BAYESIAN NETWORK BASED PATHWAY ANALYSIS OF HIGH THROUGHPUT BIOLOGICAL DATA

by

Melike Korucuoğlu

B.S., Mathematical Engineering, Işık University, 2010

B.S., Computer Engineering, Işık University, 2011

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2013

BAYESIAN NETWORK BASED PATHWAY ANALYSIS OF HIGH THROUGHPUT BIOLOGICAL DATA

APPROVED BY:

Assist. Prof. Arzucan Özgür (Thesis Supervisor)	.....
Assist. Prof. Hasan H. Otu (Thesis Co-supervisor)	.....
Assoc. Prof. Haluk Bingöl	.....
Assoc. Prof. A. Taylan Cemgil	.....
Prof. Uğur Sezerman	.....

DATE OF APPROVAL: 02.09.2013

ACKNOWLEDGEMENTS

First and foremost, I would like to express my very great appreciation to my thesis supervisors Assist. Prof. Hasan H. Otu and Assist. Prof. Arzucan Özgür for their patient guidance and continuous support. They inspired me greatly to work in this project.

I would like to thank to Assoc. Prof. Haluk Bingöl, Assoc. Prof. Taylan Cemgil and Prof. Uğur Sezerman for accepting to be in my thesis committee and provide valuable suggestions.

I am deeply indebted to Şenol İşçi for the help and support. I would like to offer my special thanks to Yavuz Mester to encourage me all the way.

I would also like to thank The Scientific and Technological Research Council of Turkey (TUBITAK) supporting this thesis study with the research project number 111E042.

Most of all, I feel myself lucky to have such a great family. They understand and support me all the time. Words cannot express my gratitude to my mother and sisters.

ABSTRACT

BAYESIAN NETWORK BASED PATHWAY ANALYSIS OF HIGH THROUGHPUT BIOLOGICAL DATA

Biological data production has been increasing at an unprecedented pace with the advancements of microarrays and next-generation sequencing technologies. Such High Throughput Biological Data (HTBD), requires detailed analysis methods. From a life science perspective, data analysis results make most sense when interpreted within the context of biological pathways. Bayesian Networks (BNs) capture both linear and nonlinear interactions, and handle stochastic events in a probabilistic framework accounting for noise. These properties make BNs excellent candidates for HTBD analysis. A recent study by Isci *et al.* [1] proposes an approach, called Bayesian Pathway Analysis (BPA), for analyzing HTBD using BNs in which known biological pathways are modeled as BNs and pathways that best explain the given HTBD are found. In this thesis, we have the following two fundamental aims. Our first aim is to improve the BPA system. In the data processing phase, fold changes between two groups (i.e., cancer and normal) were calculated for genes and discretized using hard cut-off levels to be used in the network scoring module. We evaluated six different discretization methods with various numbers of levels. In the scoring phase, we applied three scoring methods and compared the results with the Bayesian-Dirichlet Equivalent scheme currently applied in the system. The statistical significance assessment phase was improved by obtaining randomized data sets at the gene signal level to overcome the cases where the current BPA fails to provide random data sets. We provide a web portal where the optimized software can be downloaded and used for various organisms including human. Our second aim is to apply the improved pathway analysis approach on various real cancer microarray data sets in order to investigate the pathways that are commonly and differently active. We compared our findings with a comparable approach, SPIA [2].

ÖZET

YÜKSEK ÇIKTILI BİYOLOJİK VERİLERİN BAYES AĞLARI KULLANILARAK PATİKA DÜZLEMİNDE ANALİZİ

Biyolojik veri üretimi, mikrodiziler ve yeni nesil sekanslama teknolojilerinin oluşumu ile baş döndürücü bir hızla artmaktadır. Yüksek Çıktılı Biyolojik Veri (YÇBV) olarak adlandırılan bu sonuçlar kapsamlı analiz metotlarına ihtiyaç duymaktadır. Yaşam bilimleri penceresinden bakılınca veri analizi sonuçları en fazla biyolojik patikalar düzleminde yorumlanınca fayda sağlamaktadır. Bayes Ağları (BA) hem doğrusal hem de doğrusal olmayan ilişkileri modelleyebilmekte ve stokastik olayları olasılıksal bir çerçevede gürültüye tolere ederek inceleyebilmektedir. Bu özellikler BA'ları YÇBV analizine uygun bir yöntem kılmaktadır. Isci *v.d.* [1] tarafından yapılan yakın tarihli bir çalışma, Bayes Patika Analizi (BPA), YÇBV'yi BA kullanarak analiz eden bir yaklaşım önermektedir. Biyolojik patikalar BA olarak modellenip YÇBV'yi en iyi açıklayan patikalar bulunmaktadır. Bu tezin iki temel amacı vardır. Birinci amaç, BPA sistemini geliştirmektir. Veri işleme aşamasında, her gen için iki grup (kanser ve normal) arasındaki ifade değişim oranı hesaplanmakta ve ağ skorlama modülünde kullanılmak üzere sert eşik seviyeleri ile ayrıklaştırılmaktaydı. Buna ek olarak, çeşitli seviyelerde altı farklı ayrıklaştırma metodu denedik. Skorlama aşamasında, üç farklı skorlama yöntemi uygulayıp Bayes-Dirichlet eş yöntemiyle mukayese ettik. İstatistiksel belirginliği ölçme aşaması, rastsallaştırılmış veri kümelerini gen sinyal değerleri seviyesinde elde ederek bu konuda mevcut BPA yaklaşımının başarısız olduğu durumların üstesinden gelmek için geliştirdik. Optimize edilmiş yazılımın indirilip insan dahil çeşitli organizmalara uygulanabilmesi için web erişimi sağladık. İkinci amacımız, geliştirilmiş patika analizi yaklaşımını çeşitli gerçek kanser mikrodizi verilerine uygulayıp aktif patikaları belirlemektir. Sonuçlarımızı kıyaslanabilir bir yaklaşım olan SPIA [2] ile karşılaştırdık.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

a_{ijk}	Dirichlet distribution hyper-parameters
C_1	Number of samples in the first group in a dataset
C_2	Number of samples in the second group in a dataset
c	Constant
D	Dataset
d	Number of parameters in BN model
E	Expression Data
G	Graph
g_{i1j}	Values of the i^{th} node in j^{th} sample in G
g_{i2k}	Values of the i^{th} node in k^{th} sample in G
I	Indicator function in bootstrapping
k	Number of intervals used in discretization
M_{ij}	Count of parents' j configuration cases of node i
N	Number of nodes in BN model
N_{ij}	Sum of Dirichlet distribution hyper-parameters a_{ijk}
O	An ordered set of observations
o_l	The l^{th} element of O
$P(D)$	Probability of observed data
$P(G)$	Prior probability of the graph
q_i	Number of different states of node's parents
R	Randomized data sets in bootstrapping
r_i	Set of values a node can take on
S_n	BDe score calculated for n^{th} BN in bootstrapping
s_{ijk}	Total count for node i is observed
X_i	Random variable in a Bayesian network
θ_i	Parameters of the local probability distribution

LIST OF ACRONYMS/ABBREVIATIONS

AIC	Akaike Information Criterion
BDe	Bayesian Dirichlet Equivalent
BIC	Bayesian Information Criterion
BN	Bayesian Network
BPA	Bayesian Pathway Analysis
CAM	Cell Adhesion Molecule
CO	Cut-off Values
CPD	Conditional Probability Distribution
CPT	Conditional Probability Table
DAG	Directed Acyclic Graph
DNA	Deoxyribonucleic Acid
EFD	Equal Frequency Discretization
EWD	Equal Width Discretization
FC	Fold Change
FDR	False Discovery Rate
fNML	Factorized Normalized Maximum Likelihood
GEO	Gene Expression Omnibus
GI	Gene Interaction
GO	Gene Ontology
GRN	Gene Regulatory Network
GSA	Gene Set Analysis
GSEA	Gene Set Enrichment
GSEA	Gene Set Enrichment Analysis
HTBD	High Throughput Biological Data
IGA	Individual Gene Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAS	Microarray Analysis Suite
ML	Maximum Likelihood

MM	Mismatch
MRF	Markov Random Field
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
NML	Normalized Maximum Likelihood
PM	Perfect Match
RNA	Ribonucleic Acid
rRNA	ribosomal RNA
SNP	Single Nucleotide Polymorphism
SPIA	Signaling Pathway Impact Analysis
SynTReN	Synthetic Transcriptional Regulatory Networks
TCA	Tricarboxylic acid
tRNA	transfer RNA

1. INTRODUCTION

Bayesian Network (BN) models have gained popularity for learning biological pathways from microarray gene expression data [7,8]. BNs represent dependency structure for a set of random variables using directed acyclic graphs and have been used with increasing popularity in mathematics and computational sciences over the past 20 years. BNs model both linear and non-linear interactions between sets of random variables, handle stochastic events in a probabilistic framework accounting for noise, and emphasize only strong relations in noisy data. These properties make BNs excellent candidates for HTBD analysis, where nodes represent genes and edges represent interaction relations between genes. However, current BN applications are limited to structure learning using observed data and therefore work only on a few hundreds of data points as structure learning algorithms are computationally complex. This, in turn, results in inefficient use of HTBD, which contain a much larger number of data points. Furthermore, BNs are able to focus on local interactions, where each node is directly affected by a relatively small number of nodes, and interactions defined by a BN can be related to causal inference [7]. These properties are similarly observed in biological networks justifying the use of BNs in exploring pathways in the setting of gene interaction networks, using HTBD.

From life sciences perspective, data analysis results make most sense when interpreted within the context of biological networks and pathways. Previously established individual gene analysis (IGA) based methods have been extended to network and pathway scale, which has led to establishment of a discipline called systems biology that governs a holistic analysis approach specifically analyzing data within the context of networks. This alternative approach, called pathway analysis, functional enrichment analysis or gene set analysis (GSA) [9], which focuses on determining predefined gene sets or classes that are significantly regulated, has received a great deal of attention. GSA methods score groups of genes and can identify genes that exhibit subtle changes at an individual level, but show concordant enrichment within a set [10].

Gene Set Enrichment (GSE) or Gen Ontology (GO) based approaches that analyze microarray data within the context of pathways or functional groups consider the genes in a pathway or group as a list, calculate some sort of a score for each list representing the pathway's or group's significance without involving in their model the topology via which genes in a given pathway or group interact with each other. There have been methods proposed to take into account the GO graph topology [11], overlap between GO categories in the GO hierarchy [12] or modeling interactions between GO categories [13] in assessing the significance of enrichment of a GO term based on experimental data. However, none of these methods takes into account the network or structure defining the relations between the genes in each category.

A recent paper by Isci *et al.* [1] proposes an approach, called Bayesian Pathway Analysis (BPA), for analyzing HTBD using BNs. They describe a comprehensive study of pathway analysis of high-throughput biological data within a Bayesian Network framework and introduce an algorithm that models biological pathways as BNs and identifies pathways that best explain a given HTBD by scoring the fitness of each network. In addition to the BPA, GSEA, and GO based methods; more recent studies include the following. Sales *et al.* [14] presented a package that gathers pathway information from four different databases and resulting networks represent a uniform resource for pathway analysis. Edwards *et al.* [15] proposed a method that can be seen as an extension of conventional methods of gene expression analysis that takes network structure into account. Another method that Li *et al.* [16] performed is distance-based gene set analysis assisting in finding functional gene sets. Wang *et al.* [17] extended the network-based principles and presented a method to functionally characterize whole sets of genes and identify associations of gene sets. Mieczkowski *et al.* [18] proposed a method for detecting deregulated signaling pathways using microarray gene expression data, which incorporates information regarding pathway topology, as well as data on the position of every gene in each pathway. Martini *et al.* [19] developed an algorithm for pathway analysis to identify the signal paths, within a pathway that is mostly involved in a particular biological problem. Finally, Drier *et al.* [20] introduced an algorithm that infers pathway deregulation scores for each tumor sample on the basis of expression data.

In the BPA approach, pathways are retrieved from the KEGG database [21]. Each entry (node) in the pathway is mapped to an internal unique ID and input gene expression data with IDs that belong to NCBI's GenBank, Unigene, Gene, Refseq databases, SwissProt or Affymetrix are accepted by the system. A conversion module carries out the necessary mapping between the input gene expression IDs and pathway node IDs. Repeating entries in the pathway are merged and represented as a single node while conserving edge relations. BN theory utilizes Directed Acyclic Graphs (DAG) but there may exist cycles in the biological pathways. This is overcome using Spirtes' method [22] where graph representation of structural equation models are converted to collapsed acyclic graphs such that d-separations in the collapsed graph entail same independency relations defined by the model. To this end, a biological pathway is modeled as a BN, which now can be tested against input data to assess its fitness.

BPA assumes normalized gene expression data as input. First, IDs used in the array platform corresponding to a given node in the pathway representation are pooled and one representative signal value per node is calculated using one-step Tukey's bi-weight algorithm [23]. This approach is robust to outliers and performs a weighted average of input signal values. Currently BPA addresses experimental designs consisting of two groups of samples (e.g. cancer vs. normal). The observation matrix to be used scoring each DAG is obtained by generating the FC (Fold Change) values for each pair of samples in the two groups. In this matrix, columns represent genes in the DAG and rows represent pairwise comparisons. If there are G_1 and G_2 samples in the two groups, the observation matrix consists of G_1xG_2 rows. Each column represents the FC for the corresponding gene in each of the G_1xG_2 pairwise comparisons. These continuous FC values are discretized using a cut-off of 2. If the FC value is greater than 2 or less than 0.5 (i.e. the gene is deregulated), it is converted into 1, and otherwise it is converted into 2.

The degree to which a pathway explains given HTBD is measured using Bayesian Dirichlet equivalent (BDe) score with the equivalent sample size method [24]. In this phase, the BN is updated with the observation matrix during the score calculation. Statistical significance of this measurement is assessed by testing it against datasets

generated by applying randomization via bootstrapping where the observed score is ranked against scores obtained from randomized data sets. Bootstrapping is applied to the columns of the observation matrix. Therefore, a randomization of the rows, which are used in scoring, is achieved. The results are evaluated in terms of nominal p-values and false discovery rate (FDR) values correcting for multiple hypotheses testing.

