Analysis of time-series gene expression data with dynamic Bayesian networks

Research Plan

C.F.H.A. Gilissen

November 4, 2005

University Medical Center Nijmegen
Department of Antropogenetica
Supervisor: Michael Egmont-Petersen

Radboud University Nijmegen
Information and Knowledge Systems
Supervisors: Peter J.F. Lucas
Perry C. Groot

Contents

1 Introduction

This document intends to give an overview of a research project that will be performed at the UMC St. Radboud Nijmegen. This document will discuss the project environment, problem description, justification, activities and products of the project, as well as the organization of the project. The research is conducted in the light of a master thesis project for the education of Computing Science at the Radboud University Nijmegen.

2 Environment

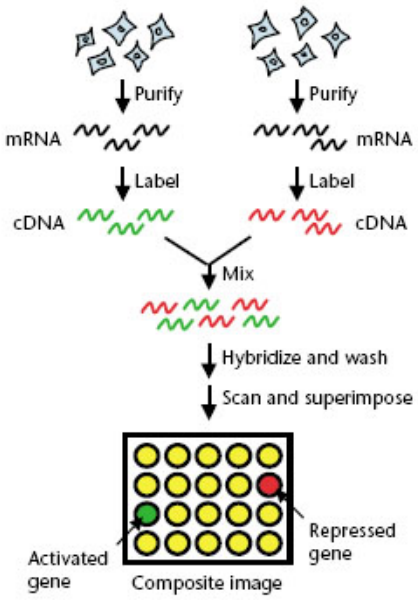
Within each living cell mRNA copies genetic information (genes) from the DNA and transports this information to sites of protein synthesis where the information is used for the production of proteins. This transcription of DNA into mRNA is known as gene expression. The amount of mRNA tells us something about the activity of certain genes in certain circumstances. Microarray facilities are designed to measure gene expression for thousands of genes in parallel.

2.1 Microarrays

The two main types of microarrays are cDNA microarrays (a.k.a. spotted microarrays) and oligonucleotide microarrays. With cDNA microarrays single strands of DNA with known sequence are immobilized on spots arranged as a grid (or array) on a (mostly glass or nylon) slide (the microarray) [4, 11, 25, 28, 29].

The most common cDNA microarray experiment is a comparative hybridization experiment where mRNA from two cells is extracted and reverse transcribed into more stable cDNA. One is a control cell that is used as a reference (control cDNA) for the cell under study (test cDNA). The cDNA is labeled with fluorescent dyes, commonly Cyanine 3 and Cyanine 5 (Cy3, Cy5) to keep the control- and test DNA apart. Now the labeled cDNA is hybridized to the microarray and undergoes competitive binding (hybridization). The cDNA that did not bind is washed away. A scanner can measure each spot on the microarray on how much of each of our red and green labels is present. This corresponds to the amount of cDNA that has bounded. By comparing the amounts of control cDNA with the test cDNA on a certain spot on the microarray, relative transcription differences can be found between the control and test DNA for the specific gene that was on that spot.

With oligonucleotide microarrays (Affymetrix, Santa Clara, CA) the labels are directly synthesized on the microarray using a photolithographic process. For this reason a comparative hybridization experiment as described above is not possible. oligonucleotide microarrays are mostly used for conducting measurements at different time intervals under changing conditions for the test probe. For instance, the first measurement occurs at a temperature of 37 degrees Celsius while a second measurement is done at 39 degrees.

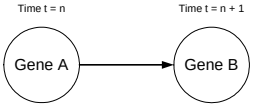


(a) The cDNA Microarray process

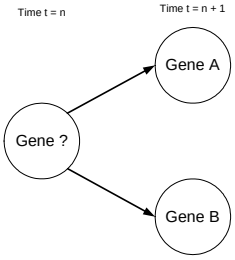
(b) A DNA Microarray, the different colours indicate relative expression of different genes

Figure 1: The microarray process

The data obtained from measuring the expression of test DNA at more than one point in time is referred to as time-series expression data. Time-series can be used for studying gene control regulation and co-regulation. When the expression of one gene influences the expression of another gene this is called control regulation. When two genes are controlled by a third gene, and thus always show expression patterns at (nearly) the same time, this is called co-regulation. Finding control regulated genes is a first step in finding regulatory pathways [5, 17].



