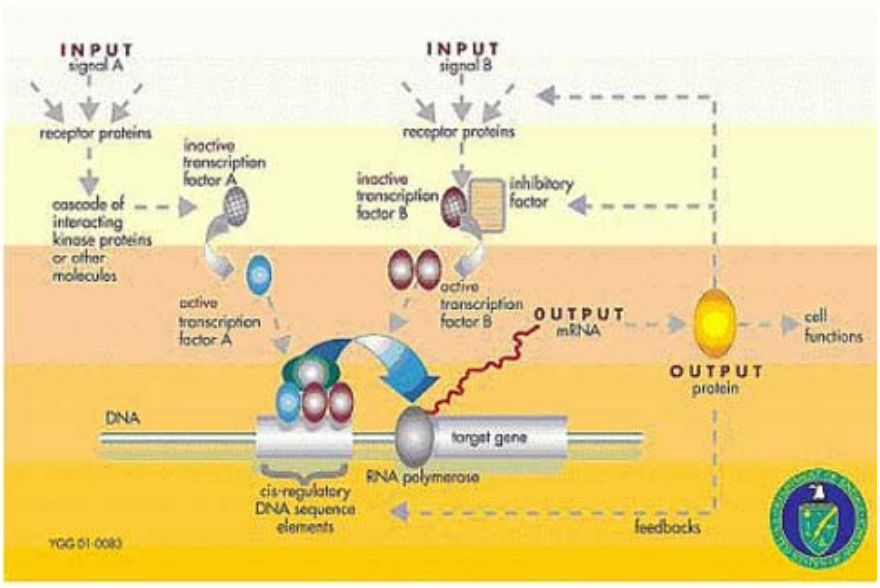

Figure 2.5: Explanation about the regulatory relationship between genes.

2.2.1. The Boolean Network Model

Originally introduced by Kauffman (Kauffman, 1969, 1974; Kauffman and Glass, 1973) the boolean network model has recently received much attention from the biology community [28,29]. In this model, each gene exists in only two state ON or OFF. The state of each gene at the next time step is determined by boolean function of the state of several genes at the current time step. The boolean network is represented as a pair (V,F) where $V = \{x_1, \ldots, x_n\}$ denotes the states of n genes and $F = (f_1, \ldots, f_n)$ is a list of Boolean function. Each state x_i may be in one of two possible values 1 or 0 corresponding to the expressed state or not respectively of gene i. Boolean networks were shown useful to understand insights in the behavior of large interconnected networks. Even though boolean networks use a binary representation for continuous domain, a variety of algorithms have been developed for infering boolean networks, Somogyi et al. [27]. Moreover, several extensions of boolean networks such as noisy boolean network were proposed in the work of Akutsu et al. [1,2].

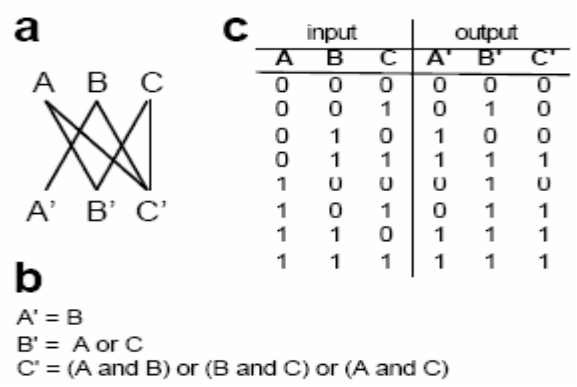

Figure 2.6: A simple Boolean Network Model

2.2.2. Probabilistic Boolean Networks

With a boolean network the state of the target gene is determined with full certainty from the state of the regulatory genes. However, there always exist uncertainty in the gene expression data. Based on boolean networks, the probabilistic boolean networks (PBN) [28,29] is capable of adapting these uncertainties not only in the data but also in the model selection. The PBN is obtained by extending the boolean network to have more than one boolean function for each node. We denote $F_i=\{ f_j \}$, $j=1,..., l_i$ as the set of all possible boolean functions and l_i as the number of such functions for gene x_i. At a given time point the PBN is characterized by a vector of boolean functions, where the i^{th} element is selected as the predictor at that time point for gene x_i. This vector is actually a standard boolean network operating at that time and called "realization" of the PBN. Here, the main problem is to learn the individual selection probabilities C_j of each element in a realization and the mechanism to calculate the probability P_i that the i^{th} boolean network or realization is selected. Shmulevich *et al.*,(2002) pointed out that P_i can be easily expressed in terms of the individual selection probabilities C_j.