Despite its novel contributions, there are certain areas that need improvement in the BPA approach:

- *Data Preprocessing and Discretization*: The discretization method utilized in BPA uses hard and arbitrary cut-offs. An automated method, which is optimized through simulations, is more desirable.
- *Scoring Method*: In addition to the BDe scoring method, newly established non-parametric and other well-known metrics can be used to improve the overall performance.
- *Significance Assessment*: BPA's randomization via bootstrapping method fails to generate randomized observation matrices for cases where a given column consists of one particular level. Randomization is also hampered even when more than one level is observed in a given column but the distribution is skewed. Moreover, a randomization at the signal level, rather than discretized levels, is more desirable.
- *Web Application*: A website with improved features where the user can choose among the aforementioned improvements would enable the scientific community to use the proposed algorithm more easily and effectively. BPA, in its original form, only handles human microarray data and pathways. An extension to other organisms is needed.

1.1. Objective

In this thesis, we have two fundamental aims. Our first aim to improve on the BPA system by using the following strategies. During the discretization phase, we tried Equal Width, Equal Frequency, K-means, Column K-means, Bi-directional K-means, and Automatic Threshold Discretization [25, 26] in addition to the hard-cut-off levels offered by BPA. In the proposed discretization schemes, we tried various numbers of levels as the discretized output. In the scoring phase, we applied Akaike Information Criterion (AIC) [27], Bayesian Information Criterion (BIC) [28], and Factorized Normalized Maximum Likelihood (fNML) [29] and compared the results with the BDe scoring scheme. The significance assessment phase was changed so that random data sets were obtained at the gene signal level. In this approach [30], samples in each of the two classes are randomly permuted to provide new data sets. Each new data set (with new class assignments for each sample) is subjected to the complete workflow and a score value is calculated. This way, we overcome the cases where the current BPA approach fails to provide randomized data sets. In testing these new approaches, we generated synthetic microarray data that simulates gene expression from N pathways where a subset, Na, of these pathways is active. A performance criterion is assessed by the accuracy of predicting active and passive pathways. Finally, we provided a web portal where the optimized software can be downloaded and used human and other organisms as well. A tutorial and examples of how to use the system is also provided.

Our second aim is to apply the improved pathway analysis approach on real cancer data sets. For this purpose, we downloaded real microarray data sets from the NCBI's GEO database regarding bladder, brain, breast, colon, liver, lung, ovarian and thyroid cancers. We investigated the pathways that are commonly and differently identified as active in these various cancer microarray data sets. We compared our findings with a comparable approach, SPIA [2].

1.2. Road Map

In Chapter 2 of this thesis, we provide a brief introduction to gene regulation and microarray technology. Chapter 3 serves as a primer for Bayesian Networks as required by the current approach and Chapter 4 summarizes background work and details of the BPA approach. In Chapter 5, we explain the improvements that we propose for the BPA system in detail. We present our results and conclusive remarks in Chapter 6 and 7, respectively.

2. GENES AND MICROARRAYS

2.1. Genes

The gene is the basic physical and functional unit of heredity. It consists of a specific sequence of nucleotides at a given position on a given chromosome that codes for a specific protein (or, in some cases, an RNA molecule). Genes consist of three types of nucleotide sequence:

- coding regions, called exons, which specify a sequence of amino acids
- non-coding regions, called introns, which do not specify amino acids
- regulatory sequences, which play a role in determining when and where the protein is produced (as well as its amount)

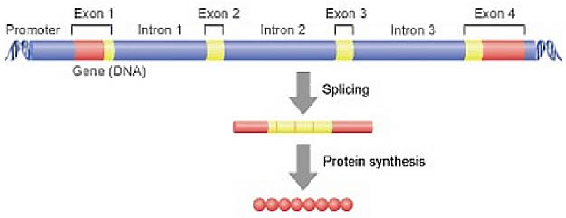


Figure 2.1. The structural components of a gene [3].

The 25,000 genes scattered throughout the human chromosomes comprise only about 3 percent of the total genome. These genes hold information critical to all human life. While all the component bases in a gene are copied as information leaves in the nucleus, not all of this information is kept. This is because within a gene there are both coding and noncoding stretches of bases. For example, in split genes, coding sections called exons supply the genetic instructions that are copied to direct protein building. These sections are preserved, but other noncoding sections within the gene, called introns, are rapidly removed and degraded. Close to each gene is a "regulatory" sequence of DNA, which is able to turn the gene on or off. Farther away, there are enhancer regions, which can speed up a gene's activity. The massive DNA molecules known as chromosomes also have many noncoding regions located outside the genes.

These contain large stretches of repetitive sequences. Some of the sequences in these locations are involved in the regulation of gene expression, and others simply act as spacers. Still other regions have functions as yet undiscovered.

A human being has 20,000 to 25,000 genes located on 46 chromosomes (23 pairs). These genes are known, collectively, as the human genome.

2.1.1. Genes to mRNA to Protein

When a gene switches on, it eventually makes a protein, but it does not do so directly. First, the gene codes an intermediary molecule called messenger RNA (mRNA). To transfer a gene's information from DNA to mRNA, base pairing is used. However, there is one change: An adenine base (A) in the DNA matches with a new base called uracil (U) in the mRNA. This difference helps to distinguish mRNA from DNA.

mRNA travels from the nucleus into the cytoplasm to cell organelles called ribosomes. There it directs the assembly of amino acids that fold into a unique protein.

Gene Expression: the process by which the genetic code - the nucleotide sequence - of a gene is used to direct protein synthesis and produce the structures of the cell.

Genes that code for amino acid sequences are known as structural genes. The process of gene expression involves two main stages: transcription and translation.

Transcription: the production of mRNA by the enzyme RNA polymerase, and the processing of the resulting mRNA molecule.

Transcription is the process of RNA synthesis, controlled by the interaction of promoters and enhancers. Several different types of RNA are produced, including mRNA, which specifies the sequence of amino acids in the protein product, plus transfer

RNA (tRNA) and ribosomal RNA (rRNA), which play roles in the translation process.

Translation: the use of mRNA to direct protein synthesis, and the subsequent post-translational processing of the protein molecule.

In translation, the mature mRNA molecule is used as a template to assemble a series of amino acids to produce a polypeptide with a specific amino acid sequence. The complex in the cytoplasm at which this occurs is called a ribosome. Ribosomes are a mixture of ribosomal proteins and ribosomal RNA (rRNA), and consist of a large subunit and a small subunit.

Some genes are responsible for the production of other forms of RNA that play roles in translation, including tRNA and rRNA.

A structural gene involves a number of different components:

- *Exons:* Exons code for amino acids and collectively determine the amino acid sequence of the protein product. It is these portions of the gene that are represented in final mature mRNA molecule.
- *Introns:* Introns are portions of the gene that do not code for amino acids, and are removed (spliced) from the mRNA molecule before translation.
- *Gene control regions*

  (i) *Start site:* A start site for transcription.

  (ii) *Promoter:* A region a few hundred nucleotides upstream of the gene (toward the 5' end). It is not transcribed into mRNA, but plays a role in controlling the transcription of the gene. Transcription factors bind to specific nucleotide sequences in the promoter region and assist in the binding of RNA polymerases.

  (iii) *Enhancers:* Some transcription factors (called activators) bind to regions called enhancers that increase the rate of transcription. These sites may be thousands of nucleotides away from the coding sequences or within an

intron. Some enhancers are conditional and only effective in the presence of other factors as well as transcription factors.

(iv) *Silencers:* Some transcription factors (called repressors) bind to regions called silencers that depress the rate of transcription.

2.1.2. Gene Regulation

Gene regulation is a label for the cellular processes that control the rate and manner of gene expression. A complex set of interactions between genes, RNA molecules, proteins (including transcription factors) and other components of the expression system determine when and where specific genes are activated and the amount of protein or RNA product produced.

Some genes are expressed continuously, as they produce proteins involved in basic metabolic functions; some genes are expressed as part of the process of cell differentiation; and some genes are expressed as a result of cell differentiation.

Mechanisms of gene regulation include:

- Regulating the rate of transcription, which is the most economical method of regulation.
- Regulating the processing of RNA molecules, including alternative splicing to produce more than one protein product from a single gene.
- Regulating the stability of mRNA molecules.
- Regulating the rate of translation.

Transcription factors are proteins that play a role in regulating the transcription of genes by binding to specific regulatory nucleotide sequences.

2.2. Microarrays

Molecular Biology research evolves through the development of the technologies used for carrying them out. It is not possible to perform research on a large number of genes using traditional methods. DNA Microarrays is one such technology, which enables the researchers to investigate and address issues that were once thought to be non-traceable. One can analyze the expression of many genes in a single reaction quickly and in an efficient manner. DNA Microarray technology has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of anomalies occurring in the functioning of the human body.

A typical microarray experiment involves the hybridization of an mRNA molecule to the DNA template from which it is originated. Many DNA samples are used to construct an array. The amount of mRNA bound to each site on the array indicates the expression level of the various genes. This number may run in thousands. All the data is collected and a profile is generated for gene expression in the cell.

2.2.1. Microarray Technique

An array is an orderly arrangement of samples where matching of known and unknown DNA samples is done based on base pairing rules. An array experiment makes use of common assay systems such as microplates or standard blotting membranes. The sample spot sizes are typically less than 200 microns in diameter and a typical microarray usually contains thousands of spots.

The intensity and color of each spot on the microarray image represent the relative expression level of that gene expressed in the test sample compared to the control sample. Thousands of spotted samples known as probes (with known identity) are immobilized on a solid support (a microscope glass slides or silicon chips or nylon membrane). The spots can be DNA, cDNA, or oligonucleotides. These are used to determine complementary binding of the unknown sequences thus allowing parallel

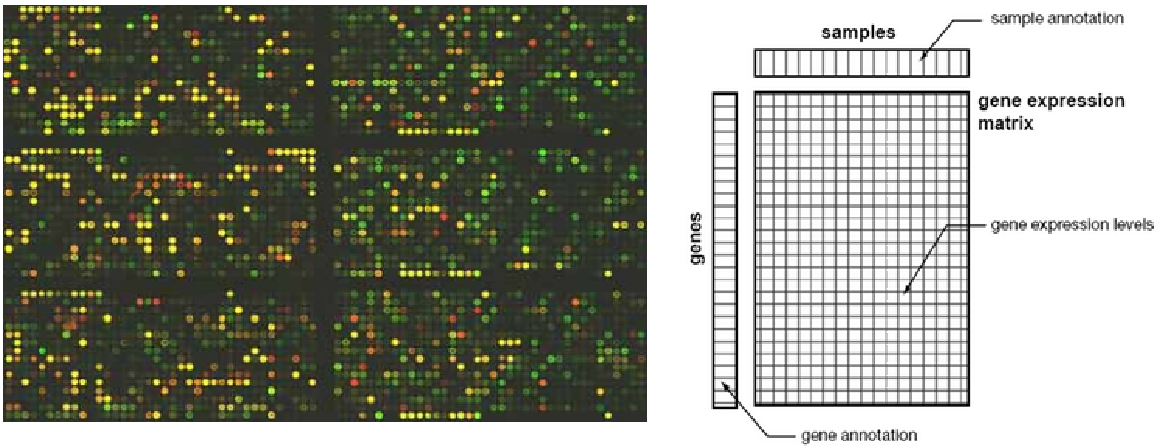


Figure 2.2. A portion of a two-color DNA microarray image and data format of DNA microarray [4].

analysis for gene expression and gene discovery. An experiment with a single DNA chip can provide information on thousands of genes simultaneously. An orderly arrangement of the probes on the support is important as the location of each spot on the array is used for the identification of a gene.

2.2.2. Types of Microarrays

Depending upon the kind of immobilized sample used to construct arrays and the information fetched, the Microarray experiments can be categorized in three ways:

(i) *Microarray Expression Analysis:* In this experimental setup, the cDNA derived from the mRNA of known genes is immobilized. The sample has genes from both normal as well as diseased tissues. Spots with more intensity are obtained for a diseased tissue gene if the gene is over-expressed in the diseased condition. This expression pattern is then compared to the expression pattern of a gene responsible for a disease.

(ii) *Microarray for Mutation Analysis:* For this analysis, the researchers use gDNA. The genes might differ from each other by as less as a single nucleotide base. A single base difference between two sequences is known as Single Nucleotide Polymorphism (SNP) and detecting them is known as SNP detection.

(iii) *Comparative Genomic Hybridization:* It is used for the identification of the in-

crease or decrease of the important chromosomal fragments harboring genes involved in a disease.

Affymetrix is one of the leading providers of DNA microarrays. There are several features that make Affymetrix high density oligonucleotide arrays popular. The most distinguishing feature of Affymetrix arrays is that on Affymetrix GeneChips, the expression intensity of a gene is reported by a probe set that comprises 11-20 individual probe pairs. In the latest generation GeneChips the number of probe pairs has stabilized and is now 11. Each probe pair contains a perfect match (PM) 25 mer oligonucleotide probe, which is designed to hybridize specifically to a unique gene transcript, and a mismatch (MM) probe of the same length, which differs from the PM probe by one single base in the center of the sequence. The MM probe is intended to measure non-specific hybridization. Probe-set algorithms developed by Affymetrix interpret the signals from each 22-oligonucleotide probe set, and derive a single value (signal) from the patterns of hybridization to the 22 individual probes.

Each individual PM-MM probe pair are in close proximity but as the 11 PM-MM probe pairs are spread over the whole array, this results in a very robust assay as small irregularities in background intensities will only affect a few of these probe sets. A second advantage is the fact that the signal results from the subtraction of the MM signal from the PM signal, which makes it possible to have an unbiased determination of signal and non-specific signal for each of the 11 probes. It has been shown that the signal produced by the MM probe is not exclusively a non-specific signal. In particular, when a transcript is present at high levels, the labeled target will also hybridize to the MM probe. As the subtraction of mismatch signal will increase the error and the fact that the target will also hybridize although with reduced affinity, using the PM-MM feature of the Affymetrix arrays will lower sensitivity and increase variability. On the other hand, using the PM-MM feature will reduce false positive signals. In order to reduce manufacturer's bias, for the real microarray data sets used in this thesis, we opted to use the Affymetrix platform.

2.2.3. Applications of Microarrays

Gene Discovery: DNA Microarray technology helps in identifying new genes, as well as determining their function and expression levels under different conditions.

Disease Diagnosis: DNA Microarray technology helps researchers learn more about different diseases such as heart diseases, mental illness, infectious disease and especially the study of cancer. Until recently, different types of cancers have been classified on the basis of the organs in which the tumors develop. Now, with the evolution of microarray technology, it will be possible for the researchers to further classify the types of cancers on the basis of the patterns of gene activities in the tumor cells. This will tremendously help the pharmaceutical community to develop more effective drugs, as the treatment strategies will be targeted directly to the specific type of cancer.

