Data Mining with Decision Trees in the Gene Logic Database – A Breast Cancer Study
(HS-IDA-MD-02-207)

Neda Rahpeymai (a98nedra@student.his.se)
Department of computer science
Högskolan in Skövde, Box 408
S-54128 Skövde, SWEDEN

Masters Dissertation in bioinformatics, Spring 2002.
Supervisor: Björn Olsson

Data Mining with Decision Trees in the Gene Logic Database – A Breast Cancer Study

Neda Rahpeymai (a98nedra@student.his.se)

Abstract

Data mining approaches have been increasingly used in recent years in order to find patterns and regularities in large databases. In this study, the C4.5 decision tree approach was used for mining of Gene Logic database, containing biological data. The decision tree approach was used in order to identify the most relevant genes and risk factors involved in breast cancer, in order to separate healthy patients from breast cancer patients in the data sets used. Four different tests were performed for this purpose. Cross validation was performed, for each of the four tests, in order to evaluate the capacity of the decision tree approaches in correctly classifying 'new' samples. In the first test, the expression of 108 breast related genes, shown in appendix A, for 75 patients were used as input to the C4.5 algorithm. This test resulted in a decision tree containing only four genes considered to be the most relevant in order to correctly classify patients. Cross validation indicates an average accuracy of 89% in classifying 'new' samples. In the second test, risk factor data was used as input. The cross validation result shows an average accuracy of 87% in classifying 'new' samples. In the third test, both gene expression data and risk factor data were put together as one input. The cross validation procedure for this approach again indicates an average accuracy of 87% in classifying 'new' samples. In the final test, the C4.5 algorithm was used in order to indicate possible signalling pathways involving the four genes identified by the decision tree based on only gene expression data. In some of cases, the C4.5 algorithm found trees suggesting pathways which are supported by the breast cancer literature. Since not all pathways involving the four putative breast cancer genes are known yet, the other suggested pathways should be further analyzed in order to increase their credibility.

In summary, this study demonstrates the application of decision tree approaches for the identification of genes and risk factors relevant for the classification of breast cancer patients.

Keywords: Data mining, Decision trees, C4.5, Breast cancer

Table of contents

1. Introduction

Data mining has been defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data"(Frawley et. al. 1992). It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form which is easily comprehensible to humans.

The data mining concept has been popularly treated as a synonym of *knowledge discovery in databases* where intelligent methods are applied in order to extract data patterns (Han, 1999). However, mining the human genome to identify genetic mutations that cause complex diseases is like looking for needles in a haystack. In this process the coding regions in the genome must first be identified so that researchers can find disease-related sequences within these regions. Bioinformatics, that is, the use of information-technology and software developed for biological studies, makes it possible for researchers to look through the whole genome in order to find these genetic defects. Years of research have shown that genetic defects, whether caused by mutagens or inherited as defective gene copies, are inherent to the onset of cancer (Mort, 2000). Cancer's characteristic uncontrolled cell growth usually involves some combination of an impaired DNA repair pathway, the transformation of a normal gene to an oncogene and/or a malfunctioning tumor suppressor gene (Mort, 2000).

One of the questions that may come to mind is if mining cDNA libraries might be a promising way of searching for breast cancer related genes in order to find genetic weapons to combat the disease. The answer to this question could be 'yes' due to the fact that data mining facilitates the management of the huge amount of information (data) accessible in the cDNA libraries. A key data mining technique is considered to be classification where database tuples, acting as training samples, are analysed in order to produce a model for the given data (Kamber, 1997). This technique has several applications including disease diagnosis. A well-accepted method for classification is the induction of decision trees. A decision tree is a flow-chart-like structure consisting of internal nodes, leaf nodes, and branches (Kamber, 1997). Each internal node represents a decision, or test, on a data attribute and each outgoing branch corresponds to a possible outcome of the test. Each leaf node represents a certain class. In order to classify an unlabeled data sample, the classifier tests the attribute values of the sample against the decision tree. A path is traced from the root to a leaf node which holds the class prediction for that sample. Decision trees can easily be converted into IF-THEN rules and used for decision-making (Kamber, 1997). ID3 and C4.5 are algorithms introduced by Quinlan in 1993 for induction of decision trees.

The main goal of this project is to evaluate the induction of decision trees for identifying genes and risk factors responsible for the establishment and growth of breast cancer. Achieving this goal involves a large data set including both gene expression data and clinical data. In this project the Gene Logic database will be used, containing this kind of information from more than 6,800 samples of both diseased and healthy patients. About 2,700 of these samples are cancer related. The clinical data available in the database is shown in Table 1 and the expression data that will be used in this project is from chipset HG-U95.

The hypothesis of this project is that decision trees are a useful classification technique in order to find genes and risk factors involved in breast cancer. Genes linked to the onset or prevention of breast cancer may be found by analysis of the gene expression pattern of cells. By comparing the production of gene products in normal and diseased cells one can decide which genes are over-expressed in the cancer cells. An example gene expression profile is shown in Figure 1. The over-expressed genes may then in different ways be involved in breast cancer and may therefore be possible drug targets.

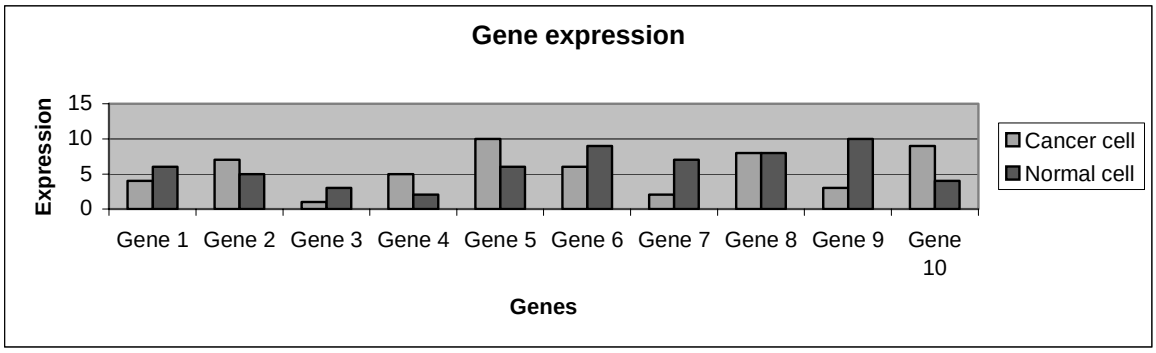


Figure 1. An example of how a set of genes can differ in expression in cancer cells compared with normal cells.