2.2.3. Bayesian Networks

Based on the assumption of conditional independence, Bayesian networks (Pearl, 1988) are becoming a promising tool for analyzing gene expression patterns including reconstruction of gene regulatory network. A Bayesian network constructed for a set of random variables X is a pair $B=(G,Q)$. The first component, G, is a directed acyclic graph (DAG) in which each vertex corresponds to the

random variables $x_1, \ldots, x_n$, and each edge represents a direct dependency between two variables. The graph G sastifies the Markov assumption, that each variable x_i is independent of its nondescendants given its parents G. This allows the joint distribution to be factorized according to the conditional independence encoded in G. As the result of the factorization process, the number of required parameters are reduced. That is

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1}) = \prod_{i=1}^{n} P(X_i | \mathbf{Pa}_i)$$

where $\mathbf{P}a_i$ denotes the set of all parents of variable X_i in the graph G.

The second component Q represents the set of the network's parameters describing the conditional distribution for each variable, given its parents in G. That is $\theta_{ijk} = P(X_i = x_i^k \mid Pa_i = pa_i^j)$ for each possible state x_i^k of X_i and each configuration pa_i^j of Pa_i.

There are two main problems concerned the reconstruction of Bayes network, that are learning the Bayes Network and Inference. The former, learning the Bayesian networks, is the process of *inferring the topology for the Bayesian network that may have generated the data together with the corresponding uncertainty distribution given the data.* More specifically, there are two phases in learning Bayesian networks. The first phase is to construct the network structure as an acyclic directed graph (DAG) whereas the second is to learn parameters of that network. It is obvious that inferring a suitable structure is more important than estimating accurate parameters. The number of Bayesian network topologies grows exponentially with the number of variables. Particularly, the number of different DAGs over n nodes is given by the Robinson's formula [16]

$$G(n) = \begin{cases} 1 & \text{if } n = 0 \\ \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-1)} G(n-1) & \text{if } n > 0 \end{cases}$$

Chickering et al.(1994) showed that determining an optimal Bayesian network structure is NP-hard [8]. Therefore, the greedy search strategies like hill climbing or beam search are often used to generate the initial topology. The key idea behind the hill climbing search is to obtain the best network from the neighbors of the current network using *traversal operators*. Instead of going only in one search direction in hill climbing, the beam search strategy uses several directions, each in turn being a hill climbing search by itself. The greedy strategies have the tendency to be trapped into a local maxima and depend on the step size. Hencen, several variations have been proposed to overcome that disadvantage. For example, the background knowledge such as the whole or partial structure of gene network or an ordering among the variables can be considered as a part of the learning algorithms. The genetic algorithm is another strategy that can be applied to Bayesian network structures [16].

2.2.4. Additive regulation models

The additive regulation models [2] assume that the change in each variable over time is given as a linear weighted sum of all other variables (Figure 2.7) as follows:

$$\Delta y_i = \sum_j w_{ji} y_j + b_i$$

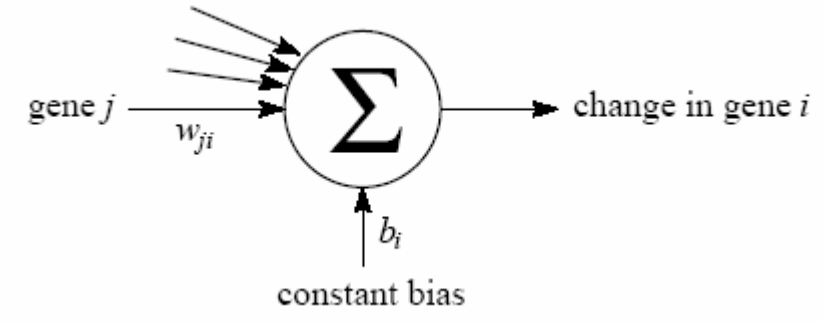

Figure 2.7: Additive regulation models

Where y_i is the level of the i^{th} variable, b_i is a bias constant indicating whether i is expressed or not in the absence of regulatory inputs, and weight w_{ji} represents the influence of variable j on variable i. For a continuous-time system we get the following differential equation:

$$\frac{dy_i}{dt} = \sum_j w_{ji} y_j + b_i$$

Since most genes exhibit a sigmoidal dose response curve, i.e., the gene activation at first increases slowly, then more rapidly, and finally saturates at a maximum level. The above formula is therefore modified with sigmoidal transfer function:

$$\frac{dy_i}{dt} = S\Big(\sum_j w_{ji} y_j + b_i \Big)$$

where S(.) is the sigmoidal function, e.g.

$$S(x) = (1 + e^{-x})^{-1},\ S(x) = \tanh(x)$$

Moreover the decay rate of gene products, D_i for gene i, plays an important role in their regulation. Hence, this factor also needs to be integrated to the model for gene regulatory network. As the result of this, the extension of the additive regulatory model becomes:

$$\frac{dy_i}{dt} = S\Big(\sum_j w_{ji} y_j + b_i \Big) - D_i y_i$$

Chapter 3

Real data analysis and discussion

3.1. The proposed scheme for gene selection in sample classifying problem

Every genetic disease is caused by sequence variant of many genes. However, among these genes, only a few of them play an important role in causing disease. For the purpose of specifying disease genes, Lude Franke *et al*.(2006) developed a functional human gene network that comprises known interaction derived from the Biomolecular Interaction Network Database (BIND), the Human Protein Reference Database (HPRD), Reactome, and the Kyoto Encyclopedia of Genes and Genomes (KEGG). After that an empirical p-value is calculated for each gene based on the gene network and then used as a score to rank all genes ralated to this type of disease.

Starting from the view point of graph based ranking algorithms, the thesis proposes a method for ranking the genes causing a specific type of disease, special characteristics of samples under a particular condition. It employs the graph-based ranking algorithms on the gene regulatory networks to rank all genes. Supposed that if gene A is regulated by gene B then gene B will be said to play more important role than gene A. Therefore, we will use the Hub score as the importance measurement of each gene. This means that the more number of genes a particular gene regulates, the more important this gene is.

In gene expression analysis, a gene is considered as a factor causing a particular types of samples, after the processing of comparision expected value of this gene in this particular class type over that in the other type. Large difference denotes gene with differential expression, it means that this gene is decisive factor. However the high expected value may be yielded from genes with the difference between expression values is high. There is a fact that with the same expected value

in different class types of samples, one gene may be informative for a particular class type while not for the other. Starting from this idea, we assume that genes whose expression values are stable over all samples belonging to a particular type. Being stably expressed is statistically represented through the variance. The smaller the variance value is, the more stable the gene is expressed. We then apply hypothesis tesing to confirm whether the variance value is less than or equal to a given small threshold value θ. This hypothesis is accepted or rejected based on the value of the sample variance S_x^2 derived from a random sample of size n with the test statistic defined by $x = (n-1)\frac{S_x^2}{\theta}$. This follows a Chi-Square distribution with n - 1 degrees of freedom. From this test, we obtain a list of all genes whose expression values are stable in all samples of each class. Furthermore, we take the intersect of the above list and the list of top-ranking genes as decisive factors causing the disease type. The remainder of this set is filled in with a gene in the set of genes with high rank score whose correlation coefficient with another one belong to set of stable genes set is greater than predefined threshold value (Figure 3.1).