Drug Discovery: Microarray technology has extensive application in Pharmacogenomics, which is the study of correlations between therapeutic responses to drugs and the genetic profiles of the patients. Comparative analysis of the genes from a diseased and a normal cell will help the identification of the biochemical constitution of the proteins synthesized by the diseased genes. The researchers can use this information to synthesize drugs, which combat with these proteins and reduce their effect.

Toxicological Research: Microarray technology provides a robust platform for the research of the impact of toxins on the cells and their passing on to the progeny. Toxicogenomics establishes correlation between responses to toxicants and the changes in the genetic profiles of the cells exposed to such toxicants.

All in all, microarray technology is a developing technology used to study the expression of many genes at once. It involves placing thousands of gene sequences in known locations on a glass/nylon/silicon slide called a gene chip. A sample containing DNA or RNA is placed in contact with the gene chip. Complementary base pairing between the sample and the gene sequences on the chip produces light that is measured.

Areas on the chip producing light identify genes that are expressed in the sample.

In the recent past, microarray technology has been extensively used by the scientific community. Consequently, over the years, there has been a lot of generation of data related to gene expression. This data is scattered and is not easily available for public use. For ease the accessibility to this data, the National Center for Biotechnology Information (NCBI) has established the Gene Expression Omnibus or GEO. It is a data repository facility, which includes data on gene expression from various sources.

2.2.4. Data Normalization

Normalization is an essential procedure in the analysis of DNA microarrays to compare data from different arrays or colour channels. Measurements on microarrays may be systematically biased by diverse effects such as efficiency of RNA extraction, reverse transcription, label incorporation, exposure, scanning, spot detection, and etc. Furthermore, there are systematic effects due to characteristics of the array, such as effects of different probes (i.e. cDNAs or oligos), spotting effects, region effects, and pin effects. Normalization attempts to compensate for such effects through use of internal controls [31].

There are three types of internal controls that can be used for normalization:

(i) Most commonly normalization is based on all genes on the array. The assumption is that between two conditions the majority of genes do not change in terms of their expression level.
(ii) Previously known Housekeeping-genes.
(iii) Spiked-in control genes.

Usually one of these methods is used for normalization and the other two can be used to validate the results. In most cases, normalization is most stable using the majority of genes, at least for microarrays where “all” genes in a genome are represented. For specialized arrays, e.g. with only disease specific genes, it might be

necessary to spot a sufficient number of housekeeping genes or spiked-in controls and use these for normalization. Special care should be taken, when using housekeeping genes for normalization, as they often vary depending on conditions.

It is, however, necessary to make sure that the data being compared is actually comparable. Normalization generally makes data look increasingly similar. However, this doesn't make sense if there are fundamental reasons why data is not comparable.

Normalization consists of several steps:

(i) *Background Correction:*
   (a) Local background (image analysis) – if the image contains heterogeneous background, scratches, etc.
   (b) Global background (e.g. 5% quantile)
   (c) No background correction – some people avoid background correction. This usually leads to some underestimation of the ratio of differential expression, but avoids some of the problems connected to background correction.

(ii) *Transformation:* Microarray intensities are generally log transformed. This scaling should roughly adjust the variance to be the same for all intensities. Differences of log2 intensities reflect the log2 ratios (M values) for a comparison. As an alternative to log2 scaling, the variance stabilizing transformation VSN (see below) may be used (this is roughly equivalent to using the natural logarithm instead of log2).

(iii) *Robust estimation of a "rescaling" factor (e.g. median of differences):* One of the methods for internal control (all genes, housekeeping genes, or spiked-in controls) is used for this purpose.

There are many normalization methods. Which of the methods is most stable and gives the best results is dependent on the type of data, the image analysis program, etc. To determine the best method, it is a good idea to try several methods initially on a few datasets and inspect the results visually using controls.

(i) *Scale normalization:* the simplest way to normalize data is simply to adjust the scale of the data, e.g. set the median of differences to 0. However, this does not consider any region or intensity dependent effects.

(ii) *Lowess (aka loess):* Local regression takes into account intensity dependent effects and might partially correct for background (additive) effects. There are also variants that take into account local effects, e.g. print-tip lowess. This type of normalization is most commonly used for two-colour arrays.

(iii) *Quantile:* This approach is similar to scale normalization but it is more drastic, as all of the various quantiles are adjusted and not only the 50% quantile (median). This type of normalization is most commonly used for affymetrix arrays.

(iv) *VSN:* Estimates an additive and multiplicative offset and transforms the data to have equal variance for all intensities. This transformation is similar to using the natural log transformation, but tries to adjust for effects that are often observed after background subtraction (i.e. high-variance for lowly expressed genes). VSN can be used with cDNA or affymetrix data, and is advisable if you observe unstable results with lowly expressed genes.

Prior to applying a normalization scheme, quality control on the experiment has to be performed. This can be achieved by several means:

(i) Negative controls, heterogonous DNA, genes from different organisms can be spotted to check background hybridization levels.

(ii) One approach is to use housekeeping genes whose expression is assumed to be constant under all conditions.

(iii) Spiked-in controls (positive dynamic range controls, negative controls, ratio controls) provide a good indication of the quality of an experiment.

(iv) Check whether the controls behave as expected after normalization.

Visual inspection of the data before and after normalization, for example, by means of a scatter plot or MA plot is essential and can help to avoid serious errors during the normalization procedure [32]. Scatter Plots (log2 red channel vs. log2 green

channel, or log2 expression level 1 vs. log2 expression level 2) or MA-plots (average log2 expression A vs. log2 difference M) are common ways to quickly inspect the data of a comparison. Further, to look at the spatial distribution of expression values or quality values on a chip or across different chips can help to detect problems with printing or hybridization.

Normalization is essential to compare the varying conditions of microarray experiments. While there are several methods for normalization, the choice of which method gives the best results really depends on your local settings. Use visual inspection of the data and comparisons before and after normalization to control that the procedure worked correctly. It is standard to display data in log2 scale. Use different types of controls, i.e. negative controls, spiked in controls, and housekeeping genes when spotting arrays, to be able to observe possible problems with the hybridizations or normalization procedure. Whenever possible normalize on the majority of the genes or use a sufficiently large number of non-differentially expressed controls to normalize the data.

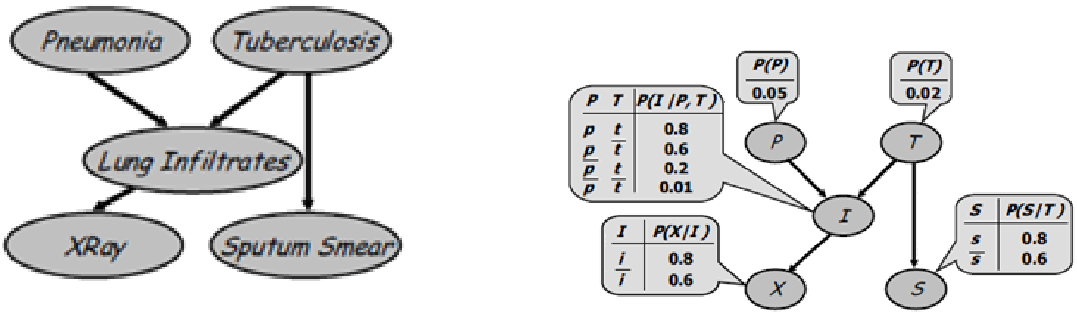
3. BAYESIAN NETWORKS

Graphical models have become an extremely popular tool for modelling uncertainty. They provide a principled approach to dealing with uncertainty through the use of probability theory, and an effective approach to coping with complexity through the use of graph theory. The two most common types of graphical models are Bayesian networks (also called belief networks or causal networks) and Markov networks (also called Markov Random Fields (MRFs)).

At a high level, our goal is to efficiently represent a joint distribution P over some set of random variables $X = X_1, \ldots, X_n$. Even in the simplest case where these variables are binary-valued, a joint distribution requires the specification of 2^n numbers — the probabilities of the 2^n different assignments of values $x_1, \ldots, x_n$. However, it is often the case that there is some structure in the distribution that allows us to factor the representation of the distribution into modular components. The structure that graphical models exploit is the independence properties that exist in many real-world phenomena.

3.1. Introduction

A Bayesian network (BN) is a tool for modeling and reasoning with uncertain beliefs. BNs have been successfully applied to create consistent probabilistic representations of uncertain knowledge in diverse fields such as medical diagnosis [33], image recognition [34], language understanding [35,36], search algorithms [37], and many others. The core of the Bayesian network representation is a directed acyclic graph (DAG) G. The nodes of G are the random variables in our domain and the edges correspond, intuitively, to direct influence of one node on another. One way to view this graph is as a data structure that provides the skeleton for representing the joint distribution compactly in a factorized way. Let G be a BN graph over the variables $X_1, \ldots, X_n$. Each random variable X_i in the network has an associated conditional probability distribution (CPD) or local probabilistic model. The CPD for X_i, given its parents in



(a) (b)

Figure 3.1. (a) A simple Bayesian network; (b) The same Bayesian network, together with the conditional probability tables [5].

the graph (denoted Pa_{X_i}), is $P(X_i \mid Pa_{X_i})$. It captures the conditional probability of the random variable, given its parents in the graph. CPDs can be described in a variety of ways. A common, but not necessarily compact, representation for a CPD is a table which contains a row for each possible set of values for the parents of the node describing the probability of different values for X_i. These are often referred to as table CPDs, and are tables of multinomial distributions.

Consider the simple Bayesian network shown in Figure 3.1. This is a toy example indicating the interactions between two potential diseases, pneumonia and tuberculosis. Both of them may cause a patient to have lung infiltrates. There are two tests that can be performed. An x-ray can be taken, which may indicate whether the patient has lung infiltrates. There is a separate sputum smear test for tuberculosis. Figure 3.1a shows the dependency structure among the variables. All of the variables are assumed to be Boolean. Figure 3.1b shows the conditional probability distributions for each of the random variables. We use initials P, T, I, X, and S for shorthand. At the roots, we have the prior probability of the patient having each disease. The probability that the patient does not have the disease a priori is simply 1 minus the probability that she has the disease; for simplicity only the probabilities for the true case are shown. Similarly, the conditional probabilities for the non-root nodes give the probability that the random variable is true, for different possible instantiations of the parents.

Definition 3.1. *Let G be a Bayesian network graph over the variables $X_1, \ldots, X_n$.*

We say that a distribution P_B over the same space factorizes according to G if P_B can be expressed as a product:

$$P_B\left(X_1, \ldots, X_n\right) = \prod_{i=1}^{n} P(X_i|\mathrm{Pa}_{X_i}). \tag{3.1}$$

A Bayesian network is a pair (G, θ_G) where PB factorizes over G, and where P_B is specified as set of CPDs associated with G's nodes, denoted θ_G.

The equation above is called the *chain rule for Bayesian networks.* It gives us a method for determining the probability of any complete assignment to the set of random variables: any entry in the joint can be computed as a product of factors, one for each variable. Each factor represents a conditional probability of the variable given its parents in the network.

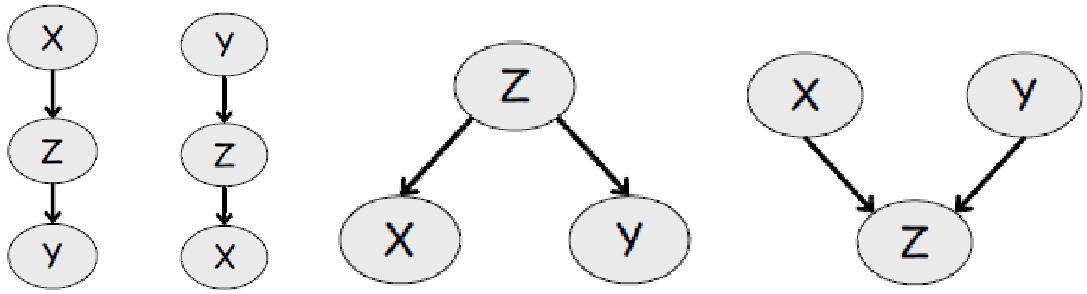
Example 3.2. *The Bayesian network in Figure 3.1a describes the following factorization:*

$$P\left(P, T, I, X, S\right) = P\left(P\right) P\left(T\right) P\left(I \mid P, T\right) P\left(X \mid I\right) P\left(S \mid T\right) \tag{3.2}$$

Sometimes it is useful to think of a Bayesian network as describing a generative process. We can view the graph as encoding a generative sampling process executed by nature, where the value for each variable is selected by nature using a distribution that depends only on its parents. In other words, each variable is a stochastic function of its parents.

3.2. Conditional Independence Assumptions in Bayesian Networks

Another way to view a Bayesian network is as a compact representation for a set of conditional independence assumptions about a distribution. These conditional independence assumptions are called the *local Markov assumptions.*



(a) (b) (c) (d)

Figure 3.2. (a) An indirect causal effect; (b) an indirect evidential effect; (c) a common cause; (d) a common effect [5].

Definition 3.3. *Given a BN network structure G over random variables $X_1, \ldots, X_n$, let NonDescendant s_{Xi} denote the variables in the graph that are not descendants of X_i. Then G encodes the following set of conditional independence assumptions, called the local Markov assumptions:*

For each variable X_i, we have that

$$(X_i \perp nondescendants_{X_i} \mid Pa_{X_i}) \tag{3.3}$$

In other words, the local Markov assumptions state that each node X_i is independent of its nondescendants given its parents.

Example 3.4. *The BN in Figure 3.1a describes the following local Markov assumptions: $(P \perp T \,|\emptyset)$, $(T \perp P \,|\emptyset)$, $(X \perp \{P, T, S\}|I)$, and $(S \perp \{P, I, X\}|T)$. These are not the only independence assertions that are encoded by a network.*

A general procedure called d-separation (which stands for directed separation) can answer whether an independence assertion must hold in any distribution consistent with the graph G. However, note that other independencies may hold in some distributions consistent with G; these are due to flukes in the particular choice of parameters of the network (and this is why they hold in some of the distributions).

Returning to our definition of d-separation, it is useful to view probabilistic influence as a flow in the graph. Our analysis here tells us when influence from X can flow through Z to affect our beliefs about Y. We will consider flow allows (undirected) paths in the graph. Consider a simple three-node path $X - Y - Z$. If influence can flow from X to Y via Z, we say that the path $X - Y - Z$ is active. There are four cases:

- *Causal path* $X \rightarrow Z \rightarrow Y$: active if and only if Z is not observed.
- *Evidential path* $X \leftarrow Z \leftarrow Y$: active if and only if Z is not observed.
- *Common cause* $X \leftarrow Z \rightarrow Y$: active if and only if Z is not observed.
- *Common effect* $X \rightarrow Z \leftarrow Y$: active if and only if either Z or one of Z's descendants is observed.