The genetic makeup and expression-profile of a cell may thus determine its fate as a normal or cancer cell. Each type of cell expresses a unique combination of genes. The degree to which genes are expressed in a cell can be quantified and displayed in a gene expression profile. Comparing the profiles of normal and breast cancer cells may elucidate which genes play a significant role in breast cancer (Mort, 2000). It is therefore important to determine which genes that are more expressed respective less expressed when cells become cancerous. In this way researchers can gain information about how the presence or absence of some genes is related to breast cancer. Bioinformatic tools must thus be developed in order to

identify different complex patterns when searching for cancer-related changes in gene expression.

This project demonstrates the application of decision tree approaches for the identification of genes and risk factors relevant for the classification of breast cancer patients. Chapter 2 describes the concept of data mining, the application of decision trees in medicine, the C4.5 decision tree algorithm, and the Gene Logic database. Chapter 2 also discusses the main genes and risk factors involved in breast cancer. In Chapter 3, the aims and objectives of this study are presented and the motivations behind this project are discussed. In Chapter 4, three works related to this project are briefly described. Chapter 5 explains the criteria that were used in order to filter the huge amount of data in Gene Logic. This Chapter further describes four different approaches performed in order to analyze the different aspects of the C4.5 algorithm. The resulting decision trees and production rules derived by the C4.5 algorithm are also shown in Chapter 5. In Chapter 6, the resulting decision trees and production rules are analyzed. The results are further discussed in Chapter 7. Finally, in Chapter 8 some conclusions are made while some possible future works are suggested in Chapter 9.

2. Background

This Chapter describes data mining with major focus on decision trees. Also breast cancer and the most central genes and risk factors involved will be discussed. Finally the C4.5 decision tree algorithm is described.

2.1 Data mining and decision trees

The term data mining refers to using a variety of techniques to identify information or decision-making knowledge in bodies of data. This concept has become a popular technology in many applications. A large amount of data is handled in these applications where the data has a low value in its raw form since the valuable parts are often hidden in the data. The process of data mining thus generates models that are later used for predictions. With an appropriate learning method, that is, a method that uses the training data set in order to correctly classify new samples, it is possible to develop accurately predictive applications.

A data mining system can accomplish several different data mining tasks. Some of these tasks are described by J. Han (1999) as:

1. *Class description*. Class description provides a concise and succinct summarization of a collection of data and distinguishes it from others. Class description should cover not only its summary properties, such as count, sum, and average, but also its properties on data dispersion, such as variance, and quartiles.

2. *Association*. Association is the discovery of association relationships or correlations among a set of items. They are often expressed in the rule form showing attribute-value conditions that occur frequently together in a given set of data.

3. *Classification*. Classification analyses a set of training data and constructs a model for each class based on the features in the data. A decision tree or a set of classification rules is generated for such a classification process. This can be used for classification of future data where for example one may predict a disease based on the symptoms of the patient.

4. *Prediction*. This mining function predicts the possible values of some missing data or the value distribution of certain attributes in a set of objects. It involves the finding of the set of attributes relevant to the attribute of interest and predicting the value distribution based on the set of data similar to the selected object(s).

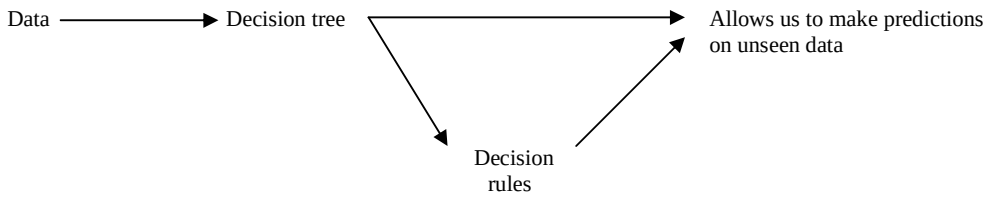
5. *Clustering*. This analysis is made to identify clusters, that is, a collection of data objects that are similar to one another, embedded in the data. Similarity can be expressed as functions

specified by the users or experts. In clustering, in contrast with classification, the clusters are not labelled.

6. *Time-series analysis*. Time-series analysis is used to analyse large set of time-series data to find certain regularities and interesting characteristics.

In this project, a decision tree model is used as the technique for mining the information in the Gene Logic database, see Section 2.5.

A decision tree is constructed by looking for regularities in data. Once a decision tree has been constructed, it is a simple matter to convert it into an equivalent set of rules. This course of action is simply shown in Figure 2.



Figur 2. The construction of decision trees and decision rules.

A decision tree is a model that is both predictive and descriptive (Kuo et. al. 2001). In other words, a decision tree has the ability to both describe available data and to predict the classification of new data. Decision trees mainly get their name from the resulting models that have the shape of tree structures with decision rules. An example of a decision tree is shown in Figure 3. This technique is mainly used for classifying which certain group a specific case belongs to. Decision trees are therefore believed to be an useful technique in this project in order to extract information about the most important genes and clinical data involved in breast cancer.

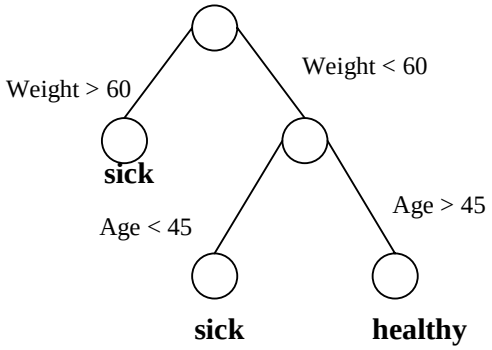


Figure 3. An example decision tree.

2.2 Decision trees and applications in medicine

Of particular value to medicine is the requested accuracy and interpretability of the results of data mining (Zupan et. al. 1998).

There are different branches of machine learning; statistical or pattern recognition methods, inductive learning of symbolic rules, and artificial neural networks (Lavrac, 1999). Probably the most promising area for medical data analysis was from the very first beginning, the symbolic learning of decision trees and decision rules. Neural networks on the other hand are more like black box classifiers lacking the transparency of generated knowledge and lacking the ability to explain the decisions (Kononenko, 2001).