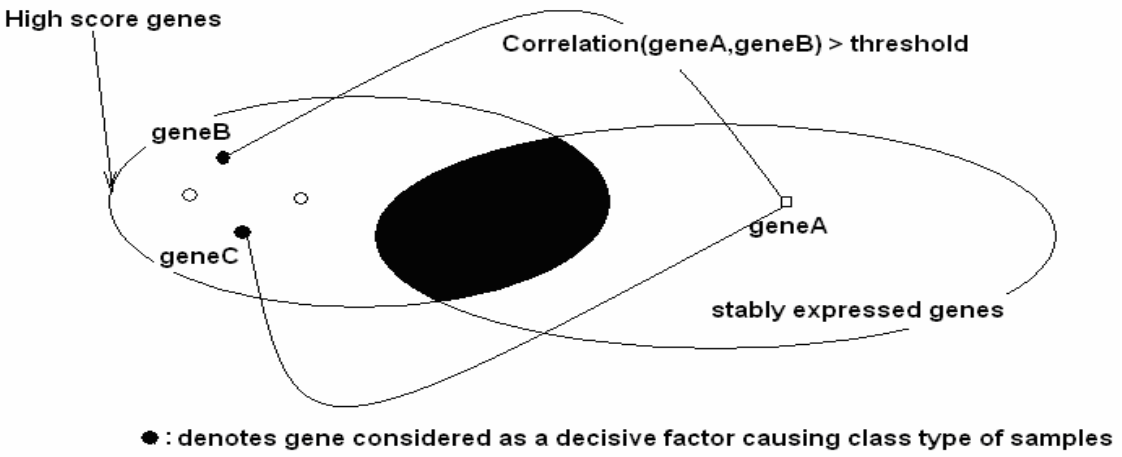

Figure 3.1: High score genes in assosiation with stability feature

For the problem of classifying the tumor samples using gene expression data, it's obvious that among a large number of genes there exist only small subset of informative genes in the determination of the sample class. Hence, the process of gene selection is needed. In the following (Figure 3.2), we explain the proposed scheme for gene selection mechanism. With each class of sample c_i (e.g., cancer type), after applying the graph-based ranking algorithm, only a predefined number

k_i of genes with highest ranking scores are selected as the representative feature set f_i of the class c_i. This is repeated over all classes of samples. Afterwards all the representative feature sets f_i are aggregated with each other in order to generate a final representative feature set. Based on the idea that genes are scored by graph based ranking algorithm as the important ones of more than one class types are likely to be not the decisive factors causing of the particular class type, the aggregation operation is chosen as the union of all complementary parts of all feature sets f_i. It's denoted as XOR operation in Table 3.1. Moreover UNION (Table 3.1) operation where the final feature set f is the union of all feature set f_i is also considered.

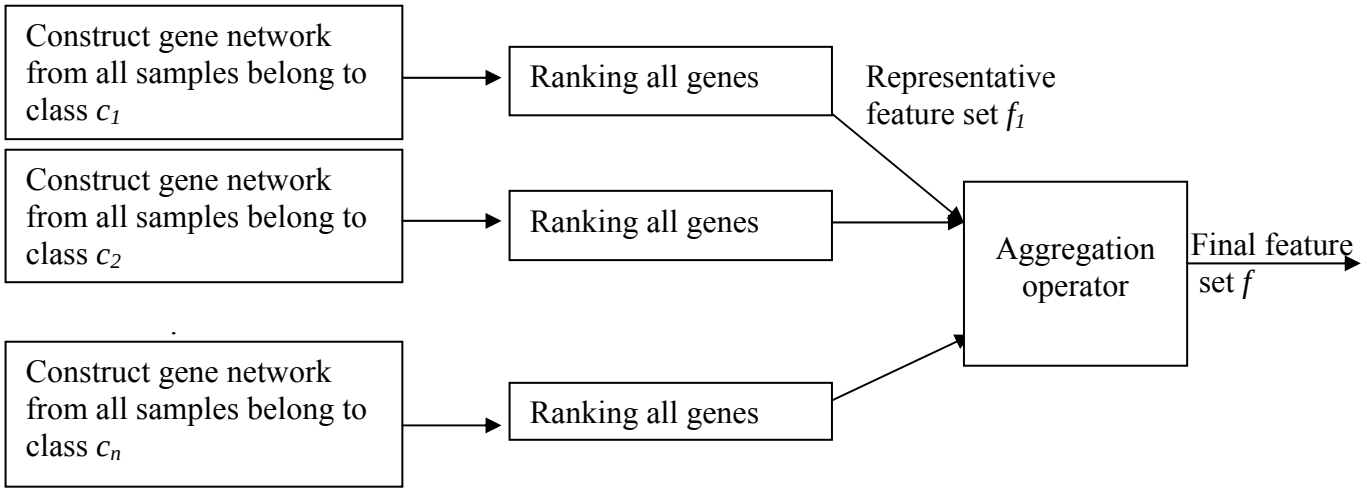

Figure 3.2. The proposed method for gene selection in sample classifying

3.2. Developing Environment

We have developed a program using C++ language. The missing values are simply pre-processed by replacement with the expected value of the others remaining expression levels in the corresponding gene profile.