A structure where $X \rightarrow Z \leftarrow Y$ (as in Figure 3.2d) is also called a v-structure.

Example 3.5. *In the BN from Figure 3.2a, the path from $P \rightarrow I \rightarrow X$ is active if I is not observed. On the other hand, the path from $P \rightarrow I \leftarrow T$ is active if I is observed.*

Now consider a longer path $X_1 - ... - X_n$. Intuitively, for influence to "flow" from X_1 to X_n, it needs to flow through every single node on the trail. In other words, X_1 can influence X_n if every two-edge path $X_{i-1} - X_i - X_{i+1}$ along the trail allows influence to flow. We can summarize this intuition in the following definition:

Definition 3.6. *Let G be a BN structure, and $X_1 - ... - X_n$ a path in G. Let E be a subset of nodes of G. The path $X_1 - ... - X_n$ is active given evidence E if whenever we have a v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, then X_i or one of its descendants is in E; no other node along the path is in E.*

Our flow intuition carries through to graphs in which there is more than one path between two nodes: one node can influence another if there is any path along which influence can flow. Putting these intuitions together, we obtain the notion of d-separation, which provides us with a notion of separation between nodes in a directed graph (hence the term d-separation, for directed separation):

Definition 3.7. *Let X, Y, Z be three sets of nodes in G. We say that X and Y are d-separated given Z, denoted $d-sep_G(X;Y \mid Z)$, if there is no active path between any node $X \in X$ and $Y \in Y$ given Z.*

3.3. Learning Bayesian Network

Learning BNs basically consists of two different components:

(i) learning the graphical structure (model selection)
(ii) learning the conditional distributions (CPT's; parameter estimation).

To learn a Bayesian network on variables $X_1, ...X_n$, we start with M measurements of n variables in a data set $D = f(X_1(1), ..., X_n(1)), ..., (X_1(M), ..., X_n(M))g$. Most learning approaches define a hypothesis space of potential network models and use the data to find ones that are most likely given the data. Learning a network model on a set $X_1, ...X_n$ of variables entails inferring the graph of dependencies between them, as well as the parameters θ consisting of the local conditional probabilities: $P(X_i|ParentsG(X_i))$. If the graph structure G is known, the parameters θ can be estimated from the available data D by maximum likelihood estimation if M is large, or by Bayesian estimation, if M is small and priors on the parameter vector θ are available. The likelihood of the data D given parameter vector θ for a known structure G is:

$$L(\theta; D) = P(D \mid \theta) = \prod_{i=1}^{M} P((X_i(i), \ldots, X_n(i))|\theta) \tag{3.4}$$

For maximum likelihood estimation, we estimate parameters θ^*_{ML} such that

$$\theta^*_{ML} = argmax_\theta L(\theta; D) \tag{3.5}$$

When the number of samples M is small, the above approach tends to overfit the model parameters to the available data. Bayesian methods reduce overfitting by representing and using the available knowledge about the parameters in the form of a prior distribution $P(\theta)$. The data D then serves to update the prior $P(\theta)$ to yield the posterior probability distribution $P(\theta|D)$. By Bayes Rule,

$$P\left(\theta \mid D\right) = \frac{P\left(D|\theta\right) P\left(\theta\right)}{P\left(D\right)} \tag{3.6}$$

Then, the estimated parameter θ^*_{MAP} is:

$$\theta^*_{MAP} = argmax_\theta P\left(D \mid \theta\right) P(\theta) \tag{3.7}$$

Since P(D) is independent of θ, it is treated as normalizing constant, and the scoring function is simply the product of the likelihood of the data given the parameter vector θ and the prior $P(\theta)$.

Learning the structure G of the Bayesian network from data is a very challenging problem. The most common approach to discovering the structure of Bayesian networks from data is to define a space of graph models to consider, and then set up a scoring function that evaluates how well a model explains the available data. Then, an optimization algorithm is used to search for the highest-scoring model. The scoring

function is the logarithm of the posterior probability of the network structure given the data

$$Score\,(G;D) = logP\,(G \mid D) = logP\,(D \mid G) + logP(G) \tag{3.8}$$

where

$$P\,(D \mid G) = \int_{\theta} P\,(D \mid G, \theta)\, P\,(\theta \mid G)\, d\theta$$

We average over all parameters θ associated with a graph structure G. Learning a Bayesian network that provably maximizes the above scoring function is NP-hard. Thus, learning optimal Bayesian networks for high throughput datasets is computationally infeasible [38].

4. BAYESIAN PATHWAY ANALYSIS(BPA)

The overall workflow of the BPA system is shown in Figure 4.1.

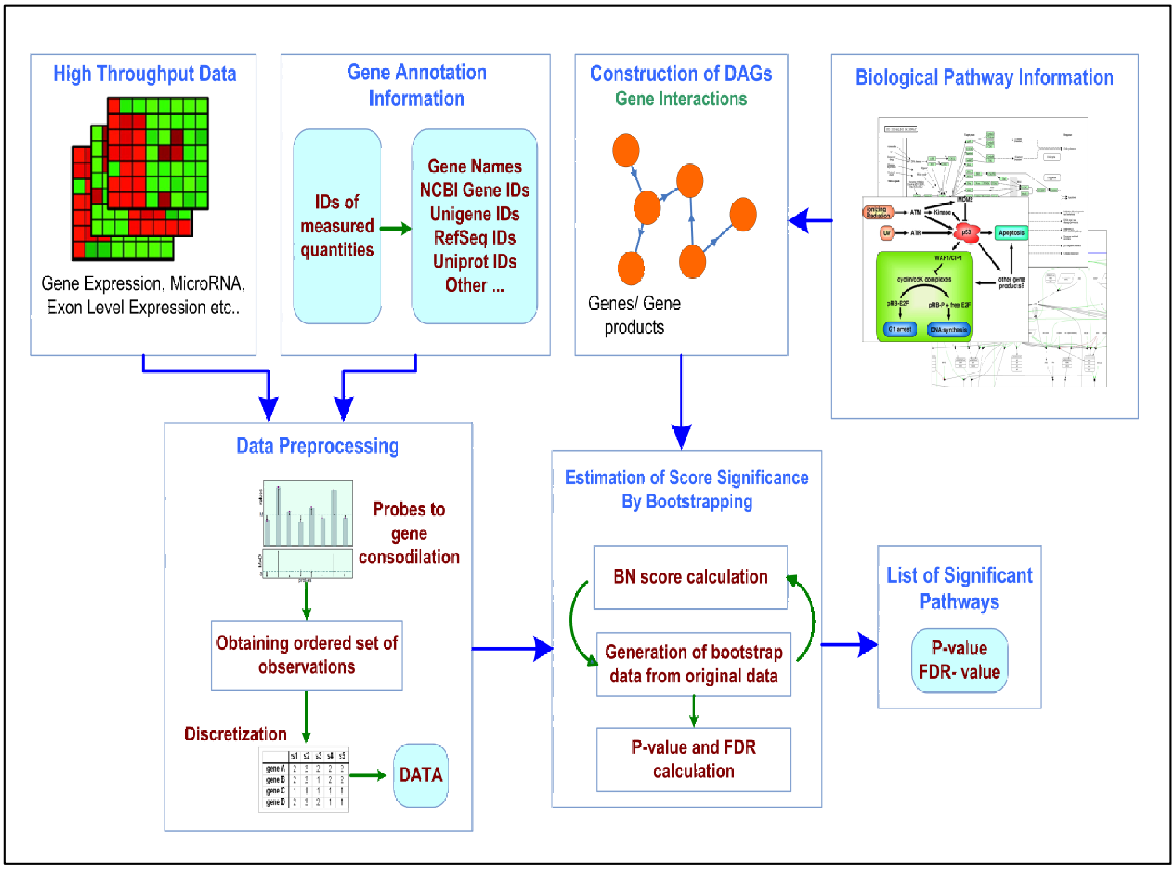


Figure 4.1. Outline of BPA.

Outline of BPA is the following. DAGs are created from pathways in external databases where, if necessary, multiple occurrences of the same nodes are merged and cyclicity is resolved preserving the dependencies entailed by the original pathway. The nodes in the DAGs and microarray probes are mapped using gene annotation information. Multiple probes mapped to the same node are replaced by a robust average value followed by the discretization of all signal values. A score metric, which is a measure of the probability of the data given the fixed network structure, is calculated. The significance of this score (p-value) is assessed by the randomization of the data via bootstrapping. The FDR values are calculated to account for multiple hypothesis testing.

Biochemical network data of pathways are retrieved from KEGG representing molecular interaction and reaction networks for metabolism, genetic information processing, environmental information processing, cellular processes and diseases. Nodes in a typical biological pathway often represent ortholog elements and complexes that can be mapped to more than one gene. For example, TGF ortholog element in a KEGG map can be a representation of several ortholog genes, which are transforming growth factor beta1, beta2, and beta3 genes and protein complexes such as activin receptor complex located on the membrane of a cell may have several subunits that are encoded by different genes. The IDs of the nodes are converted to Entrez Gene ID used by NCBI and mapped with the input gene expression data sets that can come from the GenBank, Unigene, Gene, Refseq and SwissProt databases, or Affymetrix.

4.1. Construction of Directed Acyclic Graphs (DAGs)

First, repeating entries are merged and represented as a single node in the DAG while conserving the edge relations. Since we are interested in building BN models, which utilize DAGs, cyclic paths are eliminated using Spirtes' method [22]. In this procedure, graph representation of structural equation models [39] are converted to collapsed acyclic graphs such that d-separations in the collapsed graph entails same independency relations defined by the model.

The method is employed using the following steps:

(i) remove all edges between members of the cyclegroup (equivalent to strongly connected components)
(ii) arbitrarily number the vertices in the cyclegroup
(iii) add an edge from each lower number vertex to each higher number vertex
(iv) for each parent A of a member of the cyclegroup that is not itself in the cyclegroup, add an edge from A to each member of the cyclegroup.

We use Tarjan's algorithm to find strongly connected components (cyclegroups) of the cyclic graph [40]. For a given BN, all d-separations are conditional indepen-

dencies and every conditional independency implied by the BN is identified by d-separation [24]. Therefore, the way of solving cyclicity preserves the distributional features explained by the pathway after it has been converted to a DAG. To this end, we have modelled a biological pathway as a BN, which now can be tested against input data to assess its fitness.

4.2. Microarray Data Preprocessing and Discretization

BPA assumes normalized gene expression data as input. First, IDs used in the array platform corresponding to a given node in the pathway representation are pooled and one representative signal value per node is calculated using one-step Tukey's biweight algorithm [23]. This approach is robust to outliers and performs a weighted average of input signal values. BPA addresses experimental designs consisting of two groups of samples (e.g. cancer vs. normal). Let SG_1 and SG_2 represent two groups of samples in the data set with C_1 and C_2 samples in each group, respectively. Let g_{i1j} and g_{i2k} be the expression values of the $i-th$ node in the pathway in $j-th$ and $k-th$ samples in the sample groups SG_1 and SG_2, respectively, where $1 \leq j \leq C_1$ and $1 \leq k \leq C_2$. Let X_i, $1 \leq i \leq N$, represent the random variable for the $i-th$ node in a BN with N nodes. An ordered set of observations, O, for the data set is obtained by pairwise comparison of all samples in sample groups SG_1 and SG_2. The $l-th$ element of O, o_l, corresponds to the comparison of the $j-th$ and $k-th$ samples in the sample groups SG_1 and SG_2 such that $l{=}j-1\times C_2+k$, where $1 \leq j \leq C_1$ and $1 \leq k \leq C_2$. Thus the cardinality of O is $C_1 \times C_2$. Each o_l is a vector with dimension equaling the number of nodes in the pathway, N, such that the $i-th$ element of o_l, o_{li}, equals g_{i2k}/g_{i1j}, where j and k are related to l as described above. The data matrix D, with elements d_{li} are obtained from O such that d_{li} equals 1 if $o_{li}{<}0.5$ or $o_{li}{>}2$ (i.e. a gene is deregulated) and 2 otherwise (i.e. a gene is not changed). BNs are initialized using the equivalent sample size method for prior beliefs [24]. D, which consists of N columns and $C_1 \times C_2$ rows, is sequentially evaluated row by row, where each row is used to update the Dirichlet distribution parameters driving the nodes of BN. Upon conclusion of evaluation of D, the score for the BN is calculated as described below.

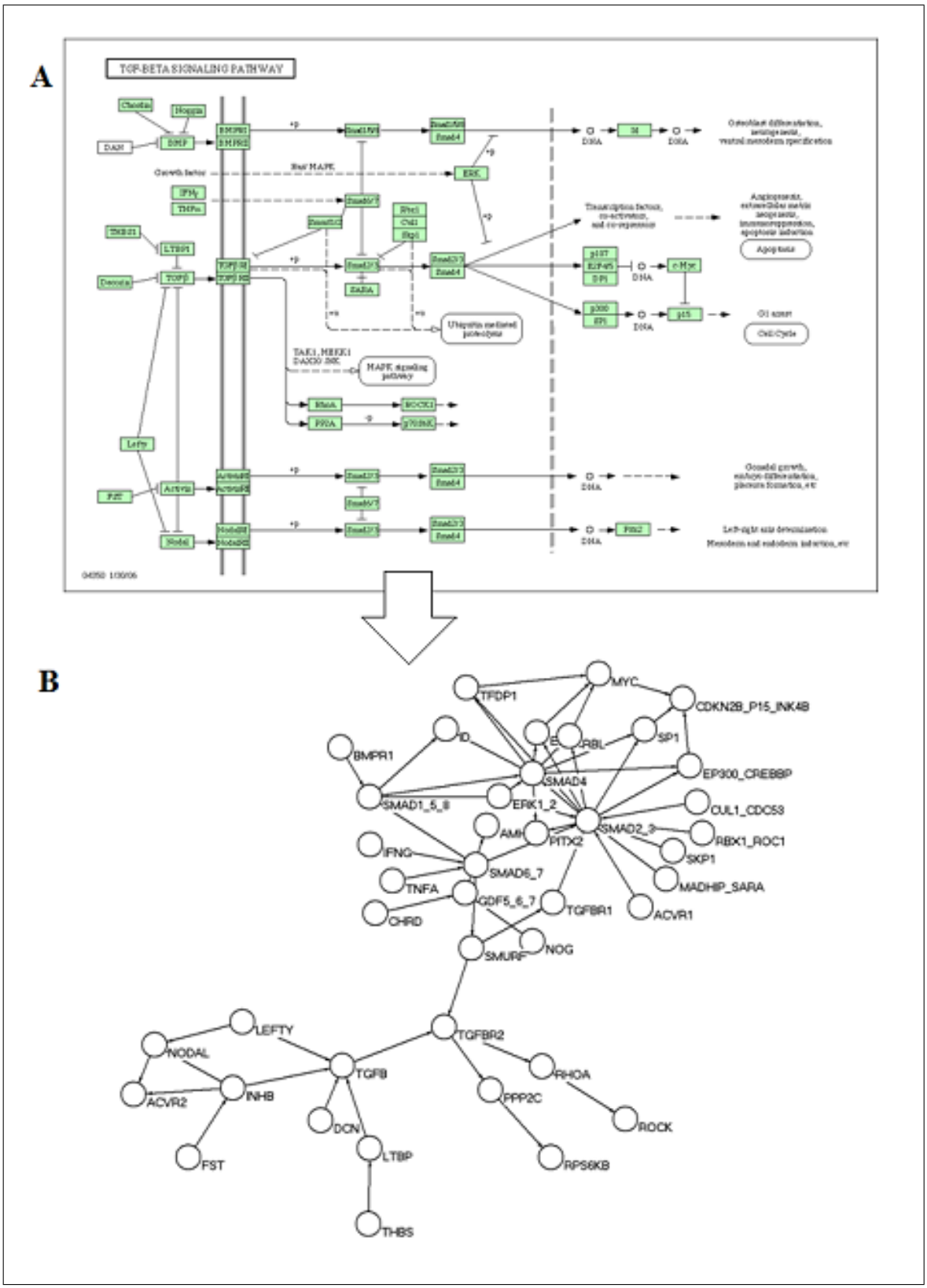


Figure 4.2. Construction of directed acyclic graphs. (A) TGF-β Signaling Pathway as retrieved from the KEGG database [6]. (B) DAG produced from TGF-pathway map. Each node is identified by the corresponding gene symbol (e.g. DCN for Decorin).