(a) Control-regulation, gene *A* controls gene *B*



(b) Co-regulation, some gene controls genes *A* and *B* (at the same time)

Figure 2: Gene expression patterns

2.2 Analysis of Microarray Data

Microarrays have provided us with large quantities of data for finding gene functionality and gene regulatory pathways. However, extracting information from the data is not a straightforward process.

When comparing data, it must be normalized for different variances that can occur. There can be variances due

to unequal quantities of starting RNA and differences in labeling or detection efficiencies. For all these different variances different normalization techniques exist, mostly based upon statistics. For instance, ANOVA techniques are widely used [13, 14, 24].

When analyzing the (normalized) data for finding gene functionality there are some well-known problems [8, 10, 26]:

- The data contains a lot of noise which makes it difficult to analyze.
- The number of genes is very large while the number of samples is usually very small, which makes conventional statistics less reliable.
- Interest lies in gene functionality instead of mRNA levels. mRNA levels only give a partial view of gene functionality. mRNA is broken down relatively fast, while the protein it encoded for may be present for a far longer time, influencing the expression of other genes.

2.3 Methods

Several methods have been used to extract information from gene expression data. Clustering methods were one of the first to succeed in this. These methods attempt to classify genes into groups with similar expression patterns. Several different clustering methods like hierarchial clustering, K-nearest neighbour, K-means, voting and Self-Optimizing Maps [1, 6, 27] have been applied. These methods have been very useful in finding co-regulated genes. They are however less suitable for the finding of regulatory pathways.

Other promising methods employed for analyzing gene expression are Neural Networks (NN) [20] and Support Vector Machines (SVM) [2].

Relatively new methods are now based upon Bayesian Networks (BN) [8, 12]. Bayesian networks is a promising tool for analyzing expression data [8], especially when looking for gene regulatory pathways in times-series data [15, 17]. Some known advantages of Bayesian networks are:

- The mathematical foundations for Bayesian networks are well understood
- Bayesian networks provide models of causal influence
- Bayesian networks allow for a useful graphical representation
- Bayesian networks are whitebox models (contrary to NNs and SVMs) [18]

2.4 Bayesian Networks

Bayesian networks (BNs) are graphical models for reasoning under uncertainty. The nodes in the model represent variables, while arcs between these nodes represent direct dependencies between variables. The strength of the dependencies is quantified by a conditional probability distribution associated with each node. The graphical structure is often referred to as the qualitative part of the BN. In graph theory, such a network is called a Directed Acyclic Graph (DAG). If the variables are discrete, the conditional probability distribution of each node can be specified in a conditional probability table (CPT) [16, 23].

Say $\{X_1, \ldots, X_N\}$ is a set of random variables and $\{x_i, \ldots, x_n\}$ is the set of possible instantiations. The joint probability distribution is represented by $P(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_n)$. Calculation of this joint probability distribution is an exponential task. However, we can simplify the distribution by using the independence structure of the variables.

