National College of Ireland

Gene Expression Analysis using Bayesian Networks for Breast Cancer Prognosis

MSc Research Project
Data Analytics

Tanvi Shirke
x15006255

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Tanvi Shirke
Student ID:	x15006255
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Vikas Sahni
Submission Due Date:	29/08/2016
Project Title:	Gene Expression Analysis using Bayesian Networks for Breast Cancer Prognosis
Word Count:	4853

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	21st December 2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Gene Expression Analysis using Bayesian Networks for Breast Cancer Prognosis

Tanvi Shirke
x15006255
MSc Research Project in Data Analytics

21st December 2016

Abstract

Hypothesis testing using Bayesian networks has been proven time and again to be very useful for various applications. One of these application areas is gene expression analysis. Gene expression analysis using Bayesian networks is widely researched topic since the early '90s. Gene expression can be used for prognosis of various diseases including cancer. This paper proposes modeling gene expression data using Bayesian networks for breast cancer prognosis with the help of DNA microarray data. Gene expression data has been used to build a Bayesian Network to study gene regulation in tumor samples. The model has been built using Grow-Shrink algorithm, Hill Climbing algorithm and Incremental Association Markov Blanket algorithm. The Markov blanket of the outcome of the Bayesian network can assist with breast cancer prognosis as well help in deciding the right therapy for patients.

Contents

1 Introduction

Genes contain information for the synthesis of proteins and other functional products used by the body. The process of reading this information from the gene and synthesizing a functional product is known as gene expression. This process is divided into two main parts: transcription and translation. Transcription is a process where DNA is copied into RNA. In some cases the non-coding RNA such as ribosomal RNA (rRNA) and transfer RNA (tRNA) is the finished product. But the messenger RNA (mRNA) carries protein sequences that can be used to synthesize proteins. This process is known as translation. Gene regulation is a process which is responsible for regulating gene expression. The cells in an organism hold the same genomic data but gene regulation can cause the protein makeup to widely differ. The process of gene expression is made up of many steps. Gene regulation can control mechanisms at any one of these steps to increase or decrease the synthesizing of products. mRNA transcription is the most common mechanism where regulation can occur.

Gene expression levels can be measured and analyzed using DNA microarrays. DNA microarray contains spots of DNA strands. Each spot contains multiple DNA strands, each with a unique sequence. DNA microarrays allow measuring thousands of mRNA sites simultaneously. Analysis of microarray data can assist in understanding the underlying mechanisms of gene expression and help in prognosis. Clustering algorithms are frequently used for analyzing microarray data. An attempt to identify correlated gene expression patterns can be made using clustering algorithms. Co-regulated genes can be discovered using clustering as well. Other than clustering, Bayesian networks have proven to be useful for analyzing microarray data (Friedman et al.; 2000). The purpose of this study is to model microarray data using Bayesian networks for breast cancer prognosis.

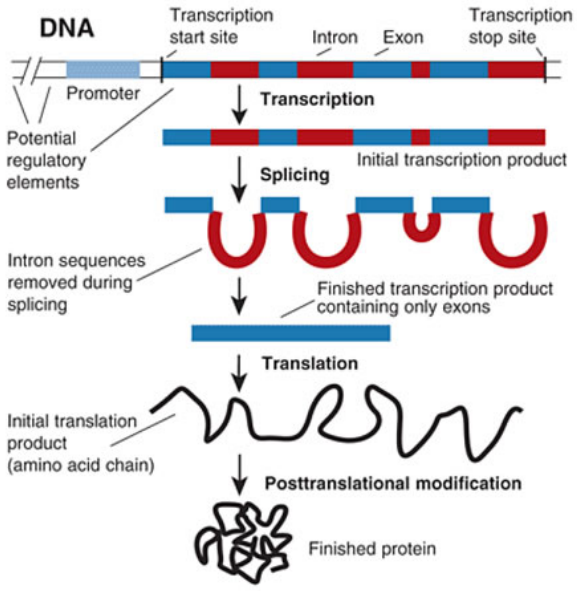

Figure 1: Gene expression mechanism: transcription and translation (Mandal; 2009)

1.1 Bayesian Networks

A Bayesian network is a Directed Acyclic Graph (DAG) that displays a set of conditionally independent random variables. The DAG is determined by two components: nodes and directed edges. Each node in the graph represents a random variable. The edges between the nodes are the probabilistic dependencies between the variables. A Bayesian network can be described with a simple condition that given the state of its parents, a child node is independent of all non-descendent nodes. Bayesian networks are commonly used as decision support models as they can handle uncertain data (Gevaert et al.; 2006). Bayesian networks can reveal dependencies of various genes and other properties of the transcription process. They can also represent a structure of expression levels of genes. They can be used to define activities that consists of locally interacting components with directly dependent values (Friedman et al.; 2000). As there are successfully attempted applications and uses of algorithms of Bayesian networks, applying them to microarray data is simpler.