It is easily realized that the problem of reconstructing the gene regulatory networks is extremely hard and time-consuming, especially when the number of is up to thousands. Recently a lot of research works are published efficiently build the gene regulatory networks. However, these methods are still restricted for maximally hundreds of genes.

Due to this difficulty, the thorough research on reconstructing the gene regulatory network is beyond the scope of this thesis. For the purpose of simplicity, the thesis builds the gene regulatory network based on the measurement of correlation between two genes. Two strategies are implemented: one for the graph-based ranking algorithm used only and another one for taking into account the stability of gene expression values with the graph-based ranking algorithm. To assess the efficiency of the proposed gene selection scheme, the kNN (k-Nearest Neighbors) classifier is used.

The dataset is the gene expression data of yeast Saccharomyces cerevisiae (http://genome-www.stanford.edu/cellcycle/links.html)from Microarray Laboratory at Standford University. They include 2467 genes over all samples belonging to two specific class types: experiment conditions CDC and ELU. The analysis randomly splits the dataset into two subset, training and testing set. The training data contains 16 Samples in which 8 samples belong to CDC class and 8 samples belong to ELU class. The testing data consist of 13 samples in which 7 samples belong to CDC class and 6 samples belong to ELU class.

The second analysis uses the benchmark gene expression dataset of two subtypes of Leukemia cancer disease, i.e., ALB and ALT for B-cells and T-cells respectively. In this dataset, 19 samples belong to class ALB and 8 samples belong to class ALT. This dataset is freely available at http://www.broad.mit.edu/ cancer/.

3.3. Analysis results

The table 3.1 is results from the analysis conducted with kNN classifier. Each value in a cell denotes the number of false positive misclassifications occurring within the test in the corresponding row using a specific value of k for kNN classifier in the corresponding column. The table shows that the proposed gene selection scheme exhibits a lower false positive misclassification number when k is greater than 10, especially when *k* equal to 15. Moreover, combining the stability of gene expression values with graph-based ranking algorithm gained even better results than the case without considering this stability computation.

Test/ geneNo	Description	K =1	K =2	K =3	K =4	K =5	K =6	K =7	K =8	K =9	K =10	K =11	K =12	K =13	K =14	K =15
Test0 2467	No gene selection	1	1	2	2	2	4	3	3	5	6	7	6	6	7	8
Test-8 195	1. Rank score 2. UNION operation	5	6	6	8	7	7	8	8	6	10	8	3	7	6	5
Test-7 393		4	6	6	5	7	5	8	5	4	8	7	9	7	7	5
Test-6 494		8	7	6	4	5	10	5	7	10	5	7	6	7	6	7
Test-5 394		8	8	5	9	9	4	6	6	8	3	4	9	9	9	6
Test-4 390		6	10	6	7	6	8	7	8	5	4	3	7	8	8	8
Test-3 913		7	7	7	6	3	8	9	7	5	4	6	8	7	8	8
Test-2 1604		6	7	7	4	6	7	7	5	7	3	6	6	8	6	9
Test-1 2093		9	6	6	6	5	4	5	6	5	3	6	7	6	8	4
Test 1 1209	1. Rank score + Variance P-value 2. XOR operation	7	6	8	7	9	4	9	4	7	6	8	6	8	7	5
Test 2 1153		7	10	9	9	9	6	5	8	1	6	8	8	7	6	8
Test 3 969		6	7	4	4	8	7	6	8	7	9	7	7	5	9	8
Test 4 852		9	7	5	8	5	9	8	8	6	10	10	5	8	2	7
Test 5 1193		7	8	5	6	8	9	5	8	5	5	8	6	7	4	8
Test 6 1257		6	8	6	7	10	8	1	7	4	6	8	8	2	8	8
Test 7 1224		3	7	7	10	5	7	7	6	4	8	7	5	9	8	6
Test 8 586		5	7	7	4	8	6	9	8	8	7	3	9	8	6	5
Test 9 896		9	3	7	6	3	7	7	8	8	6	4	8	6	8	5
Test 10 1045		5	4	6	8	7	6	9	8	5	8	9	6	9	4	6
Test 11 1172		2	2	6	6	4	9	3	9	7	8	7	6	9	6	3
Test 12 742		4	6	8	5	7	4	6	5	9	3	4	8	7	6	6
Test 13 800		8	9	8	9	10	7	6	6	5	6	6	9	6	7	6