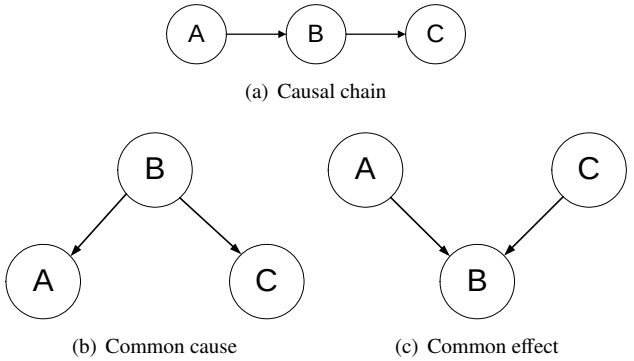
First we factorize the joint probability distribution over the values with the *chain rule*:

$$P(x_1, x_2, \ldots, x_n) = P(x_1) \times P(x_2|x_1) \times \ldots \times P(x_n|x_1, \ldots, x_{n-1}) = \prod_i^n P(x_i|x_1, \ldots, x_{i-1})$$

Now we use the independence structure by seeing that the conditional distribution of X_i depends only on his parents pa_i (the Markov assumption) and we rewrite the joint probability distribution into:

$$P(x_1, x_2, \ldots, x_n) = \prod_i^n P(x_i|pa_i)$$

Another way to simplify calculations on a Bayesian network is by making use of conditional independencies. For this we look at the following three types of networks:



(a) Causal chain

(b) Common cause

(c) Common effect

Figure 3: Conditional independencies

In Figure 3(a) we can see that given evidence on the value of B, additional evidence on the value of A will no longer influence the value of C, which has become entirely determined by B. We say C is conditionally independent of A given B.

$$A \perp\!\!\!\perp C|B$$

In Figure 3(b) we see a similar effect. Say node B stands for having a cold and is a cause for both A (= sneezing) and C (= coughing). Given evidence that a person is sneezing, the chances of the person having a cold, and thus the chances of coughing will rise. But if we already know that someone has a cold, knowing that he is sneezing will not raise our belief that he might be coughing. We say that A is conditionally independent of C given B, and vice versa.

$$P(C|A \wedge B) = P(C|B) \Rightarrow A \perp\!\!\!\perp C|B$$

In Figure 3(c) we see that given positive evidence on A will not change the chances on C. But given evidence on B, additional evidence on A will lower the chances on C. This is the opposite of conditional independence. We say A and C are conditionally dependent.

$$P(A|C \wedge B) \neq P(A|C) \Rightarrow \neg(A \perp\!\!\!\perp C|B)$$

2.4.1 Learning Bayesian Networks

Bayesian networks are not as fully applied in all sorts of businesses as one might think, considering the advantages. The reason for this is the knowledge bottleneck [16]. The domain knowledge (i.e., the qualitative and quantitative part) must somehow be available. However in lots of problems there simply is no domain knowledge available and with other problems this knowledge comes from experts in the domain. Extracting this knowledge by interviews and testing is a time consuming job. This method of Bayesian network construction is called *manual construction* [19].

A promising alternative is machine learning from domain data. With the help of datasets, an algorithm is able to construct (learn) a Bayesian network that fits the data. Learning basically consists of two different components: 1) learning the graphical structure (model selection), 2) learning the conditional distributions (CPT's; parameter estimation). Bayesian learning algorithms already exist; like the older K2 algorithm (a search and scoring algorithm) and the more recent EM and SEM algorithms [7, 19]. These algorithms all exploit the causal independencies for learning efficiently.

With a linear growth of variables, the number of possible networks grows exponentially. So, with more variables, the algorithm must consider more networks as being the right one (e.g., best fitting the data). For this reason we cannot simply feed expression data right into a learning algorithm. The time needed for considering all possibilities would leave us waiting for centuries.

2.4.2 Pre-processing

Gene expression data is not readily fit for learning Bayesian networks:

1. *Gene expression values are continues values.*
   Bayesian networks however preferably use discrete values for variables, among other things because this limits the amount of possible networks considerably. Bayesian network models for continuous variables exist, and in most cases assume a Gaussian or normal distribution for variables [18, 8] which makes the networks computationally more feasible. Whether this assumption is entirely just, remains an open issue.