1.2 Breast Cancer Prognosis

Breast cancer is a leading cause of death in women. There are 2816 new cases per year and 689 deaths occur due to breast cancer in women in Ireland (Registry; 2016). Breast cancer can be divided into 3 subtypes based on microarray analysis of breast cell lines. The subtypes are Luminal, HER-2/neu and Triple Negative type (Sorlie et al.; 2001). These subtypes of cancer are caused by over expression of various genes. A large number of genes can be considered as protein oncogenes for breast cancer. Analyzing these genes and their expression can assist in diagnosis, prognosis and even choosing the best possible therapy option for breast cancer patients.

1.3 Research Question

The purpose of this study is to analyze gene expression data using Bayesian networks for breast cancer prognosis.

1.4 Report Structure

The following section reviews research available for gene expression analysis, its application for breast cancer prognosis and the use of Bayesian networks for the analyzing gene expression. The next section describes the methodology used in this paper to build a model. This section is followed by evaluation of the methods and models used in the paper. Finally, the paper is concluded and future work has been proposed.

2 Related Work

2.1 Data Collection

Mutations in genes are a cause to a number of diseases. It is difficult to test these mutations as genes have a large number of regions which can mutate. For example, a leading cause of breast cancer, the BRCA1 gene can have around 800 mutations (Heller; 2002). These mutations can be studied with the help of DNA microarray. Heller (2002)

reviews the DNA microarray tools and applications of these tools. It talks about the experiments that can be conducted using microarray data. The use of this can determine the cause of disease that are a result of mutations. Although it concludes that DNA microarray technology can advance in to a diagnostic tool for diseases, which features can be developed for diagnostics is not mentioned.
Edgar (2002) reviews Gene Expression Omnibus, the NCBI repository that stores gene expression data. Microarray data is stored and retrieved from GEO datasets. These datasets are used by experiments and are made publicly available. GEO also facilitates tools to query and analyze the microarray data. It gives a good introduction to the GEO repository but it does not state its limitations or properly explain future developments that have been planned. This research proposes the use of DNA microarray data retrieved from GEO datasets. The datasets contain numerous genes and their expression levels that have been used for various breast cancer experiments. GEO is a reliable and simplified source of data (Edgar; 2002).
Microarray data of genes can be very huge with thousands of values. It is important to be able to select subsets of data from microarray data for accurate calculation and analysis. Typically, existing methods select the top ranked genes after ranking them according to their expressions of phenotypes. DING (2005) proposes a Minimum Redundancy Maximum Relevance (MRMR) method for selecting subsets from data. MRMR selects genes that represent broader characteristics of phenotypes and reduce noise to improve analysis. The paper contains efficient visualizations to represent the results and states the advantages and limits of the proposed method.

2.2 Gene Expression Analysis for Breast Cancer

Breast cancer can be caused by the over expression of certain genes. These genes can be measured using DNA microarray technology. This technology produces a large data of gene expression and other genomic profiles. If appropriate and effective statistical models and data mining algorithms are applied to this data, the analysis can be helpful for breast cancer prognosis.
(n.d.) proves that gene expression analysis of breast cancer cells can help improve prognosis. The paper analyses 189 invasive breast carcinomas and uses three published gene expression datasets as well. After several two-sided statistical tests, the results showed high gene expression grade index in cases with histologic 2 tumor. This can help with identifying risks of recurrence of breast cancer and improves prognosis. The article was published in 2006 and better technologies have developed in the following decade to back up their claim of improvement in prognosis.
Mehta (2009) is a PhD thesis that conducts gene expression analysis of breast cancer cells. Gene profiles of 104 tumors and 7 normal cells are used. GEO datasets are also used to compliment the in house data. Quality checks are conducted on the data using median intensity test, % array outliers and % signal outliers. Clustering algorithms are used to analyze gene expression data. Microarray data is first normalized using MAS5 technique by Affymetrix and then analyzed using hierarchical clustering and K-means clustering. Genes are identified using hierarchical clustering and t-tests are conducted. Gene lists are produced and compared with each other. These genes are then mapped using GenMAPP. A Z-score is obtained with further calculation. The Kaplan - Meir function is used to estimate the survival function. The thesis results in successfully identifying gene expressions associated with different sub-types of breast cancer and classifying them