4.3. Bayesian Score Metric

Following discretization, the nodes in the BN model represent discrete random variables with a multinomial distribution. Dirichlet distribution is chosen as the conjugate prior of the multinomial distribution, with some fixed hyper-parameters [24]. For a particular BN model, the probability of seeing the data set D is [41]

$$P(Data|Model) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(N_{ij})}{\Gamma(N_{ij} + M_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(a_{ijk} + s_{ijk})}{\Gamma(a_{ijk})} \tag{4.1}$$

where N is the number of nodes, qi is the number of different states of node's parents, and r_i is the set of values a node can take on. N_{ij} is the sum of corresponding Dirichlet distribution hyper-parameters a_{ijk}. M_{ij} is the number of times that the parents of node i take on configuration j in the dataset. Of these M_{ij} cases, s_{ijk} is the total number of times in the sample that node i is observed to have value k when its parents take on configuration j. The equation above is used as a score metric and named as Bayesian Scoring Criterion (BSC). Usually the natural logarithm of the score metric is used, which has negative values. BSC remains unchanged when likelihood equivalence is considered [42]. Hyper-parameters a_{ijk} can be determined using the equivalent sample size method (i.e. sum of the Dirichlet parameters driving each node has the same total), in which case the score is called Bayesian Dirichlet Equivalent (BDe) [24].

4.4. Estimation of score significance by randomization via bootstrapping

The statistical significance of the BDe score, S_n, is calculated for a BN by using randomization via bootstrapping. For a one-tailed test with a rejection region in the upper tail, the bootstrap p-value for S_n, $p(S_n)$, can be estimated by the proportion of randomized samples that yield a score greater than S_n. If there are B randomized data sets, then

$$p(S_n) = \frac{1}{B} \sum_{k=1}^{B} I(S_k > S_n) \tag{4.2}$$

where I is the indicator function yielding 1 if Bayesian score is better than the original network score and 0 otherwise and S_k is the score of the BN using $k-th$ randomized data set. As B goes to infinity, the estimated p-value will tend to the ideal p-value and the error in the estimation will be kept minimal [43, 44].

Data generating process for randomized samples works as follows: Suppose dataset D is composed of M cases for a total of N genes and can be considered as an MxN matrix where $l-th$ row $d_l = [d_{l1}, d_{l2}, \ldots, d_{lN}]$, $1 \leq \mathrm{l} \leq N$, and d_{li} is the value of $i-th$ node (gene) in the $l-th$ instance of input data. For each node X_i, we sample with replacement M instances from the $i-th$ column, $[d_{1i}, d_{2i}, \ldots, d_{Mi}]^T$, of the original data matrix D and obtain the newly formed column of the bootstrapped data matrix D_k. The score for this new data matrix is calculated and the whole process is repeated B times. Querying each pathway database that holds few hundreds of networks generates a multiple hypothesis testing problem. This issue is addressed by calculating FDR values for each pathway using Benjamini-Hochberg procedure [45].

5. PERMUTATION, DISCRETIZATION AND SCORING

5.1. Class Label Permutation

In the BPA approach, known biological pathways are modeled as Bayesian Networks (BNs) and pathways that best explain the given high-throughput biological data (HTBD) are identified by scoring the fitness of each network. The observation matrix is generated by pairwise comparisons of samples in each group (cancer versus normal) in microarray data. In this approach, the change of the expression value between samples for a gene in a pathway is identified via discretization of the observation matrix with a particular method.

In the current system, the degree to which a pathway explains the given HTBD is measured using the Bayesian Dirichlet equivalent (BDe) score and the statistical significance of this measurement is assessed by randomization via bootstrapping, where the observed score is ranked against scores obtained from randomized data sets. Randomized data sets are obtained by changing the structure of the columns of the observation matrix via sampling with replacement each column separately.

This method works successfully when an observation matrix as in Table 5.1a is generated. However, if the observation matrix is generated like in Table 5.1b randomizing the columns of the observation matrix will not result in any change. Therefore, the scores obtained by the randomized data sets will be the same, making the significance assessment almost impossible to achieve.

To solve this problem, we applied the permutation method previously described to randomize gene expression data sets [30]. This randomization is done by replacing the samples of each class (cancer versus normal) randomly. Suppose we have a dataset composed of 10 normal and 10 cancer samples. In one instance of the permutation, for example, the 3^{rd}, 5^{th}, and 6^{th} normal samples are replaced with the 1^{st}, 7^{th}, and 9^{th} cancer samples. The observation matrix is generated by pairwise comparison of

Table 5.1. Sample observation matrices discretized with a 2-level discretization method.

Obs.	$\mathbf{G_1}$	$\mathbf{G_2}$	$\mathbf{G_3}$	$\mathbf{G_4}$
$\mathbf{O_1}$	1	2	1	1
$\mathbf{O_2}$	1	1	2	1
$\mathbf{O_3}$	2	2	2	1
$\mathbf{O_4}$	1	2	2	1
$\mathbf{O_5}$	2	1	2	1
$\mathbf{O_6}$	2	2	1	1
$\mathbf{O_7}$	1	1	2	1

(a)

Obs.	$\mathbf{G_1}$	$\mathbf{G_2}$	$\mathbf{G_3}$	$\mathbf{G_4}$
$\mathbf{O_1}$	1	2	2	1
$\mathbf{O_2}$	1	2	2	1
$\mathbf{O_3}$	1	2	2	1
$\mathbf{O_4}$	1	2	2	1
$\mathbf{O_5}$	1	2	2	1
$\mathbf{O_6}$	1	2	2	1
$\mathbf{O_7}$	1	2	2	1

(b)

the signal values over the new order of the two classes followed by discretization. This procedure is repeated B times and the pathway scores are calculated using the discretized matrix. As a result, the statistical significance of the observed score can be assessed accurately via ranking against scores obtained from different observation matrices generated by these randomized data sets.

5.2. Discretization

The current system included a discretization method such that the fold change is represented as 1 if its value is greater than 2 or less than 1/2 (i.e. a gene is deregulated), and as 2 otherwise. Another use of the 2-level discretization is choosing a cut-off value of 3, i.e., the fold change is represented as 1 if its value is greater than 3 or less than 1/3 and as 2 otherwise. In 3-level discretization with the cutoff value 2, the fold change is represented as 1 if its value is greater than 2, as 2 if less than 1/2, and as 3 otherwise. In 3-level discretization with the cut-off value of 3, the fold change is represented as 1 if its value is greater than 3, as 2 if less than 1/3, and as 3 otherwise.

In this thesis, we propose new discretization methods to be utilized in processing the observed fold change values for use by Bayesian scoring metrics [25,26]. An N-by-M matrix E is used to denote the microarray data, where N is the number of genes, and M is the number of samples. E(n,m) denotes the expression value of gene n for the sample m. E(n,:) denotes the expression data of gene n for all samples, and E(:,m) denotes the expression data of all genes for the sample m.

5.2.1. Equal Width Discretization (EWD)

EWD divides the expression row between $E(n,:)_{min}$ and $E(n,:)_{max}$ into k intervals of equal width. Thus, the intervals of gene n have width w = ($E(n,:)_{max}$ - $E(n,:)_{min}$)/k, with boundary points at $E(n,:)_{min}$ + w, $E(n,:)_{min}$ + 2w, ..., $E(n,:)_{min}$ + (k - 1)w. k is a positive integer and is a user predefined parameter.

5.2.2. Equal Frequency Discretization (EFD)

EFD divides the sorted E(n,:) into k intervals so that each interval contains approximately the same number of expression values.

5.2.3. K-means Discretization

K-means divides E(n,:) into k intervals by k-means clustering so that similar expression values of gene n are placed in the same interval.

5.2.4. Column K-means Discretization (Co-k-means)

Co-k-means divides E(:,m) into k intervals by k-means clustering so that similar expression values for the sample m are placed in the same interval.

5.2.5. Bidirectional K-means Discretization (Bi-k-means)

In the bi-k-means method both k-means and co-k-means are respectively implemented with parameter k+1, giving every expression value two discretized values. If the product of the two values is equal to or greater than x^2, and less than $(x+1)^2$, the final discretized value of this expression value is x, where x is a positive integer ranging from 1 to k. Finally, the expression values are divided into k intervals.

5.2.6. Automatic Threshold Discretization

There are two options for the automatic threshold discretization [26], which iteratively determines the cut-off values by minimizing the variance. The whole expression data E is divided into two intervals according to a certain cut-off value in global option. Local option of this method divides E(:,m) into two intervals according to the cutoff values defined for each column (gene) separately.

5.3. Scoring

In addition to the BDe scoring scheme, we propose the following scoring metrics to be used in the BPA.

5.3.1. Akaike Information Criterion (AIC)

One of the most commonly used information criterion is Akaike Information Criterion (AIC) [27]. AIC selects the model that minimizes the negative likelihood penalized by the number of parameters as given in the following equation:

$$AIC\left(M, D\right) = \log \hat{P}\left(D \mid M\right) - p \tag{5.1}$$

where $\hat{P}(D \mid M)$ is the maximum likelihood of the model and p is the number of parameters in the model.

5.3.2. Bayesian Information Criterion (BIC)

Another widely used information criteria is Bayesian Information Criterion (BIC) [28]. Unlike AIC, BIC is consistent and improves in performance in large sample sizes. BIC is defined as:

$$BIC(M, D) = \log \hat{P}(D \mid M) - \frac{p}{2} \log N \tag{5.2}$$

Superficially, BIC differs from AIC only in the second term which now depends on sample size N. From a Bayesian perspective, BIC is proposed to find the most probable model given the data.

5.3.3. Factorized Normalized Maximum Likelihood (fNML)

Silander *et al.* [29] developed the factorized normalized maximum likelihood (fNML) score based on the normalized maximum likelihood (NML) distribution [46] [47]. Given a data set D, the NML model selection criterion choses the model M for which the $P_{NML}(D|M)$ is largest.

$$P_{NML}(D \mid M) = \frac{\hat{P}(D \mid M)}{\sum_{D}^{'} \hat{P}(D' \mid M)} \tag{5.3}$$

After taking the logarithm, the score is in a form of penalized log-likelihood given $G = \{G_1, ..., G_m\}$as the parent set:

$$S_{fNML}(D_i, D_{G_i}) = \log P_{NML}(D_i \mid D_{G_i}) = \log \hat{P}(D_i|D_{G_i}) - \log \sum\nolimits_{D_i} \hat{P}\left(D_i^{'} \mid D_{G_i}\right) \tag{5.4}$$

where the normalizing sum goes over all the possible D_i column vectors.

Even though the penalty term has an exponential number of terms, it can be evaluated efficiently using a linear-time algorithm introduced in [48]. By calculating the penalty term for each variable in the dataset, the NML becomes factorized.

6. RESULTS AND DISCUSSION

6.1. Synthetic Data

We generated synthetic transcriptional regulatory networks and produced simulated gene expression data with noise using SynTReN v1.12 [49]. We created 55 synthetic networks with sizes ranging from 7 to 200. We randomly selected 20 out of 55 pathways to be active and SynTReN generated the corresponding expression datasets for 20 cancer and 20 normal samples with 2249 synthetic genes adding a 4% noise level. For all simulations, we used 1000 permutations and chose a nominal p-value and FDR cutoff values of 0.05 and 0.25, respectively. We assessed the accuracies (if a network-or corresponding gene set-is correctly called active/inactive) of the algorithms for 10 simulated datasets (Data-1, Data-2, ..., Data-10) and provide the results in the tables below. The most successful approach, which is given in Table 6.2, is underlined.

Table 6.1. Prediction Accuracy of hard-cutoff and EWD discretization methods on synthetic datasets. AVG: average; STDEV: standard deviation.

	level_2 FC_2	level_2 FC_3	level_3 FC_2	level_3 FC_3	EWD_2	EWD_3	EWD_4	EWD_5
Data-1	0.836	0.764	0.764	0.818	0.818	0.782	0.782	0.800
Data-2	0.818	0.855	0.764	0.836	0.818	0.818	0.855	0.745
Data-3	0.855	0.855	0.745	0.764	0.855	0.745	0.764	0.764
Data-4	0.855	0.855	0.745	0.764	0.855	0.745	0.764	0.764
Data-5	0.855	0.855	0.745	0.764	0.855	0.727	0.764	0.745
Data-6	0.855	0.855	0.745	0.745	0.855	0.836	0.727	0.745
Data-7	0.855	0.855	0.745	0.745	0.855	0.745	0.764	0.764
Data-8	0.855	0.855	0.745	0.745	0.855	0.745	0.745	0.745
Data-9	0.855	0.855	0.727	0.745	0.855	0.855	0.745	0.764
Data-10	0.855	0.745	0.745	0.745	0.855	0.745	0.764	0.764
AVG	0.849	0.835	0.747	0.767	0.847	0.775	0.767	0.760
STDEV	0.012	0.042	0.010	0.033	0.015	0.046	0.034	0.017

Table 6.2. Prediction Accuracy of EFD and kmeans discretization methods on synthetic datasets. AVG: average; STDEV: standard deviation.