   We choose however to work with discrete variables, which means that the continues microarray data must be made discrete, which causes some loss of information. The amount of information lost for a continuous variable depends on the number of discrete values we choose the variable to have and the procedure used for assigning discrete values to them. The more discrete values a variable can take, the more accurate its representation of the original continuous value of the variable, but also the more possible networks we must consider. Hence converting the continues values into discrete values is a crucial step which will determine the outcome of all subsequent steps.

2. *Gene expression data contains noise and fluctuations.*
   Gene expression data even after it has been normalized, still contains noise and fluctuations. For this reason we need to smoothen the data somewhat.

We plan to use convolution techniques to achieve both at once [5]. By finding maxima and minima expression rates, we can easily assign values to a group (for instance 'up regulation', 'normal regulation' etc.).

One positive remark that can be made about modeling regulatory pathways from a computational point of view is that given a gene it is assumed that no more than a few other genes regulate it [8].

2.4.3 Gene Selection

To limit the amount of genes we must first make a selection of the genes that seem most interesting. Several different approaches can be adapted. For instance, we could look at the expression moments of pairs of genes and find pairs where one gene always shows a change in expression right after the other gene has changed its expression. Furthermore we might use the gene ontology information and consider whether transcription of the specific genes occurs at the same locations in a cell.

2.4.4 Dynamic Bayesian Networks

When modeling time-series data, time of course plays an important role. Although Bayesian networks model causal relationships, they do not explicitly model time. An extension of Bayesian networks are dynamic Bayesian networks (DBN) that explicitly model change over time.

Within a DBN, variables are represented as different nodes for each time step. The relationships between nodes at the same time step are called intra-slice arcs, while the arcs between nodes from one to another time slice are called inter-slice arcs or temporal arcs. We assume that the state of a slice at time t depends only on the state of the previous slice, time $t-1$ (the Markov assumption). Furthermore we assume that intra-slice relationships do not change over time i.e., the network is stable [10, 15, 16, 19].

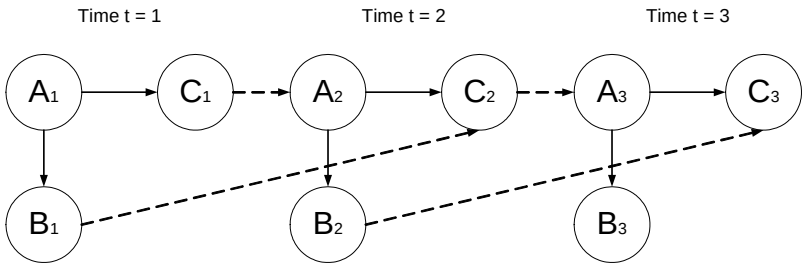


Figure 4: A dynamic Bayesian network consisting of three time slices. Inter-slice arcs are dashed; Intra-slice arcs are solid.

2.4.5 Noisy Or-Gate Models

Learning the Conditional Probability Tables (CPT) is an exponential task. If a node Y has parent nodes $X_1, X_2, \ldots, X_n$, its CPT will have 2^n entries. However, in many cases the outcome of the CPT for the child node will not be based on all available information. Such situations might approximated by a noisy OR-gate model. A binary noisy-OR model with causes $X_1, X_2, \ldots, X_n$ of Y assumes [21, 22]:

1. each of the causes X_i has a probability p_i of causing Y in the absence of all other causes.
2. the ability of each X_i to cause Y is independent of the presence of other causes.

Now the conditional probability distribution of Y can be specified with only n parameters: $p_1, p_2, \ldots, p_n$.

3 Research

The goal of this project is formulated as follows:

The research of an automated method to analyze temporal gene expression data for gene co-regulation and control regulation (regulatory pathways) with the help of dynamic Bayesian network technology and the use of expert domain knowledge.

The use of expert domain knowledge (or background information) could provide a cutting edge to the performance of a dynamic Bayesian network. For example the use of knowledge about known genes or regulatory pathways could be used to bias the Bayesian network and increase performance and accuracy.