according to these sub-types. The results of in house data was compared with the results of analysis of GEO datasets and were found to be common. A model that can be used for breast cancer prognosis is also developed. The results of the analysis are displayed using effective visualization graphs and diagrams. The results are explained in detail. The thesis is very detailed and all the measures taken during the research are mentioned. The research conducted in Mehta (2009) can prove to be of importance for further research in the same area and can help develop prognostic tools for breast cancer.
Gene expression analysis can not only help with breast cancer prognosis but also help with identifying appropriate treatment for patients. Clustering can help classify patients that require similar treatment. van 't Veer et al. (2002) analyzed 117 patient with breast cancer and conducted DNA microarray analysis. Supervised learning techniques were applied for classification. Gene expression signatures were identified that could predict poor prognosis in patients. This research can perform better than clinical parameters used during 2001 for disease diagnosis. The analysis and methods could also help in determining patients that could be benefited by adjuvant therapy. The paper does not explain steps in details and is therefore quite complicated. Research can be deemed helpful for cancer patients.
Van't Veer et al. (2002) talks about using gene expression analysis for predicting good and poor prognosis for breast cancer. The research identified 70 genes which can cause early distant metastasis in breast cancer patients. A study was conducted with patients in the Netherlands Cancer Institute. The research showed that breast cancer possesses the property to metastasize early and it is an inherent property. This claim contradicts the previously accepted idea of metastasizing occurring during late multistep tumorigenesis. The paper develops a model for predicting poor prognosis based on microarray data and identifies genes that can cause distant metastasis.

2.3 Bayesian Networks for Gene Expression Analysis

The use of Bayesian networks for gene expression analysis has proven to be effective by various researchers for the past decade. The most efficient research done on this topic is the paper by Friedman et al. (2000). Friedman et al. (2000) propose the Bayesian net- work framework for determining the interaction between genes. Gene expression data is used for this framework and statistical dependencies are mapped. Mining gene expression data from microarray datasets can be very difficult. These datasets are very large in size and capture hundreds if not thousands of genes at once. Friedman et al. (2000) explains a method of uncovering useful gene expression data from micro array datasets using learning algorithms of Bayesian networks. The paper also explains how interactions between genes can be represented using Bayesian networks. These methods are applied to microarray data. The paper was published in 2000 when genomic research technologies were under development. Although technologies have vastly changed and developed, the research is still relevant. The paper has served as a reference for developing many algorithms for Bayesian networks especially applied for gene expression analysis. Not only does Friedman et al. (2000) state methods to map gene expression data using learning Bayesian networks, it also performs biological analysis of the data and results using Markov relations and order relations. The data in this paper is effectively visualized. The explanation of methods, techniques and algorithms is simplified and explain using appropriate equations wherever necessary. Friedman et al. (2000) explains and makes use of various probability distributions and models as well as different types of variables.

Bayesian networks' application for gene expression analysis is effectively explained and modelled in this paper.
A few advantages of using Bayesian networks are:

- A statistical hypothesis is directly related to gene expression levels with the help of DAGs.
- Type I and Type II errors can be eliminated using Bayesian networks.
- Effective algorithms for Bayesian networks are easily available as they have been researched for a long time.
- They use Markov and other statistical models and stochastic elements and processes can be introduced using Bayesian networks (Sprites et al.; 2000).

Sprites et al. (2000) reviews existing algorithms and models for analyzing microarray data. They also mention shortcomings of data mining algorithms and conclude that while generalizing algorithms a lot of the difficulties seemed to be ignored instead of solved.
A large amount of algorithms have been proposed for modeling Bayesian networks to gene expression data for analysis. Murphy (1999) reviews these algorithms and states that they are all part of a single family called Dynamic Bayesian Networks (DBNs). The paper mentions and reviews various ways of learning Bayesian networks. It mentions how DBNs are related to Boolean networks, Hidden Markov Models (HMM) and compares them to each other. The paper talks about various techniques of Bayesian networks but does not try to apply them to gene expression data for analysis. It does not include any results of the review and also does not conclude properly.
There are a few challenges that are faced with Bayesian network modeling. Learning Bayesian networks is complicated and time consuming. Genes can be linked in various ways and all these topologies should be ideally explored i.e. asses all sets of DAGs as there can be multiple combinations and structures. A possible solution to this problem is suggested by Djebbari (2008) is to use general purpose search algorithms like greedy hill climbing. Djebbari (2008) also suggests introducing a bias with the use of preliminary topologies to seed the search. Another challenge could be that microarrays show subtle relations between genes and not easily defined. It is difficult to model a network with genes without determining exact relationships and dependencies.
Djebbari (2008) solves all the challenges that are faced while modeling Bayesian networks by bootstrapping and customizing general purpose algorithms used for searching. The paper proposes seeded Bayesian networks for analyzing gene expression data from a microarray. It is able to successfully model a Bayesian network for Leukemia. It contains visualization of data and results that is easily understood. Calculating the accuracy of their algorithm would have highly benefited them. This paper can be deemed as a good reference for modeling Bayesian networks for breast cancer.
Jiang et al. (2014) is a paper that proposes modeling gene expression in Signal Transduction Pathways (STPs) using Bayesian networks. It states "Our central hypothesis is that the expression levels of genes that code for proteins on a signal transduction network (STP) are causally related and that this causal structure is altered when the STP is involved in cancer." The paper uses gene expression data for HER2/neu type breast cancer. STPs form an intercellular network. Information flow through this network is initiated when extracellular molecules bind to cell receptors. These STPs can be modelled as a