	EFD_2	EFD_3	EFD_4	EFD_5	kmean_2	kmean_3	kmean_4	kmean_5
Data-1	0.782	0.873	0.709	0.873	0.945	0.909	0.927	0.909
Data-2	0.836	0.855	0.855	0.836	0.982	0.945	0.909	0.945
Data-3	0.836	0.891	0.855	0.873	0.982	0.927	0.982	0.873
Data-4	0.855	0.818	0.836	0.945	0.964	1.000	0.964	0.891
Data-5	0.873	0.836	0.836	0.909	0.945	0.964	0.964	0.909
Data-6	0.873	0.800	0.800	0.873	0.945	0.945	0.927	0.945
Data-7	0.855	0.836	0.873	0.855	0.982	0.927	0.964	0.909
Data-8	0.836	0.818	0.836	0.818	1.000	0.945	0.982	0.891
Data-9	0.836	0.873	0.927	0.818	0.982	0.964	0.945	0.964
Data-10	0.836	0.891	0.800	0.891	0.891	1.000	0.964	0.909
AVG	0.842	0.849	0.833	0.869	0.962	0.953	0.953	0.915
STDEV	0.026	0.032	0.057	0.040	0.031	0.030	0.025	0.028

Table 6.3. Prediction Accuracy of cokmeans and bikmeans discretization methods on synthetic datasets. AVG: average; STDEV: standard deviation.

	column kmeans_2	column kmeans_3	column kmeans_4	column kmeans_5	bikmeans_2	bikmeans_3
Data-1	0.636	0.636	0.636	0.636	0.636	0.636
Data-2	0.636	0.636	0.636	0.636	0.636	0.600
Data-3	0.636	0.636	0.636	0.636	0.618	0.636
Data-4	0.636	0.636	0.636	0.636	0.636	0.636
Data-5	0.636	0.636	0.636	0.636	0.636	0.636
Data-6	0.636	0.636	0.636	0.636	0.636	0.636
Data-7	0.636	0.636	0.636	0.636	0.636	0.636
Data-8	0.636	0.636	0.630	0.636	0.636	0.636
Data-9	0.636	0.636	0.630	0.636	0.636	0.636
Data-10	0.636	0.618	0.636	0.636	0.636	0.636
AVG	0.636	0.635	0.635	0.636	0.635	0.633
STDEV	0.000	0.006	0.003	0.000	0.006	0.011

Table 6.4. Prediction Accuracy of bikmeans and automatic threshold discretization methods on synthetic datasets. AVG: average; STDEV: standard deviation.

	bikmeans_4	bikmeans_5	automatic threshold-global	automatic threshold-local
Data-1	0.636	0.600	0.873	0.636
Data-2	0.636	0.618	0.782	0.636
Data-3	0.636	0.636	0.764	0.636
Data-4	0.618	0.564	0.764	0.636
Data-5	0.636	0.636	0.764	0.636
Data-6	0.636	0.636	0.764	0.636
Data-7	0.636	0.636	0.764	0.636
Data-8	0.636	0.618	0.764	0.636
Data-9	0.636	0.636	0.764	0.636
Data-10	0.636	0.618	0.800	0.636
AVG	0.635	0.620	0.780	0.636
STDEV	0.006	0.023	0.035	0.000

According to the simulation results, the best discretization method is the 2-level k-means discretization applied to the rows of the observation matrix. This approach estimates whether a pathway is active with the accuracy of 0.962 ± 0.031. Therefore, 2-level k-means method is used as the discretization method for the experiments to determine the best scoring criterion. The datasets, which are used for the performance measurement of discretization methods, are also used for the scoring methods. The obtained prediction accuracies are in the table below and the best approach is underlined.

Table 6.5. Prediction Accuracy of different scoring methods on synthetic datasets. AVG: average; STDEV: standard deviation.

	BDe	AIC	BIC	fNML
Data-1	0.945	0.964	0.909	1.000
Data-2	0.982	1.000	0.927	0.964
Data-3	0.982	1.000	0.945	1.000
Data-4	0.964	0.982	0.982	1.000
Data-5	0.945	1.000	0.891	1.000
Data-6	0.945	0.982	0.982	0.964
Data-7	0.982	0.982	0.927	0.982
Data-8	1.000	0.982	0.964	0.982
Data-9	0.982	0.982	0.927	0.982
Data-10	0.891	0.945	0.945	0.964
AVG	0.962	0.982	0.940	0.984
STDEV	0.031	0.017	0.030	0.016

According to the simulation results, the best scoring method is the fNML method, which estimates whether a pathway is active with the accuracy of 0.984 ± 0.016. Therefore, 2-level k-means discretization method and fNML scoring method are used for the real microarray data analysis since they give the best combined result.

6.2. Real Microarray Data

To test the optimized and improved BPA performance on real data sets, we used 1 bladder, 2 brain, 2 breast, 1 colon, 2 liver, 1 lung, 1 ovarian, and 2 thyroid cancer data sets. In choosing the data sets, we fixed the platform to be Affymetrix to prevent bias and used data sets where tumor and normal samples are clearly defined and the cancer samples are as homogenous as possible.

Table 6.6. Cancer microarray datasets used on BPA real microarray analysis.

GEO #	**Cancer Type**	**Affymetrics Chip Type**	**# of Samples**	**# of Significant Pathways**
GSE 7476	bladder	HG-U133_Plus_2	12 (9C, 3N)	57
GSE 12907	brain	HG-U133A	25 (21C, 4N)	81
GSE 15824	brain	HG-U133_Plus_2	35 (30C, 5N)	46
GSE 8977	breast	HG-U133_Plus_2	22 (7C, 15N)	16
GSE 22544	breast	HG-U133_Plus_2	18 (14C, 4N)	66
GSE 41328	colon	HG-U133_Plus_2	20 (10C, 10N)	36
GSE 14520	liver	HG-U133A_2	43 (22C, 21N)	59
GSE 14323	liver	HG-U133A_2	66 (47C, 19N)	77
GSE 10799	lung	HG-U133_Plus_2	19 (16C, 3N)	58
GSE 14407	ovarian	HG-U133_Plus_2	24 (12C, 12N)	5
GSE 3678	thyroid	HG-U133_Plus_2	14 (7C, 7N)	4
GSE 6004	thyroid	HG-U133_Plus_2	18 (14C, 4N)	10

The Affymetrix HG-U133 Plus 2.0 GeneChip is composed of more than 54,000 probe sets representing over 47,000 transcripts representing a comprehensive picture of the human transcriptome. HG-U133A and HG-U133A_2 includes approximately 22,000 probesets. Prior to application of the proposed approach, raw microarray data has been normalized using Affymetrix Microarray Analysis Suite (MAS) 5.0 algorithm [50].

For each data set, we applied the proposed analysis method with 500-1000 permutations and assessed significant pathways with a nominal p-value of 0.05 and an FDR of 0.25. In Table 6.6, we list the database accession numbers, cancer type and number of samples utilized in each data set along with the number of significant pathways identified by the proposed workflow.

In Table A.2, we list the pathways found significant in the analyzed cancer data sets. The occurrence of a pathway as significant in a data set is marked with an "X".

In total, we have identified 171 pathways that have been found significant in at least one of the data sets. 15 of these pathways, given in Table A.1, have been found to be significant in at least half of the data sets and therefore potentially represent mechanisms common to different cancer types.

We also investigated the commonality of significant pathways in cancer types represented by two data sets except for the thyroid cancer, which has resulted in very few significant pathways. These results are summarized in Figure 6.1.

In the case of brain and liver cancer data sets, the common pathways consist of 52% and 59% of the dataset with smaller number of pathways. In the breast cancer data sets, we see a less degree of agreement ($\sim$31%).

In addition, we compared our results with a comparable approach, SPIA [2] that combines the evidence obtained from the classical enrichment analysis with a novel type of evidence, which measures the actual perturbation on a given pathway under a given condition. Assessing the capabilities of any pathway analysis method is a

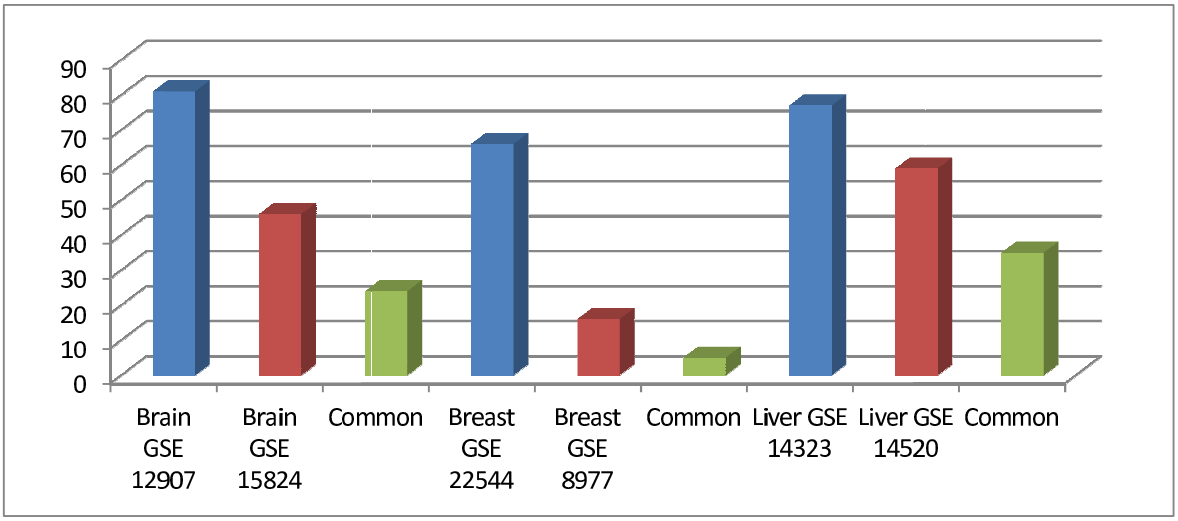


Figure 6.1. Commonality of significant pathways in same cancer types; y-axis: the number of significant pathways.

challenge in real experiments because there is no gold standard and the ground truth is never known. The absence of a definitive answer regarding the involvement of a given pathway doesn't make it possible to calculate exact values for performance metrics such as sensitivity and specificity. However, the methods can be compared in terms of the number of pathways that are found to be significant in a given condition.

The comparison is done on 72 pathways which are shared by BPA and SPIA analysis. Figure 6.2 shows the number of pathways that are found to be significantly active in real microarray data sets detailed in Table 6.6. BPA identified more pathways in two datasets which belong to brain and liver cancer types. There is also a situation, in GSE 14520 liver cancer data set, that our method and SPIA found the same number of pathways (17) as significant, 7 of which are common.

In total, we have identified 15 of these pathways to be significant in at least half of the data sets while SPIA found 29 as significantly active. There are three pathways in common: leukocyte transendothelial migration, tight junction and small cell lung cancer pathways. Moreover, in the remaining 12 pathways identified by BPA, citrate (TCA/tricarboxylic acid) cycle, complement and coagulation cascade and Adipocytokine signaling pathways are not included in the analysis list utilized by SPIA.

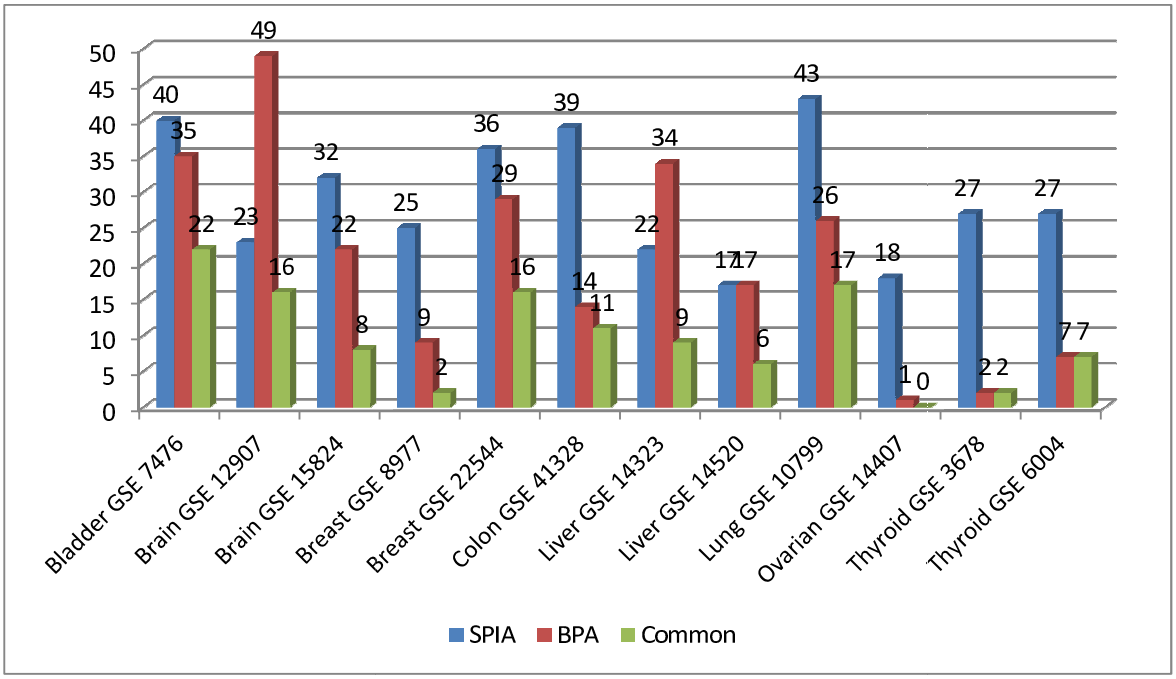


Figure 6.2. Number of pathways found significant in real microarray data sets.

We also compared the commonality of significant pathways in cancer types represented by two data sets except for the thyroid cancer, which has very few significant pathways.

Table 6.7. Commonality comparison of significant pathways in same cancer types.

	BPA (%)	**SPIA** (%)
Brain	72.7	60.9
Breast	44.4	52.0
Liver	70.6	41.2

For brain and liver cancer data sets, the common pathways consist of at least 70% of the dataset with smaller number of pathways identified by BPA, where we found more pathways significant in common than SPIA identifying 60.9% and 41.2%. In the breast cancer data set, we have a less degree of agreement than SPIA.