According to a study performed by Igor Kononenko (2001), there are some specific requirements that any machine learning system has to satisfy in order to be used in the development of applications in medical diagnosis. These specific features include:

1. *Good performance*, which refers to the ability of the algorithm to extract significant information from the available data. The diagnostic accuracy on new cases should thus be as high as possible.
2. *The ability to appropriately deal with the missing data and with noisy data*. This is due to the fact that the description of patients in patient records often lacks certain data. Medical data also often suffers from uncertainty and errors.
3. *The transparency of diagnostic knowledge*. It is very important for the explanation of decisions to be transparent for the user. It should therefore be possible to analyze and to understand the generated knowledge.
4. *The ability to explain decisions and the ability of the algorithm to reduce the number of tests necessary to obtain reliable diagnosis.* The system must be able to explain decisions when diagnosing new patients. Also, since the collection of patients in medical practice often is very expensive and time consuming and sometimes also harmful for the patients, it is very desirable for the classifier to reliably diagnose with a small amount of data about the patients.

In the same study Kononenko (2001) compares the appropriateness of various algorithms, including decision trees and neural networks, for medical diagnosis. Among the compared algorithms, only decision tree builders were shown to be able to select the appropriate subset of attributes. This means that, with respect to the criterions of reduction of the number of tests, these algorithms have clear advantage over other algorithms. With respect to

transparency and explanation ability criteria, it turned out to be a great difference between the algorithms tested. The back-propagation neural networks showed to have a non-transparent knowledge representation, which means that they in general cannot easily explain the decisions made. Decision trees on the other hand were fairly easy to understand since the paths from the root to the leaves were shorter, containing few but most informative attributes.
In many cases however, one might feel that such a tree, containing only a few nodes after pruning, describes very poorly the diagnoses and is therefore not sufficiently informative. This remains thus to be seen during this project where the decision trees will be tested on breast cancer data from the Gene Logic database.

2.3 Cancer

Cells are components of the body and each cell has a fixed lifespan that is described by the cell cycle. The cell cycle has two major periods. In the first period the cell grows and performs its specific tasks in the body while in the second period it reproduces and splits into two new cells (Lodish et. al. 1995). The cell cycle is a very organized process and the cells that are produced are exact copies of the original cell. The growth and division of normal cells goes on continuously in order to produce sufficient amount of new cells that can replace the old and damaged cells (Lodish et. al. 1995).
In cancer cells the growth and dichotomy goes on without any control. This accelerated and uncontrolled growth of cancer cells results in a lump that is called a *tumor*. Tumors can be either benign or malign. A benign tumor is often harmless and is covered by a membrane that keeps the tumor isolated from the normal surrounding cells (Bristol-Myers Squibb, 2002). Usually a benign tumor grows slowly.
A malign tumor is almost never encapsulated or covered by a membrane. It usually grows faster than a benign tumor and invades the surrounding tissues. This is called *infiltration* (Bristol-Myers Squibb, 2002). Malign cells also have a capacity to metastasize through the blood and lymphatic systems to distant organs.
Cancer cells deprive nourishment from the normal cells which causes weight- and strength loss among cancer patients (Bristol-Myers Squibb, 2002).

2.3.1 Breast cancer

Breast cancer starts in the breast tissue and like most cancers it is named after the part of the body where it first starts but can later spread to other parts of the body.
As it is shown in Figure 4, the breast is made up of lobules, ducts, and fatty, connective, and lymphatic tissue. *Lobules* are the glands that make the milk when a woman has a baby and the tubes that connect them to the nipple are called *ducts* (American Cancer Society, 2002).

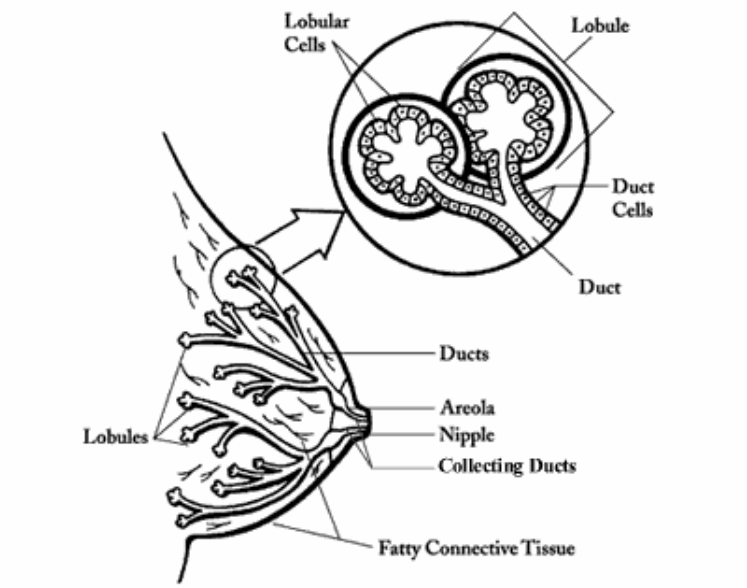


Figure 4. A close-up image of the female breast and its constituents (Reprinted by permission from Anthony Saffioti, American Cancer Society, 2002).

A fluid called *lymph* is carried in lymph vessels and eventually leads to a collection of tissues called *lymph nodes*, mostly located under the arm. The lymph fluid contains immune system cells and tissue waste products. If breast cancer cells reach the underarm lymph nodes, there is a greater possibility for the cancer to spread to other organs of the body (American Cancer Society, 2002).

2.3.1.1 Risk factors

The cause of breast cancer is not yet known but there are certain risk factors that are linked to the disease. While all women have the risk of getting breast cancer, a risk factor is anything that increases a person's chance of getting the disease (American cancer society, 2002). Some risk factors can be controlled while others can not be changed.