	The number of misclassifications less than or equal to the original number of misclassifications when not using gene selection
	Greater than by 1
	Greater than by 2

Table 3.1: The experiment result of kNN classifier in partner with gene selection

A good scheme for gene selection especially makes sense in the case of cDNA microarray datasets where number of genes usually exceeds one thousand. And almost all classification algorithms require a great amount of computation when processing data objects with large number of variables (genes). As the result of proposed scheme (Table 3.1), the number of genes are significantly reduced to hundreds of genes compared to the original set of more than one thousand genes. Table 3.1 shows that even when the number of genes is below 1.000, the accuracy of the classification in some cases is still equivalent or better than that of the case when not using gene selection.

The scheme mentioned in Figure 3.2 is implemented based on the undirected network of genes. Definitely this structure does not really reflect the regulatory gene network where all edges are directed. This might be the main reason why the proposed scheme (Table 3.1) is not always improve the result but only in some case. Hence, the proposed scheme is meaningful in literature.

Apart from reducing the computational burden by decreasing in the number of genes (Golub,T.R., 1999), the proposed method also makes sense in the biological aspect. The method can be applied to find important genes causing a particular disease type. Due to the global information is considered by the graph-based algorithms, the method is especially well-fitted where the expression level of a particular gene depends on the occurrences of all other genes. Therefore, the obtained gene scores are meaningful to some certain extent. Genes with highest scores may be considered as the key factors leading to the development of cancer type in the studied samples. This will clearly help medical researchers find out an appropriate method to best prevent the affect of the diseases.

The analysis ranke 999 genes within two gene expression datasets of two subtypes of Leukemia cancer disease, that is ALB and ALT for B-cells and T-cells respectively and identified 10 genes with highest hub ranking scores in both. Within both sets of 10 genes, there is one gene in each which is observed to play an important role in causing the Leukemia cancer disease [36]. Those genes are Cytoplasmic dynein light chain 1 with accession number U32944 and NADPH-flavin reductase with accession number D26308 for ALB and ALT subtypes respectively (Figure 3.1). Although gene network characteristic is still limited, we

recovered one gene that causes the Leukemia cancer disease in the top ten of genes with highest hub ranking score. This shows the benefit of the proposed method in discovering the important genes causing the cancer type.

Highest ranking score genes for ALB disease type

1	D50310_at	4031.	4058.	3835.	3147.	3151.	3052.	5117.	4212.
2	D83776_at	417.	391.	428.	204.	101.	93.	621.	527.
3	L19437_at	990.	2016.	1912.	1021.	1187.	2073.	1651.	1829.
4	M31303_rnal_at	3828.	3780.	3563.	3330.	4212.	2290.	4751.	3347.
5	M62762_at	835.	935.	1665.	764.	1323.	1030.	1482.	1306.
6	U10323_at	3215.	3110.	3157.	3150.	4100.	1751.	3412.	3862.
7	U32944_at	3349.	1625.	3502.	1530.	1041.	277.	2069.	2952.
8	U60325_at	678.	369.	631.	267.	412.	176.	504.	771.
9	U73737_at	258.	477.	676.	453.	454.	76.	739.	910.
10	X56468_at	2689.	1894.	4091.	1545.	2011.	1521.	3147.	2927.