Bayesian network for analysis. Jiang et al. (2014) analyzes 5 STPs involved in breast cancer. These are modelled into a Bayesian network for HER2/neu type breast cancer using STPs. This study concludes that gene expression on STPs can be learned using Bayesian networks and that it is altered by tumorous tissues.
Gevaert et al. (2006) is another research that has modeled breast cancer microarray data using Bayesian Networks. They have integrated clinical data with microarray data for prognosis. The paper proposes methods for this integration: decision integration, partial integration and full integration (Gevaert et al.; 2006). The research classifies data into good and poor prognosis. Poor prognosis is the relapse of breast cancer within 5 years while good prognosis would be being disease free for at least 5 years. The results include a Bayesian network modeling the partial integration of clinical and microarray data with the Markov blanket of outcome variable. Gevaert et al. (2006) concludes that Bayesian networks can be used to model microarray data for various cancers. Clinical data can also be integrated with the microarray data to improve prognosis.

3 Methodology

3.1 Bayesian Networks

A Bayesian network consists of a set of random variables and their conditional dependencies that are graphed using a DAG. To model the data, first the state of the system should be described using random variables. In this case, the genes can be denoted as random variables. We can now attempt a joint distribution over these variables to understand its structure and features. The expression values of genes help us understand the direct or indirect dependencies between the genes and which genes can affect dependencies.
A joint distribution that satisfies the Markov property can be broken down in the product of

$$P(X) = \prod_{i=0}^{n} P(X_i \mid Pr^g(X_i)) \tag{1}$$

Here X is a set of random variables and Pr^g is a set of parents of X_i in G.
There are two types of algorithms that can be used to create a Bayesian network: constraint based algorithms and score based algorithms. Constraint based algorithms analyze the relations between variables by using conditional independence tests. A DAG can then be built according to the results of these tests. These are known as causal models (Pearl; 1988).
Score based algorithms assign a score to each variable and then learn a network using these scores. The problem with learning Bayesian networks is that given a set D = {x_1, ..., x_n} of instances of X, we have to find $B(G, \theta)$ that is the best match for D. (Friedman et al.; 2000) states a solution to this problem can be solved by the scoring function

$$S(G : D) = logP(G \mid D) = logP(D \mid G) + P(G) + C \tag{2}$$

Here C is a constant which is independent of G and

$$P(D \mid G) = \int P(D \mid G, \theta) P(\theta \mid G) d\theta \tag{3}$$

A high score can be achieved if large samples are available and if graphs of the distribution capture the exact dependencies. The hybrid networks combined with conditional

Gaussian distributions will be used. By using the scoring function, a Bayesian network can be modeled using the hill climbing algorithm.
Partial models like Markov relations and order relations can be used for analysis. In a Markov relation, variables are directly linked. Protein activations and gene expressions can be easily observed by using Markov relations. In a DAG modeling a Bayesian network, if a node is a parent of another node, the first can be said to be the cause of another. This assumption, although correct for Bayesian networks, is not correct in the case of gene expression analysis. Order relations models can be used to solve this.

3.2 Markov Blanket

Markov blanket is an important concept of Bayesian networks.A set of variables that encloses a node with it's parents and children is the Markov blanket of a variable(Gevaert et al.; 2006). Each variable which is a part of the Bayesian network is conditionally independent of every other variable given its Markov blanket. Conditional independency implies that knowing the Markov blanket of a variable keeps the probability of that variable constant even if knowledge is added to other variables in the network. It is enough to know the Markov Blanket of a variable to predict its behavior which makes it an important concept. This paper will focus on the Markov blanket of a variable for prognosis.

3.3 Software and Tools Used

- Data was stored in .csv files. Some data was also loaded from .Rdata files.
- Microsoft Excel was used for preprocessing the data. R was used to clean and subset data and select relevant genes.
- Model building for the data has also been performed in R for all the subsets of data.

4 Implementation

4.1 Data

The dataset was obtained from Integrated Tumor Transcriptome Array and Clinical data Analysis database (ITTACA). The data has been submitted as a part of van 't Veer et al. (2002). This data set consists of 98 primary breast cancers. It is further divided into various sets. The first set consists of 34 tumors with poor prognosis that is patients who developed distant metastases within 5 years. The next set contains 44 tumors with good prognosis which is patients who were disease free for 5 years. The other set consists of 18 tumors with BRCA1 carries and 2 tumors with BRCA2 carriers. These three sets were merged to get the training set for the project. The test set contains unlabeled 19 samples of tumours with the same attributes as the training set.
The mRNA expression levels of 25000 genes for every patient were analyzed using DNA microarray. The tumors were compared with a reference made by accumulating identical number of RNA from each patient. The ratios of the tumor sample were compared with this reference for measurement of expression levels. This data has been normalized and log transformed (van 't Veer et al.; 2002). Each gene contains three essential values: log10 intensity, log10 ratio and P-value. The log10 intensity corresponds to the mean intensity

of red and green channels of the probe on the microarray chip. This value determines the expression of a gene. It is it greater than 1, the gene has been overexpressed. If it is less than -1, the gene has been under expressed. Anything in between 1 and -1 is considered neutral. The log10 ratio corresponds to the mean ratio of both the red and green channels. This value indicates the fold change. P-value is the confidence level of the mean ratio differing from 1.