Some of these risk factors are described by M. Jönsson (2000) and include:

- **Gender:** While breast cancer usually affects women, men can also get breast cancer, although this is rare. The risk of developing breast cancer is estimated to be 100 times higher for a female.
- **Age:** About 20% of the women that get breast cancer are under the age of 50 (Lidbrink, 2001). The chance of getting breast cancer thus increases as women get older. However, the rate of breast cancer incidence slows somewhat between the ages of 45-50 years. According to Lidbrink (2001) this must be due to the hormonal changes that arise during the menopause.
- **Family history and genetic risk factors:** Breast cancer risk is higher among women whose close blood relatives have or have had this disease. It is estimated that at least 5% of women with breast cancer have a true hereditary predisposition. These cases appear often as an effect of a single genetic abnormality in genes that are involved in predisposition to hereditary breast cancer such as BCRA1 and BRCA2.
- **Menstrual periods:** Menarche represents the development of the mature hormonal environment of a young woman. Women who began having periods early, that is, before the age of 12, or who went through the menopause after the age of 55 have a small increased risk of breast cancer.
- **Age of pregnancy:** The majority of studies have shown that a younger age at first full-term pregnancy predicts a lower lifetime risk of developing breast cancer while a first full-term pregnancy after the age of 30 increases the risk (Ranstam & Olsson, 1995).
- **Alcohol:** Alcohol consumption has gained attention as a possible risk in the development of breast cancer. Consistent findings on the relationship between alcohol and risk of breast cancer seem to exist only for a relatively high consumption (Ranstam & Olsson, 1995).
- **Smoking:** There have been conflicting evidence in the literature on the relationship between smoking and breast cancer but there is presently no strong support for this relationship (Mesko et. al. 1990).
- **Obesity:** Being overweight appears to be associated with a higher risk of post-menopausal breast cancer in most studies. However, some studies show that obesity might be slightly correlated with a decreased risk of breast cancer in pre-menopausal women (Pujol et. al. 1997).

The three last mentioned risk factors and also a patients age will be analysed and studied during this project in order to further evaluate the role of these four risk factors in breast cancer development. This is due to the fact that there has been a lot of research regarding the relations between these risk factors and breast cancer and it will therefore be interesting to find out if the resulting decision trees will reflect the relations described by different researchers.

2.3.1.2 Major genes

Breast cancer, like other cancer types is a disease of the cell cycle, see Figure 5 (Ingvarsson, 2000). Cell cycle disturbance is therefore thought be the main cause of cancer growth.

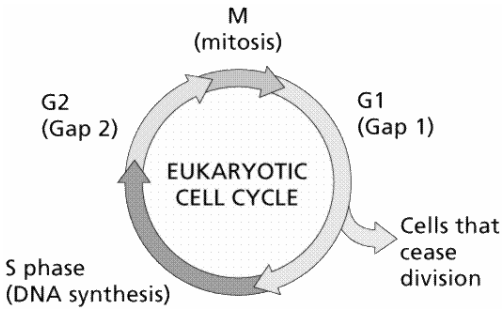


Figure 5. Illustration of the eukaryotic cell cycle (Reprinted by permission from William E. Camelet, Lake Michigan College, 2002).

Genes that are differentially expressed in tumor tissues compared with normal tissues are potential diagnostic markers and drug targets. Progress has been made toward a better understanding of breast tumorigenesis and several markers have been identified. A few of the major genes involved in breast cancer are shortly described below.

- **ErbB2**

  The tyrosine kinase receptor ErbB2, also called HER2/neu, is a member of the epidermal growth factor receptor (EGFR) family (Wang & Hung, 2001). This gene has been shown to be amplified/over-expressed in approximately 25-30% of invasive breast cancers in human (Mendelsohn & Baselga, 2000). ErbB2 encodes a transmembrane glycoprotein with tyrosine kinase activity that functions as a growth factor receptor. ErbB2 is thus important in breast cancer growth, accessible as a cell surface receptor, and is expressed at high levels in breast cancer and at low levels in normal tissue (Schnitt, 2001).

- **C-myc**

  C-myc is involved in the regulation of the transcription of other genes important for cell growth regulation (Jönsson, 2000). Over-expression of Myc transcription factor runs the cell through the cell cycle. Amplification of this oncogene has been reported in 20% of primary breast tumors (Dairkee & Smith, 1996).

- **Cyclin D**

  A small gene family encodes three different D-type cyclins called D1, D2 and D3. These are referred collectively as Cyclin D (Lodish et. al. 1995). An increase in Cyclin D eventually results in progression throughout the G1 and S phase of the cell cycle (Ingvarsson, 2000). Cyclin D is one of the most commonly over expressed proteins in breast cancer (Hynes & Dickson, 1996).

- **p53**

  p53 is a cell cycle-regulating transcription factor (Hynes & Dickson, 1996). p53 is the most common genetic change found in breast cancer (Dairkee & Smith, 1996). Losses and/or mutations of p53 occur in about 40% of the breast cancer cases (Hynes & Dickson, 1996). The lack of p53 leads to increased rates of DNA alternation-mutations in the cells (Lodish et. al. 1995).

  p53 has the ability to check cell-cycle progression and hold cells in quiescence or even lead cells to commit suicide, apoptosis, unless conditions are appropriate for cell-cycle progression (see Figure 6). This gene is therefore considered to be a tumor-suppressor gene which means that it can prevent cells from becoming cancerous (Lodish et. al. 1995).

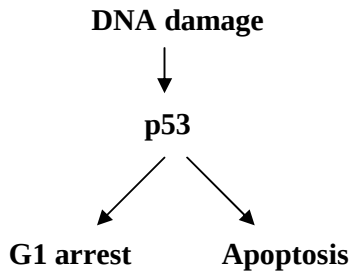


Figure 6. p53 has a central role in cell division, maintenance of genetic integrity and regulation of programmed cell death.

- **Rb**

  The Retinoblastoma (Rb) gene is also considered to be a tumor-suppressor gene (Lodish et. al. 1995). The Rb gene product is a nuclear phosphoprotein with DNA-binding properties (Dairkee & Smith, 1996). This product is associated with several other proteins as a means of regulating cell growth primarily at the G1 phase. Rb functions as a proliferation-inhibiting gene and its inactivation by a variety of mechanisms can finally lead to disease development (Jönsson, 2000).

- **Wnt-5a**

  This gene has been proposed to be a novel tumor-suppressor gene and its lost has been shown to be related with an increased risk of breast cancer. A function suggested for the Wnt-5a gene is a role in the blocking of cell movement.

  Wnt-5a can thus prevent tumor-cells from migration and metastasizing (Jönsson, 2000).

- **BRCA1 and BRCA2**

  Germline mutations in either BRCA1 or the BRCA2 gene are responsible for the majority of hereditary breast cancers (Feunteun, 1998). These two genes play a role in monitoring and/or repairing DNA lesions, see Figure 7. The relaxation of this monitoring caused by mutations of either of these two genes leaves un-repaired events, leading to accumulation of mutations and finally to cancer (Hedenfalk et. al. 2001).