3.1 Objectives

In order to achieve this goal we set the following objectives:

- *Find time-series gene expression data*
  In order to test algorithms for time-series analysis, we of course need such time-series. Such a time-serie should ideally consist of about 20 time steps or more. Gene expression data is widely available in public databases on the internet. It will be a matter of choice what time-series are most suitable and thus will be used.
- *Choose and adapt an appropriate gene selection algorithm*
  Because learning a Bayesian network is a very hard problem, it is important to limit the amount of variables used. For this we will employ a feature selection algorithm that will pinpoint the best candidates for our Bayesian network.
- *Choose and adapt a dynamic BN learning algorithm*
  Several kinds of dynamic Bayesian networks exist. A choice has to be made which structure is most appropriate for the given task. The same applies for the different techniques for learning dynamic Bayesian networks, depending on the chosen time-series. Furthermore we must decide which background information can be incorporated that will enhance the network. From the set of obtained networks we must choose by some criteria the best one. The finding of suitable criteria will be part of this research.
- *Implement and test these algorithms on time-series data*
  In order to test our developed method in practice, we need to test it on real data. Cross-validation will be most likely used for this.
- *Compare the results with other algorithms*
  If we want to say anything about the usefulness of the developed method, it should be compared to existing methods.

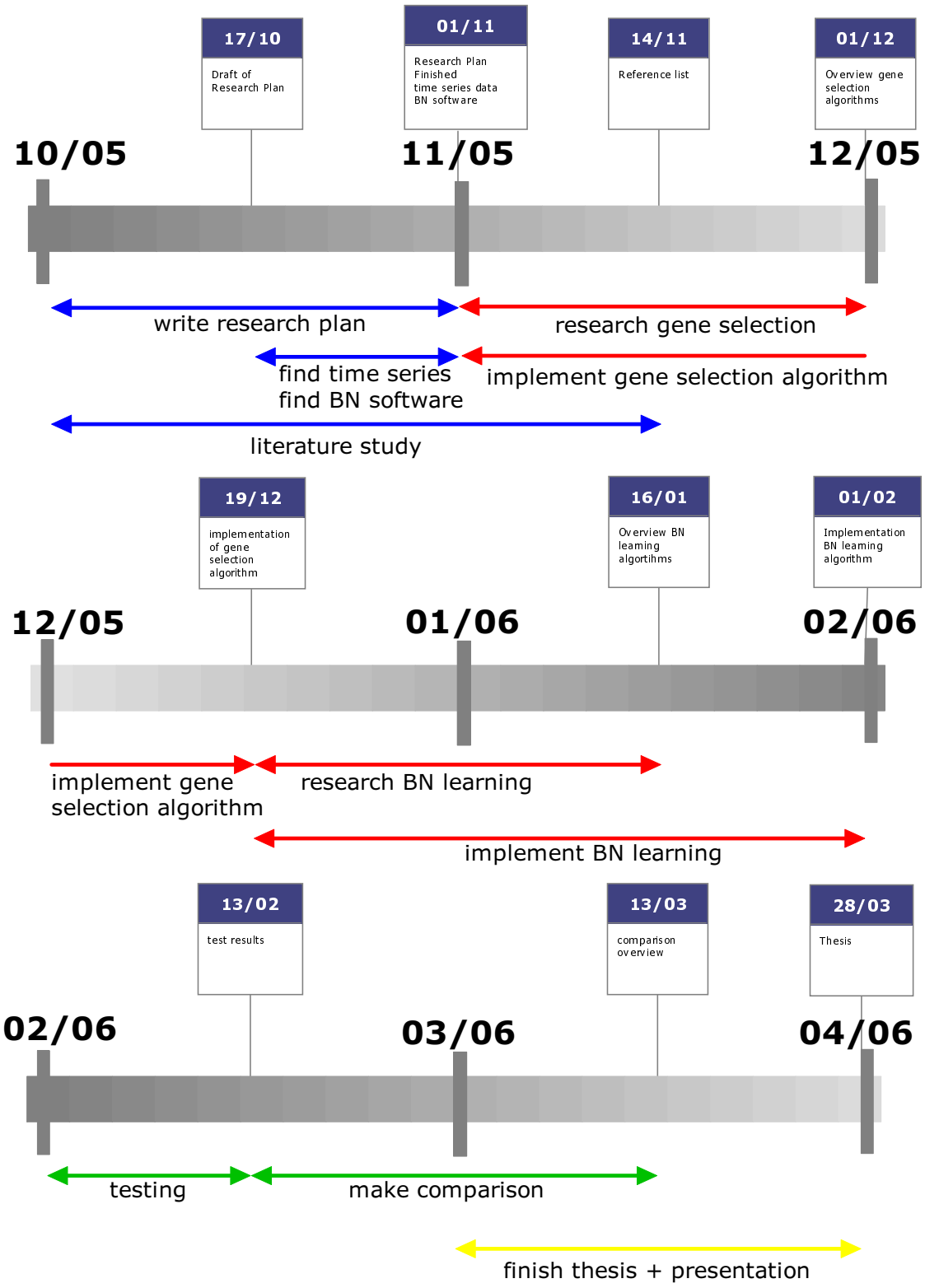
3.2 Project Activities

A distinction can be made between several activities that must be conducted:

- *Literature study*
  This activity consists of the reading of published papers and books on the research topic: microarray (time-series) data analysis, (dynamic) Bayesian networks, clustering methods etc.

- *Writing research plan*
  This activity consists of the writing of a research plan. The research plan addresses (among others) the research environment, the research question, and the research timetable.
- *Selection of suitable time-series gene regression data*