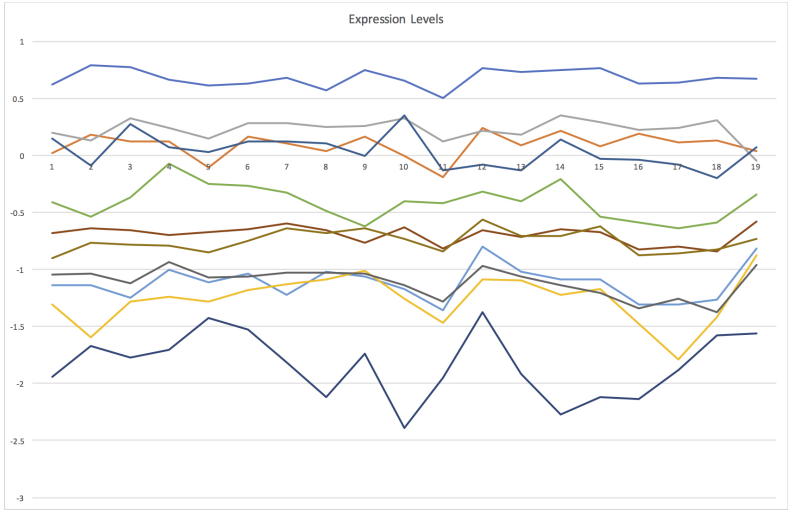

Figure 2: Data showing various expression levels of genes. Genes with expression values > 1 are over expressed.

The data contains expression levels of about 25000 genes for each tumor sample which was already background corrected, normalized and log-transformed. Microarray data is large in size and not all of it always relevant to a study. Hence, subsets of the data were selected. The data contains description of each gene that has been recorded. The first subset contains the genes with the Expressed Sequence Tag (EST). DNA microarray probes designed using ESTs are precise and can be used to measure accurate gene expression values (Nagaraj et al.; 2007). Further, selection was done by eliminating genes that did not meet the following criteria

- At least a two fold increase or decrease.
- a P-value of less than 0.01 in 4 or more tumours.

This resulted in a subset of 5000 genes. Pearson's correlation coefficient was calculated for each gene in the subset. Genes with correlation coefficient less than -0.5 and more than 0.5 were selected which resulted in approximately 230 genes. The data that resulted from these filters was used as an input for the Bayesian network.

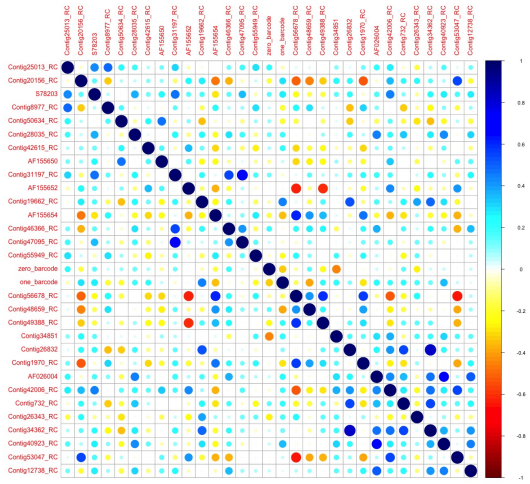

Figure 3: Plot of correlation between the genes

4.2 Model

The training set consisting of tumours with good prognosis, poor prognosis and BRCA1 tumours has been used to build the following models. This dataset contains 288 genes that are correlated. These genes were used to build a DAG. A Bayesian network that represents the gene regulatory network of the genes contained in the subset has been built. This regulatory network represents a DAG that shows the relationships between genes and their expression levels. Three types of algorithms were used to build the model: Grow-Shrink algorithm, Hill Climbing algorithm and Incremental Association Markov Blanket algorithm (Scutari; 2010).

4.2.1 Model A

This model is based on Grow Shrink algorithm. The method used to calculate the conditional independence test is Pearson's correlation. The Pearson's correlation method calculates an information theoretic distance measure which is proportional to the log-likelihood ratio test. It is related to the deviance of tested models. Initially the Markov blanket is empty. The first step of the growing phase test conditional independence between the first two genes. Then it moves on to the next genes one by one. The dependent variables (genes) are added to a set. At the end of the growing phase this set consists of all the dependent variables. The shrinking phase tests the variables from the final set which are dependent on each other i.e. they have a directed edge between them. These variables are then removed from the set to give the Markov blanket. This model has been trained using both the train sets containing tumours with good and bad prognosis as well as the set with BRCA1 tumours. It was been then fitted to the test set.