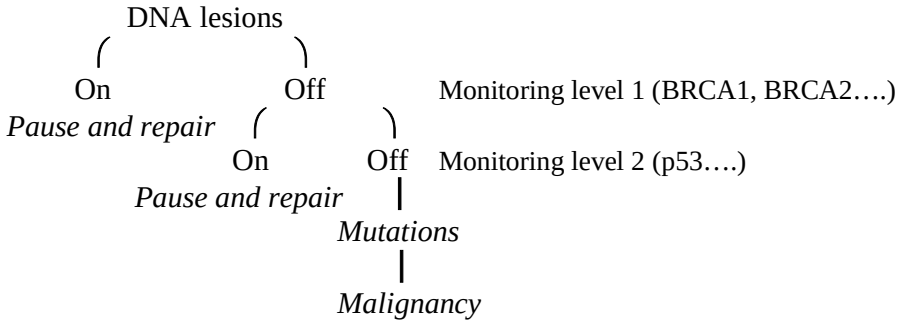


Figure 7. The above model illustrates the flow of possible events leading to malignancy (Feunteun, 1998).

In Figure 7 the BRCA1 and BRCA2 genes work as *genome caretakers*, that is, molecules that act as sensors of DNA lesions and participate in repair processes. p53 works as a *gatekeeper*, that is, a molecule that directly controls the progression of the cell cycle. The right side of the model illustrates the events leading to cancer whereas the left side illustrates the alternative pathways that are taken when the monitoring of DNA lesions is fully active (Feunteun, 1998).

2.4 The Gene Logic database

The Gene Logic database contains gene expression data and clinical data from more than 6,800 samples. About 2,700 of these samples are cancer related. Most of the expression data is gained from the Affymetrix technology, containing gene expressions from human, rat and mouse. In this project, only the information from chipset HG-U95 will be used. Gene Logic contains gene expression data from 65,000 transcripts in human. Gene Logic also contains a list of clinical data, which are shown in Table 1.

Basic Donor Information	• DOB • Age at excision • Gender • Race • Height • Weight
Obstetric Information	• Menstrual history • Last menstrual period • Pregnancy information
Family History	• Family members with significant medical conditions
Social History	• Diet information • Smoking history • Alcohol consumption history • Recreational drug use history
Medical Information	• Medical History • Surgical History • Medications
Laboratory Values	• Lab tests taken on day of surgery • Lab values taken while in hospital

Table 1. Clinical data in Gene Logic.

The HG-U95 chipset used in this project contains both quantitative and qualitative measures of the gene expression profiles. The quantitative measures are called *Average difference*. These measures are described by a real number. The measures in Gene Logic are normalized

and no negative values of average difference exist in the database. The qualitative measures on the other hand are called *Abscall*. These measures use the values 'present' (P), 'absent' (A) and 'marginal' (M) to describe the expression of a gene in Gene Logic. Abscall uses a threshold to place different genes in these three categories. The threshold is set by Affymetrix in a confidential manner but it is considered to be statistically correct. If the expression of a gene is over the threshold, the expression has been detected in the tissue and the gene is thus said to be 'present'. On the other hand, if a gene's expression has not been detected in the tissue, it is said to be 'absent'. This does however not mean that genes in the 'absent' category surely are not expressed. It only means that the expression of these genes has not been detected or that the expression has not been possible to measure. The expression of the genes in the 'marginal' category are thought to be somewhere in between the expression of the genes in the 'present' and the 'absent' categories. The 'marginal' category is thus more like a gray-zon. In a few cases, the expression of a gene can also be presented by a zero (0). This means that the expression of that gene has not been examined for the patient and thus no results are found.

2.5 The C4.5 algorithm

Several systems for learning decision trees to correctly classify samples have been proposed. Prominent among these are ID3 and its new version C4.5 (Nilsson, 1996). C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan in 1993. Given a set of classified examples, the ID3 algorithm can induce a decision tree biased by the *gain* measure. This measure is aimed to minimize the number of tests needed for classification of a new object. The gain measure is the standard information entropy difference achieved at node x and it is expressed as:

$$\text{Gain}(x) = \text{info}(T) - \text{info}_x(T)$$

Gain(x) thus represents the difference between the *information needed to identify the class of an element of the training set T* and *the information needed to identify the class of an element of T after the value of attribute x has been obtained*. In other words, this ratio represents the gain in information due to attribute x (Quinlan, 1993).

The gain measure introduced above has however been shown to have a serious weakness since it has a strong bias in favour of tests with many outcomes. To avoid this problem, the C4.5 algorithm was introduced using a combination of gain and *gain ratio* criteria (Quinlan, 1993). The gain ratio criterion uses the measure of SplitInformation defined as:

$$\text{SplitInformation}(x) = -\sum_{i=1}^{n} (|T_i|/|T|) * \log_2(|T_i|/|T|)$$

where T is the number of samples in the training set and T_i is the partition of T induced by attribute x. SplitInformation represents the potential information generated by dividing the training set T, on the basis of the value of the categorical attribute x, into n subsets.

The gain ratio measure, used by the C4.5 algorithm, is defined as

$$\text{GainRatio}(x) = \text{Gain}(x) / \text{SplitInformation}(x)$$

and expresses the proportion of information generated by the split, that is, the information that appears helpful for classification. If the split is trivial, the split information will be small and this ratio will be unstable. The gain ratio criterion therefore selects a test to maximize this ratio as long as the numerator, that is, Gain (x), is larger than the average gain across all tests examined (Quinlan, 1993).

An essential property of the C4.5 algorithm is called *pruning*. Pruning of a decision tree is done by replacing a whole subtree by a leaf node. The replacement takes place if a decision rule establishes that the expected error rate in the subtree is greater than in the single leaf. Pruning is a method most widely used for obtaining appropriately sized trees and there are different kinds of techniques for pruning. C4.5 uses pessimistic pruning to prevent overtraining. The pessimistic pruning uses a statistical method to calculate the error rate associated with each node and to adjust the tree to reflect bias. Thus, this pruning technique pessimistically increases the errors observed at each node using statistical measurements to encourage pruning.