Highest ranking score genes for ALT disease type

1	AJ000480_at	749.	1048.	752.	934.	536.	607.	592.	482.¶
2	D14657_at	1239.	2863.	1202.	3793.	2788.	2400.	1765.	1104.¶
3	D15057_at	425.	759.	122.	825.	633.	603.	449.	9.¶
4	D23673_at	2103.	2418.	1925.	2036.	1491.	1545.	1638.	1619.¶
5	D26308_at	660.	1144.	456.	955.	721.	373.	550.	591.¶
6	D49738_at	1676.	1225.	1017.	2262.	1363.	1364.	1399.	1098.¶
7	D50310_at	6862.	5526.	3819.	8149.	3074.	6096.	3989.	3820.¶
8	D63874_at	11557.	12121.	5315.	10058.	7198.	7021.	6555.	5519.¶
9	D86970_at	594.	422.	318.	573.	135.	242.	253.	237.¶
10	L10373_at	1258.	2097.	1655.	2950.	1713.	718.	1949.	271.¶

Figure 3.1: Ten highest ranking score genes obtained from experiments

Conclusion and Future Works

In summary, the thesis has introduced all concepts concerned with the cDNA microarray technology, an interesting field of study with a lot of challenges for a long term research. The cDNA microarray technology has been shown to have a lot of applications in life sciences, especially in cancer researches.

Moreover, the thesis has proposed a new approach to realize genes that play an important role in causing cancers using gene expression data. A new scheme for gene selection process was also introduced. The efficiency and biological aspects of the method were tested and discussed in chapter 3. Not only making sense in the process of classifying tumor samples, but also the gene selection scheme is meaningful in future processing of the other machine learning techniques such as clustering, discovering pattern or model, etc … It is obvious that the number of genes after gene selection process is reduced significantly.

Due to limitation of time, only HITS algorithm is implemented to evaluated the proposed method. Page-Rank algorithm needs to be deployed in the future. With the characteristic of undirected graph, the current gene network shows similarity to a social network. Therefore, we can also employ the strategy for social networks.

Moreover, analysis result is only compared with that obtained from kNN classifier. This can definitely be extended to other classifiers such as Naïve Bayes, SVM, Decision Tree.

Finally, the technology will be developed to measure the protein abundance. The proposed method is surely to analyzed the protein expression data.

REFERENCES

[1]. Akutsu, T., Miyano, S., and Kuhara, S. Identifcation of genetic networks from a small number of gene expression pattern under the Boolean network model. In Altman et al. 16, pp. 17 - 28.

[2]. Akutsu, T., Miyano, S., and Kuhara, S. Algorithms for inferring qualitative models of biological networks. *Proc. Pacific Symposium on Biocomputing*, 290-301(2000)

[**3**]. Andrew D Keller, Michel Schummer, Walter L Ruzzo, Lee Hood, "Bayesian classification of DNA array expression data", 08-01 (Computer Science and Engineering, Univ Washington, Aug 2000).

[4]. Baeza-Yates, R., Boldi, P. and Castillo, C., Generalizing PageRank: Damping functions for link-based ranking algorithms. *In Proceedings of SIGIR*, Seattle, Washington, USA, ACM Press, August 2006.

[6]. Bengtsson, H., Introduction to cDNA Introduction to cDNA microarray analysis, *Lund University, Sweden*

[7]. Brin, S., Page, L., The Anatomy of a Large-scale Hypertextual Web Search Engine, *Proceedings 7th WWW Conference*, 107–117 (1998).

[8]. Chickering, D. M., Geiger, D., and Heckerman, D.. Learning Bayesian networks is NP-hard. *Technical Report MSR-TR-94-17, Microsoft Research*, 1994.

[9]. Chow, M.L., Moler, E.J., and Mian, I.S. Identifying marker genes in transcription profiles data using a mixture of feature relevance experts. *Physiol. Genomics*, **5:** 99-111(2001)

[10]. Crammer, K., and Singer, Y., A New Family of Online Algorithms for Category Ranking, *Proceedings of the 25rd Conference on Research and Development in Information Retrieval (SIGIR)*, 151-158 (2002). Tampere, Finland.