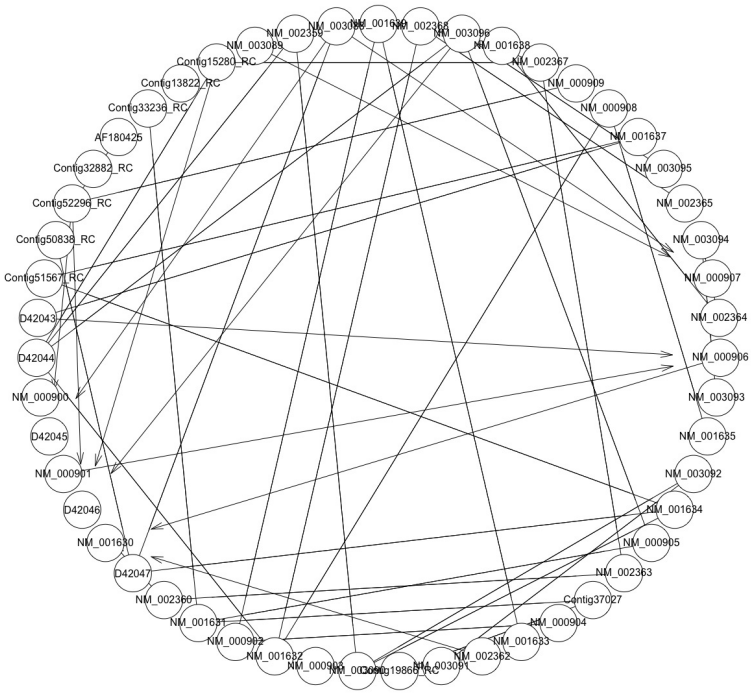

Figure 4: Bayesian network built by using Grow-Shrink algorithm

4.2.2 Model B

This model is based on the Hill climbing algorithm. The method used to determine the network score is log-likelihood scores. This model starts with identifying the children and parents of each variable. It then performs a greedy hill-climbing search on it. The edges can be redirected, deleted or formed during this search. The network which has the highest network score is selected. The model continues to do so with all the variables until the network with the highest network score has been built. The idea behind using Hill Climbing for building a Bayesian network is to find a optimum solution.

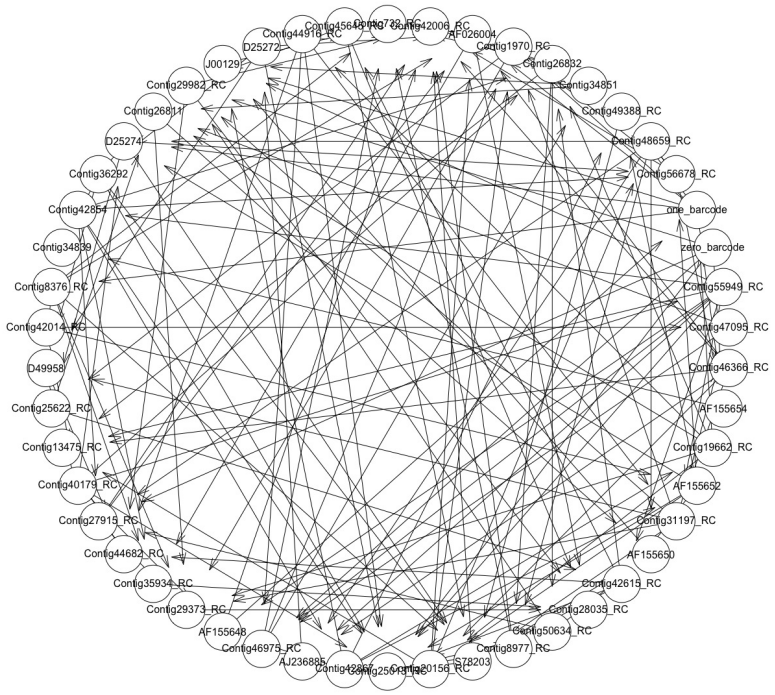

Figure 5: Bayesian network built by using Hill-Climbing algorithm

4.2.3 Model C

This model is based on Incremental Association Markov Blanket algorithm. It is a constraint based algorithm. The conditional independencies were calculated using mutual information method.The algorithm contains two phases of selection. The first phase consists of a forward selection. The second phase consists of eliminating false positives. It starts with an empty set and nodes are added in by forward stepwise selection. False positives are eliminated from this set by checking if a gene is independent of the target variable.