According to Quinlan (1993), the advantages of pessimistic pruning over other pruning methods include:

1. It builds only one tree.
2. It does not require held out training data for error estimation.
3. It provides a more reliable tree when data is scarce.

The C4.5 algorithm can also address other issues not dealt with by its predecessor ID3. Some of these issues include:

- Handling and incorporation of numerical (continuous) attributes, in this case ‘age’ and ‘BMI’.
- Dealing with incomplete information (missing attribute values), in this case some values of ‘alcohol status’ and ‘smoking status’.

- Avoiding overfitting to the data by determining how deeply to grow a decision tree.
- Post-pruning after induction of trees in order to increase accuracy and avoid overfitting.
- Improving computational efficiency.

Another property of C4.5 is rule derivation, where rules are derived from the decision tree. A principal aim of decision tree models is that it should be intelligible to human. C4.5 thus tries to achieve this goal by re-expressing a classification model as production rules, a format that appears to be more intelligible than trees (Quinlan, 1993). At first, a rule is written for each path in the decision tree from the root to a leaf. The left-hand side (L) of a rule is easily built from the label of the nodes and the labels of the arcs, whereas the right-hand side (R) is a class, L → R. The rules derived by some paths may have an unacceptably high error rate or may duplicate rules derived from other paths. The process therefore usually yields fewer rules than the number of leaves on the trees. The resulting rule set can then be further simplified and reduced. Some of the conditions on the left-hand side of the rules may be eliminated. The left-hand side of those rules will then be replaced by new conditions (L'). This replacement happens only if the proportion of the training set that satisfy respectively L and L' are equal (Quinlan, 1993). For example, consider the following simple rule:

```
AGE > 20
ALCOHOL_STATUS = NEVER_USED
AGE > 36
     ->  class  NON_CANCER
```

The C4.5 algorithm may eliminate the first condition of the above rule (Age >20). This can happen if this condition covers a single or only a few instances, for example, if only 1 patient in the data set is between the ages of 20 and 36. In that case, the first condition may not be considered, by the C4.5 algorithm, to contribute to the classification of samples and the C4.5 algorithm might therefore simplify the rule to:

```
ALCOHOL_STATUS = NEVER_USED
AGE > 36
     ->  class  NON_CANCER
```

It is however also conventional to define a fallback or default rule that comes into play when no other rule covers a case. A reasonable choice for the default class would be that class

which appears most frequently in the training set. The system therefore simply chooses the class which contains the most training cases, not covered by any rule, as the default class. The system will in this way be in favor of the class with the higher absolute frequency.

An extension of the C4.5 algorithm, called C5.0, has also been developed. The Unix program C5.0 and its Windows counterpart See5 (Quinlan, 1996) are considered to be superior to the C4.5 algorithm in several ways. For instance, C5.0 has several new data types in addition to those available in C4.5, C.5 is more than 200 times faster than C4.5, and it incorporates several new facilities such as variable misclassification costs, instead of treating all misclassifications as equal. However, since only incomplete demonstration versions of the C5.0 software are available for free, the C4.5 algorithm is instead used in this project.

3. Problem definition, hypothesis and motivations

This disease mining project will explore the possibilities of data mining on the Gene Logic database, containing gene expression data and clinical data for thousands of samples taken from both diseased and healthy tissue. The problem definition is to create a general method for the development of decision trees in order to extract valuable information regarding the major genes and risk factors involved in different diseases. The method will in this project be tested on the breast cancer data found in the Gene Logic database. The main goal is thus to develop a method that can isolate drug targets by identifying genes that are modulated in breast cancer, and which vary with clinical parameters associated with this disease.

The hypothesis is that decision trees are a useful classification technique in the disease mining project. Consequently, with this technique it may be possible to mine through an amount of gene expression data and clinical data in order to classify patients correctly. It may thus be possible to extract valuable information about the main genes and risk factors involved in breast cancer.

There are a couple of different ways to define the concept of 'useful classification technique' mentioned in the hypothesis. When is it adequate to state that the hypothesis is true respective false? Well, the resulting decision trees can differ dependent on how many genes and risk factors that are classified correctly. Even if the decision trees don't classify *all* the genes and risk factor correctly, as described by different literatures, it might still be an 'useful' classification technique depending on the amount of genes and risk factors classified correctly in comparison with the *total* amount of attributes present in the resulting tree. If the decision tree for example correctly identifies three out of five attributes, there is good reason to further analyze the two new attributes in order to find out *if* and in *what way* these are involved in breast cancer. Probably the simplest and most profitable way to evaluate the classification performance of a created decision tree in this project, is to compare the gained classification results with classification results made by a very simple randomized algorithm, where the classifications is made in a random manner. This randomized algorithm is an algorithm where the distribution of the cancer and non-cancer patients in the data set is known. If this algorithm gives better classification results than the decision trees created by the C4.5 algorithm, the created decision trees will be regarded as insignificant and the application of decision trees on the Gene Logic data will therefore be dismissed. Otherwise, the C4.5 decision tree algorithm will be considered to be a better classification technique, compared with the randomized algorithm, for the classification of data in the Gene Logic database.

It's important to remember that there is still much that is unknown about breast cancer and therefore any kind of technique that gives more significant classification results than randomly made classifications, must be considered useful. In other words, any classification technique that in some way simplifies an earlier diagnosis of the breast cancer and also simplifies the separation of cancer and non-cancer patients, is very valuable and should be further investigated and developed.

One motivation behind this disease mining project is the fact that very little attention has been given to the potential of applying decision trees to the kind of data found in the Gene Logic database, see Chapter 4. The rise in attention and focus on decision support solutions using data mining techniques has become a big interest in different classification modelling (Apte & Weiss, 1997). Decision tree and decision rule solutions offer a level of interpretability that is unique since these tree-like solutions are often easy to understand, even for the non-technical users.

Another motivation is that the data content of the Gene Logic database is quite new and has not yet been fully analysed. It is therefore interesting to examine if the conclusive results illustrated by the final decision trees will agree with the findings of the important genes and risk factors involved in breast cancer so far. The decision tree may also shed some light on new, unknown genes and risk factors involved in breast cancer.