[11]. Crammer, K. and Singer, Y., PRanking with Ranking, *Proceedings of the Fourteenth Annual Conference on Neural Information Processing Systems* (NIPS), 641-647 (2001)

[12]. Dang Thanh Hai, Nguyen Thu Trang, Ha Quang Thuy, Graph of Concepts Based Text Summarization, *The 9th National Conference on Information Technology of Vietnam*, 6/2006

[13]. Dang Thanh Hai, Nguyen Huong Giang, Ha Quang Thuy, Naive Bayes text classification algorithm and problem of specifying clasifying threshold in search engine, *Journal of Computer Science And Cybernetics* 21(2):, 152-161 (2005)

[14]. Dudoit, S. Fridlyand, J.& Speed, T. P. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data *J. Am. Stat. Assoc.* 97: 77–87(2002).

[15]. Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research* 4: 933–969(2003).

[16]. Friedman, N. and Koller, D. Being bayesian about network structure, in C.Boutilier and M.Godszmidt (eds), *Uncertainty in Articial Intelligence, Morgan Kaufmann Publishers*, 201-210 (2000).

[17]. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov, J.P., Coller,H., Loh,M.L., Downing,J. Caligiuri,M. A Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286: 531–537.

[**18**]. Hollmén,J., Tresp, V. and Simula, O., A learning vector quantization algorithm for probabilistic models. In Proceedings of EUSIPCO 2000 - X European Signal Processing Conference ,Volume II, 721 –724.

[**19**]. http://en.wikipedia.com/wiki

[**20**]. http://www.statsoft.com/textbook/stclatre.html

[21]. Kleinberg, J., Authoritative Sources in a Hyperlinked Environment, *Journal of the ACM* 46 (5): 604–632 (1999).

[22]. Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A., The Web as a Graph: Measurements, Models, and Methods, *Proceedings 5th COCOON Conference*, 1–17 (1999).

[23]. Lempel, R., Moran, S., SALSA: the Stochastic Approach for Link-structure Analysis, *ACM Transactions on Information Systems* 19 (2) 131–160 (2001).

[**24**]. Machine Learning, Tom Mitchell, McGraw Hill, 1997.

[25]. Page, L., Brin, S., Motwani, R. and Winograd, T., The PageRank citation ranking: bringing order to the Web. *Tech. report, Stanford University*, 1998.

[26]. Raychaudhuri, S., Stuart JM, Altman RB, Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.*, 455–466 (2000).

[27]. Somogyi, R., Fuhrman, S., Askenazi, M., and Wuensche, A. The gene expression matrix: towards the extraction of genetic network architectures. *Proceedings of the Second World Congress of Nonlinear Analysts (WCNA96),* vol. 30 of Nonlinear Analysis, Pergamon Press (1996).

[28]. Shmulevich, E.R. Dougherty, and W. Zhang, From Boolean to probabilistic Boolean networks as models of genetic regulatory networks, *Proceedings of the IEEE*, 90 (11): 1778-1792 (2002).

[29]. Shmulevich, E., Dougherty, R., Kim, S., Zhang, W., Probabilistic Boolean Networks: A Rule-based Uncertainty Model for Gene Regulatory Networks, *Bioinformatics*, 18 (2): 261-274 (2002).

[30]. Sidiropoulos A., Manolopoulos Y., Generalized Comparison of Graph-based Ranking Algorithms for Publications and Authors, *Journal for Systems and Software*, 79 (12): 1679-1700 (2006)

[31]. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T.,Tibshirani, R., Botstein, D., and Altman, R.B., Missing value estimation methods for DNA microarrays, *Bioinformatics*, 17(6): 520-525 (2001)

[32]. van Dijk, S., Thierens, D. and van der Gaag, L. C. Building a GA from design principles for learning Bayesian networks, *Proceedings of the Genetic and Evolutionary Computation Conference*, volume 2723 of *Lecture Notes in Computer Science*, 2003.

[33]. von Haeseler Arndt, *Introduction to Bioinformatics Course***,** *Lecture Notes***,** Hanoi, 10-2005**.**

[34]. William W. Cohen, Robert E. Schapire, Yoram Singer, Learning to Order Things. *Advances in Neural Information Processing Systems*, 1999

[35]. Werner Dubitzky, Martin Granzow, C. Stephen Downes, Daniel Berrar, "*Pratical approach to Microarray Analysis*", Oxford University Press, 1999

[36]. Zhong Guan and Hongyu Zhao**,** A semiparametric approach for marker gene selection based on gene expression data, *Bioinformatics* 21(4):529-536 (2005)