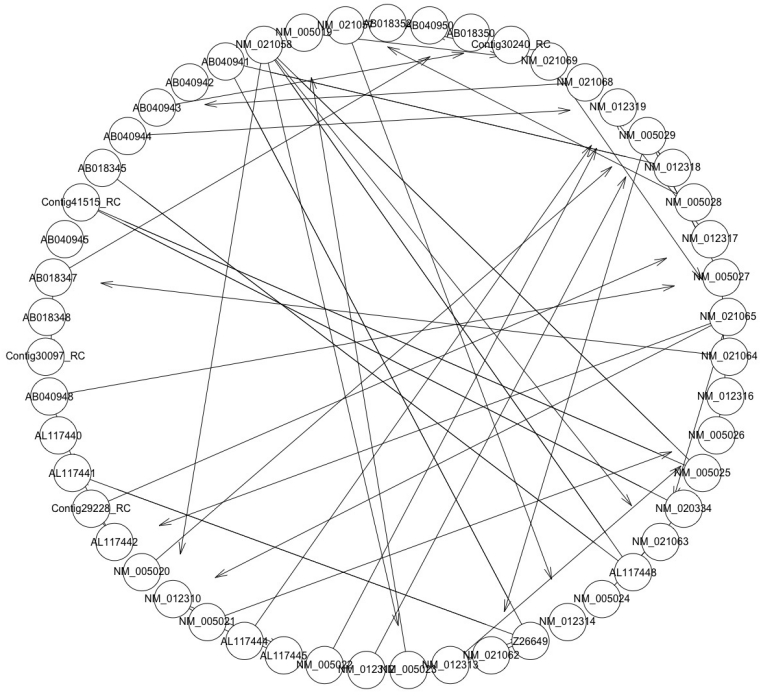

Figure 6: Bayesian network built using Incremental Association Markov Blanket algorithm

5 Evaluation

This paper built a Bayesian network using data from van 't Veer et al. (2002). Three algorithms were used to build a Bayesian network: Grow-Shrink algorithm, Hill Climbing algorithm and Incremental Association Markov Blanket algorithm. Grow-Shrink algorithm was used to build a Bayesian network. This network represents a gene regulatory network with 61 nodes. It is a partially directed graph with 12 undirected arcs and 56 directed arcs. The average Markov blanket size is 2.62.

```
Bayesian network learned via Constraint-based methods

model:
  [partially directed graph]
nodes:                                61
arcs:                                 68
  undirected arcs:                    12
  directed arcs:                      56
average markov blanket size:          2.62
average neighbourhood size:           2.23
average branching factor:             0.92

learning algorithm:                   Grow-Shrink
conditional independence test:        Pearson's Correlation
alpha threshold:                      0.05
tests used in the learning procedure: 15188
optimized:                            TRUE
```

Figure 7: Outcome of the Grow-Shrink algorithm

Hill climbing algorithm was used to build a Bayesian network. This betwork represents a gene regulatory network with 61 nodes and 232 directed arcs. The average Markov blanket size is 20.59.

```
nodes:                                61
arcs:                                 232
  undirected arcs:                    0
  directed arcs:                      232
average markov blanket size:          20.59
average neighbourhood size:           7.61
average branching factor:             3.80

learning algorithm:                   Hill-Climbing
score:                                BIC (Gauss.)
penalization coefficient:             2.178354
tests used in the learning procedure: 16590
optimized:                            TRUE
```

Figure 8: Outcome of the Hill Climbing algorithm

Incremental Association Markov Blanket algorithm is used to build a Bayesian network. This network represents a gene regulatory network with 61 nodes. This is a partially directed graph with 10 undirected nodes and 60 directed arcs. The average Markov blanket size is 2.85.

```
Bayesian network learned via Constraint-based methods

model:
  [partially directed graph]
nodes:                                61
arcs:                                 70
  undirected arcs:                    10
  directed arcs:                      60
average markov blanket size:          2.85
average neighbourhood size:           2.30
average branching factor:             0.98

learning algorithm:                   IAMB
conditional independence test:        Pearson's Correlation
alpha threshold:                      0.05
tests used in the learning procedure: 19908
optimized:                            TRUE
```

Figure 9: Outcome of the Incremental Association Markov Blanket algorithm

The Hill Climbing algorithm builds the better Bayesian network with 0 undirected arcs and a larger average Markov Blanket size.