3.1 Aims and objectives

- *Selection of an appropriate amount of data from the Gene Logic database*. In order to extract valuable information it is important to only consider the amount of data necessary for the creation of decision trees. Otherwise it might become very difficult to extract any kind of interpretable information from the resulting trees. In order to increase the probability of selecting the right amount of data, a lot of literatures concerning the main genes and risk factors involved in breast cancer, will be read. The LocusLink website, at the NCBI homepage (http://www.ncbi.nlm.nih.gov), will also be used as a tool in order to find the known breast-related genes. The expression profiles of these genes will later be used as input to the decision tree algorithm.

- *Installation of software for the creation of the decision trees*. Several different methods exist for the generation of decision trees and the results vary significantly depending on the chosen method. In this project the C4.5 algorithm, induced by

Quinlan in 1993, will be used. This software is an extension of the basic ID3 algorithm and is therefore believed to better handle situations where ID3 has shown certain weaknesses, see Section 2.5.

- *Deriving decision trees and production rules based on gene expression data*. Gene expression data will be used as input to the C4.5 algorithm in order to find relevant breast cancer markers that can separate cancer patients from non-cancer patients in Gene Logic. This approach is thus performed in order to investigate the capacity of the C4.5 algorithm in diagnosing patients on the basis of gene expression data.

- *Deriving decision trees and production rules based on risk factor data*. Risk factor data will be used as input to the C4.5 algorithm in order to investigate the capacity of the C4.5 algorithm in diagnosing Gene Logic patients based on only risk factor data. In this way it may be possible to shed light on the risk factors that are most important in the development of breast cancer.

- *Deriving decision trees and production rules based on both gene expression data and risk factor data*. Gene expression data and risk factor data will be used as input to the C4.5 algorithm in order to investigate the capacity of C4.5 in diagnosing patients when the two data sets are combined.

- *Deriving decision trees and production rules based on the expression of different probes of breast cancer related genes.* In this way it may be possible to suggest genes involved in different signalling pathways during breast cancer.

- *Evaluation of decision trees, production rules and the performance of the C4.5 algorithm.*
    - The performance of the decision tree algorithm in classifying patients correctly will be evaluated through comparisons with the classification performance made by a simple algorithm, where the sample classifications are done in a random manner. In this way it is possible to determine if the classification performance of the decision trees created by C4.5 is better

than the classification performance of the randomized classification algorithm.

- The resulting decision trees and production rules created for the illustration of different signalling pathways will be evaluated through comparisons with existing literature. In this way it is possible to find out if the resulting trees really reflect the relations and pathways described by different researches within the molecular biology of breast cancer.

4. Related work

Only a few previous works have been done regarding decision trees and breast cancer studies. Some articles concerning the use of decision tree models are described below. However, none of the articles mentioned below apply the use of decision trees on the kind of gene expression- and risk factor data that will be used in this project. It is therefore very hard to compare this project with these related works in order to analyse and evaluate the different approaches used. This situation has both its advantages and disadvantages. The advantage is that this project explores a new and unknown area that might contribute with very interesting and valuable results important in different medical approaches and for earlier patient diagnosis. The disadvantage is of course that the evaluation of the resulting trees will be more difficult because there is no previous works to compare the resulting trees with. In other words, it is not possible to get any tips or 'warnings' from studies and research done within this area since searches against PubMed to current date (2002-04-18) do not show any similar approaches compared with this project.

However, in a study performed by Pendharkar et. al. (1999) the authors attempted to illustrate how different data mining approaches can be used to predict and diagnose the occurrence of breast cancer. According to this paper researchers have previously used several statistical and artificial intelligence approaches for predicting breast cancer and these studies have indicated that artificial intelligence approaches can be successfully applied in this way. In the study of Pendharkar et. al., artificial neural networks (ANN) and data envelopments analysis (DEA), for binary classification problems, are used and compared as tools for mining breast cancer patterns. The results indicate that neural networks outperform DEA in terms of prediction accuracy. However, one of the problems described concerning data mining techniques is the wide variety of approaches that can be used. This makes the selection of a particular technique that "best" matches a given problem a difficult task. In the study made by Pendharkar et. al., the authors specially emphasize on how association rules can play an important role in the prediction of cancer development and the authors also discuss the future possibilities of using the ID3 machine learning technique for the generation of decision trees from existing breast cancer data. In order to further explore these possibilities mentioned by the authors, this project will use the C4.5 decision tree algorithm on breast cancer data found in Gene Logic. In other words, the C4.5 algorithm will be used instead of the ID3 algorithm since this algorithm is the new version of ID3 and thus introduces a number of extensions.

Another study that focuses on the growing demand for decision model approaches for problems in medicine and health care was written by Bohanec et. al. (2000). The authors present an approach to the development and application of hierarchical decision models that is based on an expert system shell for multi-attribute decision support, DEX. They demonstrate the applicability of this approach presenting different real-life applications in health care including the assessment of breast cancer risk. A prototype model is developed in order to assess the risk of breast cancer. The risk of cancer is evaluated by decision rules that are defined by the experts. In other words, the models presented in the study of Bohanec et. al. were developed "manually", that is, through the collaboration between the experts and decision analysts, who used DEX mainly as a computer-based editor and storage of models. The study of Bohanec et. al. thus only uses information about the risk factors of breast cancer for the creation of decision trees. In this project however, both risk factor data and gene expression data will be used for the creation of decision trees and production rules. In other words, this project also explores the possibilities of using gene expression data as input to the C4.5 algorithm.

In 2001, Kuo et. al. performed a study where breast masses in a series of pathologically proven tumors were evaluated using data mining with decision tree model for classification of breast tumors. Regions of interest of ultrasound images and co-variance texture parameters were used as the inputs to construct the decision tree model. C5.0 algorithm was used for the construction of decision trees. Summary of performance between an experienced physician and the proposed data mining model provided evidences for good diagnosis of breast tumors with the proposed method. Since the results from this study provided evidences for good diagnosis of breast tumors, the disease mining project performed here will also make use of the decision tree model for the analysis of breast cancer data, found in Gene Logic. In this project however, the C4.5 algorithm will be applied on different kinds of input data, compared with the study performed by Kuo et. al. (2001).

5. Methods and results

This Chapter describes the data and the methods used during this project. The resulting decision trees and production rules are also presented.