7. CONCLUSION

The BPA approach, provides a unique perspective that merges Bayesian Network theory and HTBD analysis. BPA is a tool that can be used with most common experimental settings interpreting the results within the context of known biological pathways. Moreover, existing BN approaches on HTBD generally focus on building networks from input data, which makes these approaches applicable on a few dozens of genes due to the complexity of structure learning algorithms. High-throughput platforms generate data for tens of thousands of genes. This approach makes use of relevant experimental information, and is applied to the complete data set within the context of known biological pathways. Our experiments on synthetic and real data sets show that BPA is able to successfully find molecular mechanisms that best describe underlying HTBD.

Our synthetic data simulations identified k-means clustering as the best performing discretization method among equal width, equal frequency, column k-means, bi-directional k-means, and automatic threshold discretization. We find this result reasonable as k-means uses the distribution in the data to minimize the total mean squared error with respect to the discretized values and the real FC occurrences. Also based on the synthetic data results, the scoring method that yielded the highest accuracy was the factorized normalized maximum likelihood (fNML) score [29]. This result was also expected as it has been shown that the BDe scoring scheme is very sensitive to the choice of prior hyper-parameters and AIC and BIC require some manual parameter setting and do not work well with small data sets, which is occasionally the case with HTBD [51]. fNML on the other hand, is an information theory based optimized scoring method that has no tunable parameters.

In the real microarray data analysis using BPA, the pathway that came out in most of the cancer data sets as significantly active (8/12) is the Cell Adhesion Molecules (CAMs) pathway. CAMs are located on the cell surface and participate in the activity of a cell binding with other cells. One of the primary features of cancer cells is

uncontrolled growth where the cells are immune to density-dependent inhibition. Normal cells, for example, stop dividing when the medium gets crowded with other cells. Cancer cells, on the other hand, keep on growing, forming multiple levels, even when the cell density is increased. This is mainly due to the malfunctioning in CAMs, which has been shown to play an important role in cancer progression [52] and disrupting important signal-transduction pathways [53]. Specifically, CAMs have been shown to be involved in brain [54], bladder [55], breast [56], liver [57], lung [58] and thyroid [59] cancer; the cancer data sets where the proposed system found the CAM pathway as significantly activated.

Other pathways that need to be emphasized are Citrate (TCA/tricarboxylic acid) cycle, Complement and coagulation cascade and Adipocytokine signaling pathways that are found to be significantly active in 7 cancer data sets out of 12 by BPA (and not included in the analysis list utilized by SPIA). Citrate cycle, also known as the tricarboxylic acid cycle (TCA cycle) or the Krebs cycle, is a part of cellular respiration. It is a series of chemical reactions used by all aerobic organisms to generate energy. Its central importance to many biochemical pathways suggests that it was one of the earliest parts of cellular metabolism to evolve [60]. A recent study identified this cycle as a cancer-specific metabolic pathway [61]. In wide range of tumor cells including the types of our datasets, it is found that a mutation causes this cycle to run in reverse. Complement and coagulation cascade pathway can be explained in two parts: the complement system is a proteolytic cascade in blood plasma and a mediator of innate immunity, a nonspecific defense mechanism against pathogens, and blood coagulation is another series of proenzyme-to-serine protease conversions. This pathway is identified as significant for breast cancer types and liver cancer in functional cancer map, functional expression profile of significant enriched KEGG pathway maps across different tumor entities assigned to various tumor classes [62]. Adipocytokine signaling pathway is positively correlated with leptin production, which is an important regulator of energy intake and metabolic rate. Leptin and adiponectin are the most abundant adipocytokines and the best-studied molecules in this class so far. Recent tumor biological findings on the role of the most prominent adipocytokines leptin and adiponectin, which are involved in tumor growth, invasion and metastasis, show the

effects of adipocytokines to brain and breast cancers [63], the cancer data sets where the BPA system found this pathway as significantly activated. In addition, the relation of adipocytokine signaling pathway to lung and liver cancers has been shown in [64,65], respectively.

Overall, we have established an effective method analyzing HTBD within the context of known biological pathways. This method finds activity of underlying biological mechanisms for given HTBD by modeling gene interactions and suggest hypothesis-based new frontiers in life sciences.

Future directions include the following. We have identified 15 pathways that have been found to be significant in at least half of the datasets; therefore, these pathways potentially represent mechanisms common to different cancer types. Besides, we can analyze the effects of these pathways on other diseases such as pathogenic diseases, deficiency diseases or hereditary diseases. If the pathways are not found active in these diseases, then we can conclude that the pathways, found significant in BPA analysis, can actually exist specifically in cancer diseases.

Moreover, BPA workflow can be expanded for use with time-series data which is a collection of observations of data items obtained through repeated measurements over time. From experimental design of view, time-series are classified based on different criteria: the number of time points, the number of biological conditions and the independency between each individual time point. In real time-course microarray data, the number of time points generally ranges between 3-6 time points where more than 6 are considered as long series, so a reasonable dataset size must be provided. In addition, the cases where there are more than two groups (cancer and normal) can be handled. We can also extend the system by integrating external biological knowledge and analyze the performance of the methods with or without using external biological knowledge. Discovering the structure representing how external knowledge relates to interaction is another direction that can be included as a future work.

APPENDIX A: THE SIGNIFICANT PATHWAYS

We list the pathways found significant in the analyzed cancer data sets. The occurrence of a pathway as significant in a data set is marked with an "X".

In total, we have identified 171 pathways that have been found significant in at least one of the data sets. 15 of these pathways, given below, have been found to be significant in at least half of the data sets and therefore potentially represent mechanisms common to different cancer types.

Table A.1. List of significant pathways in at least half of the cancer data sets.

Pathway Name
Adipocytokine signaling pathway
Alanine and aspartate metabolism
B cell receptor signaling pathway
Biosynthesis of steroids
Cell adhesion molecules (CAMs)
Citrate cycle (TCA cycle)
Complement and coagulation cascades
Cysteine metabolism
Drug metabolism - cytochrome P450
Fc epsilon RI signaling pathway
Leukocyte transendothelial migration
O-Glycan biosynthesis
Small cell lung cancer
Tight junction
Valine, leucine and isoleucine degradation

Table A.2. The pathways found significant in the analyzed cancer data sets.

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Cell adhesion molecules (CAMs)	X	X	X		X		X	X	X		X		8
Citrate cycle (TCA cycle)		X	X		X	X	X	X				X	7
Complement and coagulation cascades	X	X			X	X	X	X	X				7
Adipocytokine signaling pathway		X	X	X	X		X	X	X				7
Biosynthesis of steroids	X				X	X	X	X	X				6
Alanine and aspartate metabolism		X	X		X	X	X	X					6
Cysteine metabolism	X			X	X			X	X	X			6

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Valine, leucine and isoleucine degradation	X	X			X	X	X	X					6
O-Glycan biosynthesis		X	X			X	X	X	X				6
Drug metabolism - cytochrome P450			X		X	X	X	X	X				6
Tight junction	X	X			X	X	X		X				6
B cell receptor signaling pathway	X		X	X	X		X	X					6
Fc epsilon RI signaling pathway	X	X	X				X	X	X				6
Leukocyte transendothelial migration	X	X	X		X		X		X				6

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Small cell lung cancer	X	X	X		X		X					X	6
Purine metabolism		X	X				X	X	X				5
Pyrimidine metabolism			X		X	X	X	X					5
Glycine, serine and threonine metabolism		X	X		X	X	X						5
N-Glycan degradation	X		X		X		X	X					5
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis		X	X		X	X			X				5
Arachidonic acid metabolism	X	X			X			X	X				5
PPAR signaling pathway					X	X		X	X			X	5

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
MAPK signaling pathway	X	X			X	X			X				5
p53 signaling pathway	X	X			X			X	X				5
TGF-beta signaling pathway	X	X			X		X		X				5
ECM-receptor interaction	X	X			X	X		X					5
Natural killer cell mediated cytotoxicity	X	X			X		X	X					5
Circadian rhythm		X	X		X		X	X					5
Long-term potentiation		X	X		X			X	X				5
Taste transduction	X	X	X	X			X						5

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Insulin signaling pathway	X	X			X		X		X				5
Melanogenesis	X	X	X		X				X				5
Epithelial cell signaling in Helicobacter pylori infection	X	X	X			X			X				5
Colorectal cancer	X	X				X	X					X	5
Pancreatic cancer	X	X	X			X			X				5
Pentose and glucuronate interconversions					X		X	X	X				4
Fatty acid metabolism					X		X	X	X				4

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Urea cycle and metabolism of amino groups			X			X	X	X					4
Arginine and proline metabolism	X	X		X		X							4
Taurine and hypotaurine metabolism	X				X			X	X				4
Nucleotide sugars metabolism		X			X		X		X				4
Chondroitin sulfate biosynthesis		X					X	X	X				4

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Glycosphingolipid biosynthesis - ganglio series	X	X						X	X				4
Propanoate metabolism	X	X					X		X				4
Vitamin B6 metabolism	X				X				X			X	4
Nicotinate and nicotinamide metabolism		X			X			X	X				4
Calcium signaling pathway	X	X					X				X		4
Wnt signaling pathway						X	X	X	X				4

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Axon guidance	X	X			X				X				4
Antigen processing and presentation	X	X					X	X					4
Olfactory transduction		X			X		X	X					4
GnRH signaling pathway		X	X	X			X						4
Type II diabetes mellitus			X	X	X		X						4
Alzheimer's disease		X			X		X	X					4
Amyotrophic lateral sclerosis (ALS)		X		X			X	X					4

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Pathogenic Escherichia coli infection - EHEC				X	X	X	X						4
Pathogenic Escherichia coli infection - EPEC	X				X	X	X						4
Chronic myeloid leukemia	X	X	X						X				4
Non-small cell lung cancer	X	X					X					X	4
Glycolysis / Gluconeogenesis					X			X	X				3
Pentose phosphate pathway			X					X	X				3

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Galactose metabolism			X		X			X					3
Fatty acid biosynthesis	X	X					X						3
Fatty acid elongation in mitochondria						X		X	X				3
Glutamate metabolism					X		X	X					3
Methionine metabolism		X					X	X					3
Lysine biosynthesis	X	X					X						3
Histidine metabolism			X				X	X					3
N-Glycan biosynthesis	X				X				X				3

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Glycerolipid metabolism	X				X				X				3
Glycerophospholipid metabolism		X					X		X				3
alpha-Linolenic acid metabolism				X			X	X					3
Sphingolipid metabolism	X				X		X						3
Pyruvate metabolism		X			X		X						3
One carbon pool by folate	X				X	X							3
Methane metabolism		X						X	X				3
Sulfur metabolism			X				X	X					3

	blad-der	brain	brain	breast	breast	colon	liver	liver	lung	ovar-ian	thy-roid	thy-roid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Oc-cur-rence
Biosynthesis of unsaturated fatty acids							X	X			X		3
Proteasome		X	X					X					3
Protein export							X	X		X			3
ErbB signaling pathway	X	X			X								3
Cell cycle	X					X			X				3
Apoptosis	X	X							X				3
VEGF signaling pathway	X		X				X						3
Focal adhesion	X	X							X				3
Toll-like receptor signaling pathway			X			X			X				3

	blad-der	brain	brain	breast	breast	colon	liver	liver	lung	ovar-ian	thy-roid	thy-roid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Oc-cur-rence
T cell receptor signaling pathway		X			X				X				3
Long-term depression		X					X			X			3
Parkinson's disease		X	X				X						3
Vibrio cholerae infection					X		X		X				3
Endometrial cancer	X	X	X										3
Glioma		X			X							X	3
Prostate cancer	X	X						X					3
Thyroid cancer		X				X						X	3
Melanoma	X	X										X	3
Asthma		X		X			X						3
Autoimmune thyroid disease			X				X	X					3

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Allograft rejection		X	X				X						3
Fructose and mannose metabolism					X			X					2
Ascorbate and aldarate metabolism		X					X						2
Caffeine metabolism				X						X			2
Bisphenol A degradation	X				X								2
Aminophosphonate metabolism							X		X				2
Starch and sucrose metabolism					X				X				2
Inositol phosphate metabolism		X				X							2

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Ether lipid metabolism					X			X					2
Glycosphingolipid biosynthesis - lacto and neolacto series	X								X				2
Glycosphingolipid biosynthesis - globo series			X			X							2
3-Chloroacrylic acid degradation							X	X					2
Reductive carboxylate cycle (CO2 fixation)		X			X								2
Biotin metabolism							X	X					2

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Retinol metabolism		X			X								2
Terpenoid biosynthesis					X	X							2
Nitrogen metabolism		X						X					2
Alkaloid biosynthesis II		X			X								2
Aminoacyl-tRNA biosynthesis			X	X									2
Metabolism of xenobiotics by cytochrome P450	X					X							2
Drug metabolism - other enzymes						X	X						2
RNA polymerase			X					X					2

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Phosphatidylinositol signaling system		X			X								2
Neuroactive ligand-receptor interaction			X	X									2
Regulation of autophagy	X								X				2
mTOR signaling pathway		X	X										2
Notch signaling pathway		X			X								2
Adherens junction		X							X				2
Jak-STAT signaling pathway	X	X											2

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Hematopoietic cell lineage			X					X					2
Maturity onset diabetes of the young		X					X						2
Renal cell carcinoma							X		X				2
Basal cell carcinoma		X							X				2
Acute myeloid leukemia	X	X											2
Primary immunodeficiency				X			X						2
Synthesis and degradation of ketone bodies								X					1
Bile acid biosynthesis						X							1

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Ubiquinone and menaquinone biosynthesis					X								1
C21-Steroid hormone metabolism								X					1
Oxidative phosphorylation		X											1
Valine, leucine and isoleucine biosynthesis		X											1
Tyrosine metabolism				X									1
Phenylalanine metabolism					X								1

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Tryptophan metabolism	X												1
beta-Alanine metabolism									X				1
Selenoamino acid metabolism							X						1
Cyanoamino acid metabolism					X								1
D-Glutamine and D-glutamate metabolism									X				1
Glutathione metabolism							X						1

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Aminosugars metabolism			X										1
Glycosaminoglycan degradation							X						1
Keratan sulfate biosynthesis			X										1
Heparan sulfate biosynthesis									X				1
Glyoxylate and dicarboxylate metabolism						X							1
Styrene degradation						X							1
Butanoate metabolism						X							1

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Thiamine metabolism							X						1
Pantothenate and CoA biosynthesis												X	1
Folate biosynthesis										X			1
Porphyrin and chlorophyll metabolism			X										1
Caprolactam degradation							X						1
Alkaloid biosynthesis I									X				1
Glycan structures - biosynthesis 1								X					1

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Glycan structures - degradation			X										1
ABC transporters - General							X						1
Base excision repair	X												1
Nucleotide excision repair		X											1
Cytokine-cytokine receptor interaction	X												1
SNARE interactions in vesicular transport		X											1
Renin-angiotensin system							X						1

	bladder	brain	brain	breast	breast	colon	liver	liver	lung	ovarian	thyroid	thyroid	
Pathway Name	GSE 7476	GSE 12907	GSE 15824	GSE 8977	GSE 22544	GSE 41328	GSE 14323	GSE 14520	GSE 10799	GSE 14407	GSE 3678	GSE 6004	Occurrence
Regulation of actin cytoskeleton											X		1
Dentatorubropallidoluysian atrophy (DRPLA)							X						1
Bladder cancer					X								1
Graft-versus-host disease							X						1

REFERENCES

1. Isci, S., C. Ozturk, J. Jones and H. H. Otu, "Pathway Analysis of High-throughput Biological Data within a Bayesian Network Framework", *Bioinformatics*, Vol. 27, No. 12, pp. 1667–1674, 2011.