Algorithm	Nodes	Arcs	Directed Arcs	Markov Blanket Size
Grow Shrink	61	68	56	2.62
Hill Climbing	61	232	232	20.59
IAMB	61	70	60	2.85

The resulting Bayesian network is similar to a gene regulatory network, that is, it provides an insight into gene regulation and coexpression. This is an important part of forecasting good or poor prognosis for breast cancer.
Literature related to breast cancer metastatis was researched for variables in the Markov blanket of the Bayesian network. These variables are genes that are associated with cancer. MMP9, HRASLS and RAB27B have string associations with cancer (Thangapazham et al.; 2006)(Kaneda; 2004). MMP9 is associated with angiogenesis and tumour invasion. It belongs to a family of proteases, matrix metalloproteases that degrade a path through the extra cellular matrix and stroma (Pecorino; 2012). RAB27B belongs to an oncogene family. This implies a poor prognosis for the test sample tumours. Patients with these tumours may have a relapse of breast cancer within 5 years. The results were compared with the results of Gevaert et al. (2006) as the paper has successfully been able to forecast the prognosis for the same data. The results are the same as Gevaert et al. (2006) also predicted poor prognosis for the tumours.
This model is not built for classification thus, it represents a general framework for a joint probability distribution. This model can be further improved if prior knowledge of breast cancer is used for feature selection.

6 Conclusion and Future Work

The significance of breast cancer prognosis cannot be stressed enough. Breast cancer has been a cause for numerous deaths. Building a model that can assist in prognosis can help early diagnosis of breast cancer. It can also help in choosing the right cure and therapy for a patient. This paper has attempted in building a model using log trabnsformed and normalized DNA microarray data. Hill Climbing algorithm used to build a Bayesian network showed the best possible results and predicted poor prognosis

for the test sample tumours. The successful implementation of a Bayesian network in this project can propose possible use of the model for other types of cancers. This model can also be used to decide if a patient needs adjuvant systematic therapy.
Diagnosis of breast cancer can depend on both gene expression and clinical data of a patient including patient history and family history. Integrating both the micorarray data and clinical data to build a Bayesian network can lead to more accurate results. Future work proposed in this paper would be to use Hill-Climbing algorithm for building a Bayesian network that could determine which genes form a regulatory network. Clinical data can be integrated with this model for more accurate prognosis.

7 Acknowledgements

I would like to thank my project supervisor Mr.Vikas Sahni, (School of Computing) for his helpful guidance. He has guided me well during this research project. The project would not have been possible without his help and support.

References

(n.d.).

DING, CHRISPENG, H. (2005). Minimum redundancy feature selection from microarray gene expression data, *Journal of Bioinformatics and Computational Biology* **03**(02): 185–205.

Djebbari, AmiraQuackenbush, J. (2008). Seeded bayesian networks: Constructing genetic networks from microarray data, *BMC Systems Biology* **2**(1): 57.

Edgar, R. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository, *Nucleic Acids Research* **30**(1): 207–210.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). Using bayesian networks to analyze expression data, *Journal of Computational Biology* **7**(3-4): 601–620.

Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y. and Moor, B. D. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks, *Bioinformatics* **22**(14): e184–e190.

Heller, M. (2002). *DNA MICROARRAY TECHNOLOGY: Devices, Systems, and Applications*.

Jiang, X., Neapolitan, R. and Xue, D. (2014). Modeling the altered expression levels of genes on signaling pathways in tumors as causal bayesian networks, *CIN* p. 77.

Kaneda, A. (2004). Lysyl oxidase is a tumor suppressor gene inactivated by methylation and loss of heterozygosity in human gastric cancers, *Cancer Research* **64**(18): 6410–6415.

Mandal, A. (2009). What is gene expression?
URL: *http://www.news-medical.net/life-sciences/What-is-Gene-Expression.aspx*

Mehta, J. P. (2009). *Gene expression analysis in breast cancer*, PhD thesis, Dublin City University.

Murphy, KevinMian, S. (1999). *Modelling Gene Expression Data using Dynamic Bayesian Networks*, PhD thesis, University of California, Berkley.

Nagaraj, S. H., Gasser, R. B. and Ranganathan, S. (2007). A hitchhiker's guide to expressed sequence tag (est) analysis, *Briefings in Bioinformatics* **8**(1): 6–21.
URL: *http://bib.oxfordjournals.org/content/8/1/6.abstract*

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*, Morgan Kaufmann Publishers.

Pecorino, L. (2012). *Molecular biology of cancer: mechanisms, targets, and therapeutics*, Oxford university press.

Registry, N. C. (2016).
URL: *http://www.ncri.ie/data*

Scutari, M. (2010). Learning bayesian networks with the bnlearn r package, *Journal of Statistical Software* **35**(3).

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M. and Jeffrey, S. S. e. a. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *Proceedings of the National Academy of Sciences* **98**(19): 10869–10874.

Sprites, P., Glymour, C., Scheines, R., Kauffman, S., Aimale, V. and Wimberly, F. (2000). *Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data*, PhD thesis, Carnegie Mellon University.

Thangapazham, R. L., Sharma, A. and Maheshwari, R. K. (2006). Multiple molecular targets in cancer chemoprevention by curcumin, *The AAPS Journal* **8**(3): E443–E449.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J. and Witteveen, A. T. e. a. (2002). Gene expression profiling predicts clinical outcome of breast cancer, *Nature* **415**(6871): 530–536.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T. et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer, *nature* **415**(6871): 530–536.