  This activity consists of the search for appropriate time-series data. The data should contain expression data from genes of homo sapiens, and there should be at least 15 samples from different time points. Public databases are available that contain such data (i.e., http://www.ncbi.nlm.nih.gov/geo/).
- *Selection of suitable BN software*
  This activity consists of the search for suitable software to construct Bayesian networks.
- *Literature study on gene selection algorithms*
  This activity consists of a study of algorithms for gene selection in time-series data.
- *Implementation of a suitable feature selection algorithm*
  This activity consists of the implementation (and if necessary alteration/improvement) of a gene selection algorithm.
- *Literature study on dynamic BN learning*
  This activity consists of a study of learning algorithms for dynamic Bayesian networks.
- *Implementation of dynamic BN learning algorithms*
  This activity consists of the implementation (and if necessary alteration/improvement) of a learning algorithm for dynamic Bayesian networks.
- *Testing of the implementations*
  This activity consists of the testing of the implementations with test data and evaluation of the results.
- *Methods comparisons*
  This activity consists of a comparison with other well-known methods for modeling time-series gene expression data.
- *Finish thesis*
  This activity consists of finishing the thesis. The writing of the thesis is a process parallel to all other activities.

3.3 Timeline


17/10
Draft of Research Plan
01/11
Research Plan Finished
time series data
BN software
14/11
Reference list
01/12
Overview gene selection algorithms
10/05
11/05
12/05
write research plan
research gene selection
find time series
find BN software
implement gene selection algorithm
literature study
19/12
implementation of gene selection algorithm
16/01
Overview BN learning algortihms
01/02
Implementation BN learning algorithm
12/05
01/06
02/06
implement gene selection algorithm
research BN learning
implement BN learning
13/02
test results
13/03
comparison overview
28/03
Thesis
02/06
03/06
04/06
testing
make comparison
finish thesis + presentation

3.4 Resources

For this project, the following information resources are available:

- *The Radboud University library*
  The library of the Radboud university contains an extensive amount of books (http://www.ru.nl/library/).
- *Publication repositories on the internet*
  The Radboud University has a lot of subscriptions to scientific publisher sites on the internet. For instance to Springer-Verlag (http://www.springeronline.com) or ScienceDirect (http://www.sciencedirect.com) which is the world's largest electronic collection of science, technology and medicine full text and bibliographic information. With these subscriptions published articles can be freely downloaded at these sites.
- *Microarray data repositories*
  On the internet, public databases exist, containing gene expression data from microarray experiments, which can be freely used. For instance the Gene Expression Omnibus at http://www.ncbi.nlm.nih.gov/geo/, or the Standford Microarray Database at http://genome-www5.stanford.edu/. The UMC St. Radboud also has subscriptions to non-public databases containing gene expression data, as well as their own database.
- *Gene maps*
  Gene encyclopedia containing information about known genes and known regulatory pathways exist and are partially freely available on the internet. For instance KEGG: Kyoto Encyclopedia of Genes and Genomes (http://www.kegg.com/). This information can be used to strengthen confidence on found regulatory pathways and for validation purposes.
- *Gene ontology database*
  The Gene ontology database (http://www.geneontology.org/) can provide us with background information on genes which can be used for learning Bayesian networks. For instance, it is highly unlikely that genes responsible for producing proteins that are always located at the Mitochondria will control regulate genes for producing proteins that are always located at the Golgi apparatus.

4 Justification

4.1 Social Relevance

Automated analysis of time-series gene expression data can reveal new information about gene regulatory pathways, which would lead to a better understanding of gene regulation and diseases related to gene regulatory problems. Already, microarray experiments are used to classify cancer cells from patients [3, 30, 31].

4.2 Scientific Relevance

Bayesian networks have a large potential for solving this kind of problems and are being investigated a lot [8, 12]. Different gene selection algorithms and BN learning algorithms are being developed to improve automated analysis of gene expression data. There also exist some models for time series Bayesian networks, but these are still under heavy development [9, 15, 17].

5 Project Organization

This project will be supervised by Peter Lucas and Perry Groot of the Radboud University Nijmegen and Michael Egmont-Petersen of the UMC St. Radboud.

5.1 Roles and Responsibilities

Within this research, the following persons participate:

Christian Gilissen	Project team, responsible for carrying out the project.
Peter Lucas	Supervisor
Perry Groot	Supervisor
Michael Egmont-Petersen	Client, Supervisor

5.2 Contact Information

Christian Gilissen
Horizonhof 8
6133 VG Sittard, The Netherlands
Phone: +31 6 200 56 356
Email: christian.gilissen@xs4all.nl

Peter Lucas
Institute for Computer and Information Sciences,
University of Nijmegen,
Toernooiveld 1
6525 ED Nijmegen, The Netherlands
Phone: +31 24 24 3652611
Email: peterl@cs.ru.nl

Perry Groot
Institute for Computer and Information Sciences,
University of Nijmegen,
Toernooiveld 1
6525 ED Nijmegen, The Netherlands
Phone: +31 24 365 2169
Email: perry.groot@science.ru.nl

Michael Egmont-Petersen
MCN - St Radboud, Dept. of Human Genetics (R. 417),
Section Cytogenetics
6500 HB Nijmegen, The Netherlands
Phone: +31 24 36 68160
Email: egmont-petersen@ieee.org

References

[1] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4):281–297, 1999.

[2] M. Brown, W. Grundy, D. Lin, N. Christianini, C. Sugnet, M. Jr, and D. Haussler. Support vector machine classification of microarray gene expression data, June 1999.

[3] S. Bulashevska. Bayesian network models for inferring cancer pathogenetic and gene regulatory pathways. Master's thesis, University Karlsruhe (TH), July 2005.

[4] J. Dopazo. Microarray data processing and analysis. *Microarray data analysis II. Kluwer Academic. Publ.*, 2002.

[5] M. Egmont-Petersen, W. de Jonge, and A. Siebes. Discovery of regulatory connections in microarray data. *Lecture Notes in Artificial Intelligence: Knowledge Discovery in Databases*, 3202:149–160, September 2004.