2. Tarca, A. L., S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J. sun Kim, C. J. Kim, J. P. Kusanovic and R. Romero, "A Novel Signaling Pathway Impact Analysis", *Bioinformatics*, Vol. 25, No. 1, pp. 75–82, 2009.

3. "Wellcome Trust - The Human Genome", Website, 2000, `http://genome.wellcome.ac.uk`, accessed at June 2013.

4. "Seoul National University Biomedical Informatics - Xperanto", Website, 2005, `http://www.snubi.org/xperanto.html`, accessed at June 2013.

5. Getoor, L. and B. Taskar, *Introduction to Statistical Relational Learning*, The MIT Press, 2007.

6. "KEGG: Kyoto Encyclopedia of Genes and Genomes", Website, 1994, `http://www.genome.jp`, accessed at June 2013.

7. Friedman, N., M. Linial and I. Nachman, "Using Bayesian Networks to Analyze Expression Data", *Journal of Computational Biology*, Vol. 7, pp. 601–620, 2000.

8. Imoto, S., S. Kim, T. Goto, S. Aburatani, K. Tashiro, S. Kuhara and S. Miyano, "Bayesian Network and Nonparametric Heteroscedastic Regression for Nonlinear Modeling of Genetic Network", Vol. 1, No. 2, pp. 231–252, 2003.

9. Nam, D. and S.-Y. Kim, "Gene-set Approach for Expression Pattern Analysis", *Briefings in Bioinformatics*, Vol. 9, No. 3, pp. 189–197, 2008.

10. Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, "Gene-set Enrichment Analysis: A Knowledge-based Approach for Interpreting Genome-wide Expression Profiles", *Proceedings of the National Academy of Sciences*, Vol. 102, No. 43, pp. 15545–15550, Oct. 2005.

11. Alexa, A., J. Rahnenführer and T. Lengauer, "Improved Scoring of Functional Groups from Gene Expression Data by Decorrelating GO Graph Structure", *Bioinformatics*, Vol. 22, No. 13, pp. 1600–1607, July 2006.

12. Lu, Y., R. Rosenfeld, I. Simon, G. Nau and Z. Bar-Joseph, "A Probabilistic Generative Model for GO Enrichment Analysis", *Nucleic Acids Research*, Vol. 36, No. 17, p. e109, 2008.

13. Bauer, S., J. Gagneur and P. Robinson, "Going Bayesian: Model-based Gene-set Analysis of Genome-scale Data", *Nucleic Acids Research*, 2010.

14. Sales, G., E. Calura, D. Cavalieri and C. Romualdi, "Graphite - A Bioconductor Package to Convert Pathway Topology to Gene Network", *BMC Bioinformatics*, Vol. 13, p. 20, 2012.

15. Edwards, D., L. Wang and P. Sørensen, "Network-enabled Gene Expression Analysis", *BMC Bioinformatics*, Vol. 13, p. 167, 2012.

16. Li, J., L. Wang, L. Xu, R. Zhang, M. Huang, K. Wang, J. Xu, H. Lv, Z. Shang, M. Zhang, Y. Jiang, M. Guo and X. Li, "DBGSA: A Novel Method of Distance-based Gene-set Analysis", *Journal of Human Genetics*, Vol. 57, No. 10, pp. 642–653, 2012.

17. Wang, P., S. Hwang, R. Kincaid, C. Sullivan, I. Lee and E. Marcotte, "RIDDLE: Reflective Diffusion and Local Extension Reveal Functional Associations for Unannotated Gene-sets via Proximity in a Gene Network", *Genome Biology*, Vol. 13, No. 12, p. R125, 2012.

18. Mieczkowski, J., K. Swiatek-Machado and B. Kaminska, "Identification of Pathway Deregulation Gene Expression Based Analysis of Consistent Signal Transduction", *PLoS ONE*, Vol. 7, No. 7, p. e41541, 07 2012.

19. Martini, P., G. Sales, M. S. Massa, M. Chiogna and C. Romualdi, "Along Signal Paths: an Empirical Gene-set Approach Exploiting Pathway Topology", *Nucleic Acids Research*, Vol. 41, No. 1, p. e19, 2013.

20. Drier, Y., M. Sheffer and E. Domany, "Pathway-based Personalized Analysis of Cancer", *Proceedings of the National Academy of Sciences*, 2013.

21. Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu and Y. Yamanishi, "KEGG for Linking Genomes to Life and the Environment", *Nucleic Acids Research*, Vol. 36, No. Database-Issue, pp. 480–484, 2008.

22. Spirtes, P., "Directed Cyclic Graphical Representations of Feedback Models", *Proceedings of the Eleventh Conference on Uncertainty in Articial Intelligence*, pp. 491–498, 1995.

23. Hoaglin, D. C., F. Mosteller and J. W. T. (Editor), *Understanding Robust and Exploratory Data Analysis*, Wiley-Interscience, 1 edn., 2000.

24. Neapolitan, R. E., *Learning Bayesian Networks*, Prentice Hall, 2003.

25. Li, Y., L. Liu, X. Bai, H. Cai, W. Ji, D. Guo and Y. Zhu, "Comparative Study of Discretization Methods of Microarray Data for Inferring Transcriptional Regulatory Networks", *BMC Bioinformatics*, Vol. 11, p. 520, 2010.

26. Ridler, T. W. and S. Calvard, "Picture Thresholding Using an Iterative Selection Method", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 8, pp. 630–632, 1978.

27. Akaike, H., *Information Theory and an Extension of the Maximum Likelihood Principle*, Akademiai Kiado, Budapest, 1972.

28. Schwarz, G., "Estimating the Dimension of a Model", *Annals of Statistics*, Vol. 6, No. 2, pp. 461–464, 1978.

29. Silander, T., T. Roos, P. Kontkanen and P. Myllymaki, "Factorized Normalized Maximum Likelihood Criterion for Learning Bayesian Network Structures", *Proceedings of the 4th European Workshop on Probabilistic Graphical Models (PGM-08)*, pp. 257–272, 2008.

30. Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri and C. D. Bloomfield, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", *Science*, Vol. 286, pp. 531–537, 1999.

31. Bolstad, B. M., R. A. Irizarry, M. Astrand and T. P. Speed, "A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias", *Bioinformatics*, Vol. 19, No. 2, pp. 185–193, 2003.

32. Yang, Y. H., S. Dudoit, P. Luu and V. Peng, "Normalization for cDNA Microarray Data: A Robust Composite Method Addressing Single and Multiple Slide Systematic Variation", *Nucleic Acids Res*, Vol. 30, No. 4, p. e15, 2002.

33. Spiegelhalter, D. J., R. C. G. Franklin and K. Bull, "Assessment, Criticism and Improvement of Imprecise Subjective Probabilities for a Medical Expert System", *UAI*, pp. 285–294, 1989.

34. Booker, L. B. and N. Hota, "Probabilistic Reasoning about Ship Images", *UAI*, pp. 371–380, 1986.

35. Charniak, E. and R. P. Goldman, "Plan Recognition in Stories and in Life", *UAI*, pp. 343–352, 1989.

36. Charniak, E. and R. P. Goldman, "A Semantics for Probabilistic Quantifier-Free First-Order Languages, with Particular Application to Story Understanding", *IJCAI*, pp. 1074–1079, 1989.

37. Hansson, O. and A. Mayer, "Heuristic Search as Evidential Reasoning", *Proceedings of the Fifth Workshop on Uncertainty in AI*, 1989.

38. Broom, B. M. and D. Subramanian, *Computational Methods for Learning Bayesian Networks from High-throughput Biological Data*, Cambridge University Press, 2006.

39. Pearl, J., *Causality: Models, Reasoning, and Inference*, Cambridge University Press, New York, NY, USA, 2000.

40. Tarjan, R. E., "Depth-First Search and Linear Graph Algorithms", *SIAM Journal on Computing*, Vol. 1, No. 2, pp. 146–160, 1972.

41. Cooper, G. F. and E. Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks from Data", *Machine Learning*, Vol. 9, No. 4, pp. 309–347, Oct. 1992.

42. Heckerman, D. and D. M. Chickering, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data", *Machine Learning*, pp. 20–197, 1995.

43. Davison, A. C. and D. V. Hinkley, *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge, 1997.

44. Efron, B. and R. Tibshirani, *An introduction to the bootstrap*, New York, 1993.

45. Benjamini, Y. and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1, pp. 289–300, 1995.

46. Shtarkov, Y. M., "Universal Sequential Coding of Single Messages", *Problems of Information Transmission*, Vol. 23, pp. 175–186, July-September 1987.

47. Rissanen, J., "Fisher Information and Stochastic Complexity", *IEEE Transactions on Information Theory*, Vol. 42, No. 1, pp. 40–47, 1996.

48. Kontkanen, P. and P. Myllymäki, "A Linear-time Algorithm for Computing the Multinomial Stochastic Complexity", *Information Processing Letters*, Vol. 103, No. 6, pp. 227–233, 2007.

49. den Bulcke, T. V., K. V. Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. D. Moor and K. Marchal, "SynTReN: A Generator of Synthetic Gene Expression Data for Design and Analysis of Structure Learning Algorithms", *BMC Bioinformatics*, Vol. 7, p. 43, 2006.

50. Hubbell, E., W.-M. Liu and R. Mei, "Robust Estimators for Expression Analysis", *Bioinformatics*, Vol. 18, No. 12, pp. 1585–1592, 2002.

51. Silander, T., T. Roos and P. Myllymäki, "Learning Locally Minimax Optimal Bayesian Networks", *International Journal Approximate Reasoning*, Vol. 51, No. 5, pp. 544–557, 2010.

52. Okegawa, T., R. Pong, Y. Li and J. Hsieh, "The Role of Cell Adhesion Molecule in Cancer Progression and its Application in Cancer Therapy", *Acta Biochimica Polonica*, Vol. 51, No. 2, pp. 445–57, 2004.

53. Cavallaro, U. and G. Christofori, "Cell Adhesion and Signalling by Cadherins and Ig-CAMs in Cancer", *Nature Reviews Cancer*, Vol. 4, No. 2, pp. 118–32, 2004.

54. Sehgal, A., A. Boynton, R. Young, S. Vermeulen, K. Yonemura, E. Kohler, H. Aldape, C. Simrell and G. Murphy, "Cell Adhesion Molecule Nr-CAM is Overexpressed in Human Brain Tumors", *International Journal of Cancer*, Vol. 76, No. 4, pp. 451–8, 1998.

55. Griffiths, T., I. Brotherick, R. Bishop, M. White, D. McKenna, C. Horne, B. Shenton, D. Neal and J. Mellon, "Cell Adhesion Molecules in Bladder Cancer: Soluble

Serum E-cadherin Correlates with Predictors of Recurrence", *British Journal of Cancer*, Vol. 74, No. 4, pp. 579–84, 1996.

56. Li, D. and Y. Feng, "Signaling Mechanism of Cell Adhesion Molecules in Breast Cancer Metastasis: Potential Therapeutic Targets", *Breast Cancer Research Treat*, Vol. 128, No. 1, pp. 7–21, 2011.

57. Paschos, K., D. Canovas and N. Bird, "The Role of Cell Adhesion Molecules in the Progression of Colorectal Cancer and the Development of Liver Metastasis", *Cell Signal*, Vol. 21, No. 5, pp. 665–74, 2009.

58. Hase, T., M. Sato, K. Yoshida, L. Girard, Y. Takeyama, M. Horio, M. Elshazley, T. Oguri, Y. Sekido, D. Shames, A. Gazdar, J. Minna, M. Kondo and Y. Hasegawa, "Pivotal Role of Epithelial Cell Adhesion Molecule in the Survival of Lung Cancer Cells", *Cancer Science*, Vol. 102, No. 8, pp. 1493–500, 2011.

59. Chaker, S., I. Kak, C. Macmillan, R. Ralhan and P. Walfish Md, "Activated Leukocyte Cell Adhesion Molecule (ALCAM/CD166) is a Marker for Thyroid Cancer Aggressiveness and Disease-free Survival", *Thyroid*, 2012.

60. Lane, N., *Life Ascending: The Ten Great Inventions of Evolution*, W.W. Norton, June 2009.

61. Mullen, A. R., W. W. Wheaton, E. S. Jin, P.-H. Chen, L. B. Sullivan, T. Cheng, Y. Yang, W. M. Linehan, N. S. Chandel and R. J. DeBerardinis, "Reductive Carboxylation Supports Growth in Tumour Cells with Defective Mitochondria", *Nature*, Vol. 481, No. 7381, pp. 385–8, 2012.

62. Krupp, M., T. Maass, J. U. Marquardt, F. Staib, B. Tobias, R. König, S. Biesterfeld, P. R. Galle, A. Tresch and A. Teufel, "The Functional Cancer Map: A Systems-level Synopsis of Genetic Deregulation in Cancer", *BMC Medical Genomics*, Vol. 4, p. 53, 2011.

63. Lang, K. and J. Ratke, "Leptin and Adiponectin: New Players in the Field of Tumor Cell and Leukocyte Migration", *Cell Communication Signal*, Vol. 7, 2009.

64. Hegyi, K., K. Fulop, K. Kovacs, S. Toth and A. Falus, "Leptin-induced Signal Transduction Pathways", *Cell Biology International*, Vol. 28, No. 3, pp. 159–69, 2004.

65. Kelesidis, I., T. Kelesidis and C. Mantzoros, "Adiponectin and Cancer: A Systematic Review", *British Journal of Cancer*, Vol. 94, No. 9, pp. 1221–5, 2006.