[6] M. Eisen, P. Spellman, D. Botstein, and P. Brown. Cluster analysis and display of genome-wide expression patterns. *Proceedings of National Academy of Science USA*, 95:14863–14867, December 1998.

[7] N. Friedman. The bayesian structural em algorithm. 1998.

[8] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data, 1999.

[9] N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. pages 139–147, 1998.

[10] L. Göransson and T. Koski. Using a dynamic bayesian network to learn genetic interactions, 2002.

[11] P. Hegde, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. Hughes, E. Snesrud, N. Lee, and J. Quackenbush. A concise guide to cdna microarray analysis. *BioTechniques*, 29(3):548–562, September 2000.

[12] D. Husmeier. Reverse engineering of genetic networks with bayesian networks. *Biochemical Society Transactions*, 31:1516–1518, December 2003.

[13] M. Kerr and G. Churchill. Statistical design and the analysis of gene expression microarray data. *Genetical Research*, 77(2):123–128, April 2001.

[14] M. Kerr, M. Martin, and G. Churchill. Analysis of variance for gene expression microarray data. *Artificial Intelligence in Medicine*, 30(3):257–281, March 2000.

[15] S. Kim, S. Imoto, and S. Miyano. Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression. *Biosystems*, 75(1-3):57–65, July 2004.

[16] K. Korb and A. Nicholson. *Bayesian Artifical Inteligence*. Chapmann & Hall/CRC Press UK, 2003.

[17] S. Li, J. Tseng, and S. Wang. Reconstructing gene regulatory networks from time-series microarray data. *Physica A: Statistical Mechanics and its Applications*, 350(1):63–69, May 2005.

[18] P. Lucas. Bayesian analysis, pattern analysis, and data mining in health care. *Current Opinion in Critical Care*, 10:399–403, 2000.

[19] P. Lucas, L. van der Gaag, and A. Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, 30:201–214, March 2004.

[20] A. Narayanan, E. Keedwell, J. Gamalielsson, and S. Tatineni. Single-layer artificial neural networks for gene expression analysis. *Neurocomputing*, 61:217–240, Oktober 2004.

[21] R. Neapolitan. *Learning Bayesian Networks*. Pearson Prentice Hall, Upper Saddle River, NJ, 2003.

[22] A. Oniśko, M. Druzdel, and H. Wasyluk. Learning bayesian network parameters from small data sets: Application of noisy-or gates. volume 27, pages 165–182, August 2001.

[23] J. Pearl. Bayesian networks. Technical Report 980002, 1998.

[24] J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32:496–501, December 2002.

[25] R. Stears, T. Martinsky, and M. Schena. Trends in microarray analysis. *Nature Medicine*, 9(1):140–145, January 2003.

[26] E. van Someren, L. Wessels, and M. Reinders. Genetic network models: A comparative study. 2001.

[27] E. Ver Loren van Themaat. On the use of learning bayesian networks to analyze gene expression data: Classification and gene network reconstruction. Master's thesis, University of Amsterdam, June 2005.

[28] A. Watson, A. Mazumder, M. Stewart, and S. Balasubramanian. Technology for microarray analysis of gene expression. *Current Opinion in Biotechnology*, 9(6):609–614, 1998.

[29] S. Zareparsi, A. Hero, D. Zack, R. Williams, and A. Swaroop. Seeing the unseen: Microarray-based gene expression profiling in vision. *Investigative Ophthalmology & Visual Science*, 45(8):2457–2462, August 2004.

[30] X. Zhou, K. Liu, and S. Wong. Cancer classification and prediction using logistic regression with bayesian gene selection. *Journal of Biomedical Informatics*, 37(4):249–259, 2004.

[31] X. Zhou, X. Wang, and E. Doughertya. A bayesian approach to nonlinear probit gene selection and classification. *Journal of the Franklin Institute*, 341(1,2):137–156, January-March 2004.