5.1 Filtration of data

As it was mentioned in the first objective shown in Section 3.1, one of the first steps in this project is to select an amount of data from the Gene Logic database to work with when running the decision tree program. This amount should not be too large but it should not be too small either. A too small amount of data could be overfitted by the decision program and may therefore give the wrong impression about the effect of attributes on the cause and development of breast cancer. A too large amount of data on the other hand, consisting of many samples and attributes, may cause some problems when it is used as input to the C4.5 algorithm. One problem could for example be caused by long running times when creating the decision trees. This project will focus on data concerning female patients of all ages in the database. The female patients should all have values on certain specific attributes. These attributes include BMI, age, alcohol status and smoking status. Since there has been a lot of research regarding the relations between these four risk factors and breast cancer, it will be interesting to find out if the resulting decision trees and production rules will reflect the relations described by different studies. The values of height and weight in gene Logic will be used in order to calculate the body mass index (BMI) of each patient. This value is a measure that correlates with fatness. To determine BMI, weight in kilograms is divided by height in meters squared, BMI = Kg / $(m)^2$. A BMI of 25 to 29.9 is considered overweight and one of 30 or above is considered obese (National Heart, Lung and Blood institute, 2002).

The patients selected from the database must also have gene expression values (abscall values) for 108 specific genes earlier chosen from the LocusLink homepage, at NCBI. These genes were gained when the word ‘breast’ was used as a keyword against the LocusLink database. These 108 genes, shown in appendix A, are thus known to be breast-related in humans and it is therefore very interesting to study their differences in gene-expression between non-cancer patients and patients suffering from breast cancer.

The above described filtrations of the data in Gene Logic results in a total of 75 patients that in all ways match the criteria set. 53 of these patients suffer from breast cancer whereas 22 patients are non-cancer samples. In Figure 8, the age distribution of the 75 patients in the data set is shown, whereas Figure 9 shows distribution of BMI values for the 75 patients.

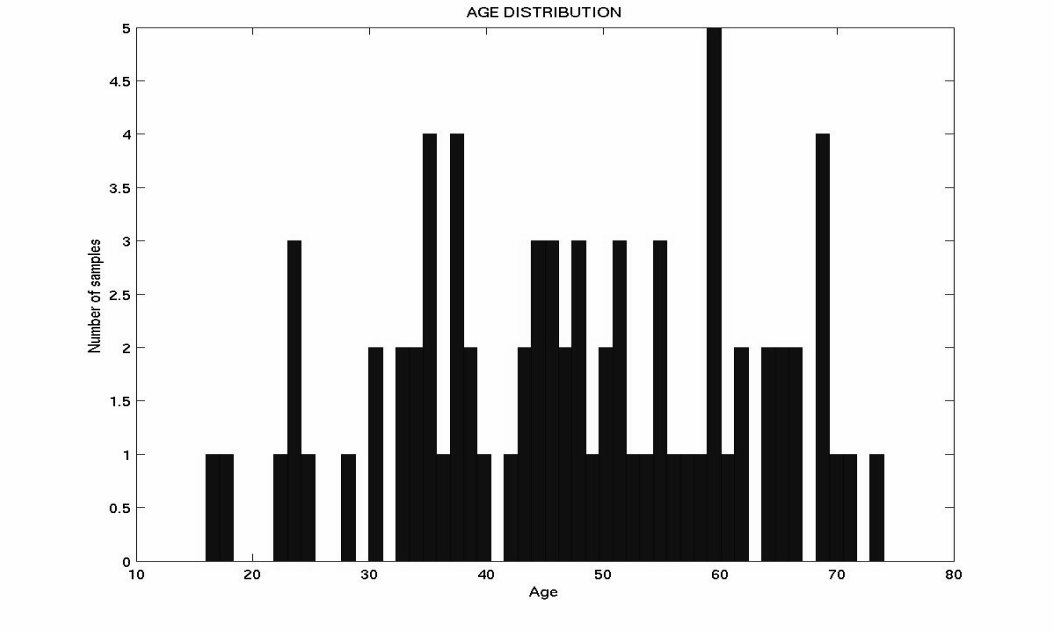


Figure 8. The age distribution of the 75 patients. The samples are distributed between the ages of 16 and 74.

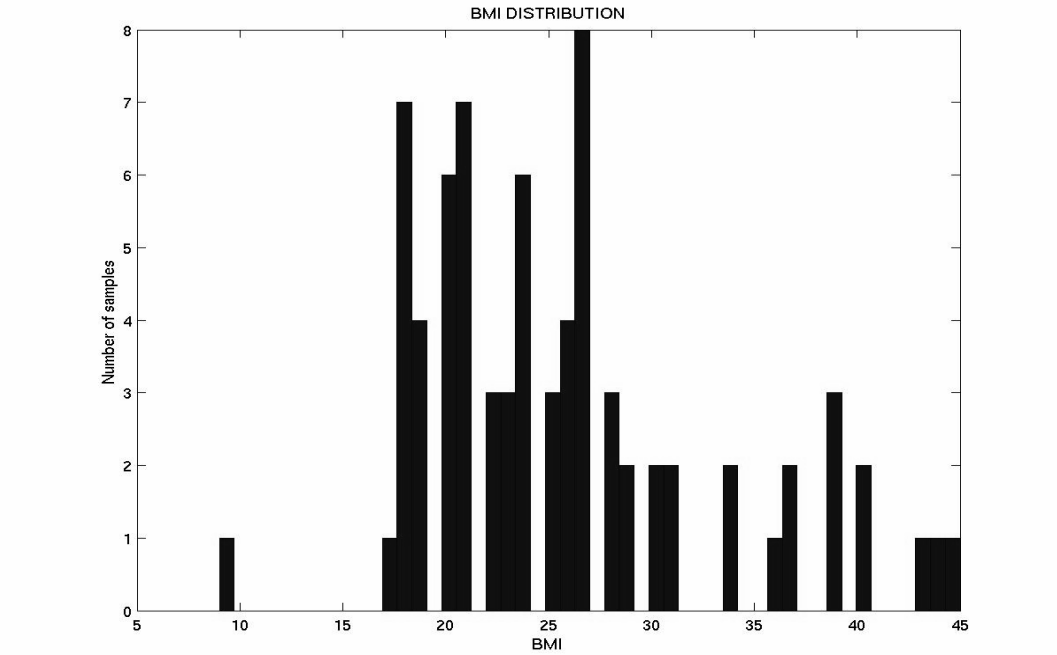


Figure 9. The distribution of BMI values for the 75 samples in the data set. The BMI varies between values 9 and 45.

The association between the ‘age’ and the ‘BMI’ values in the data set, containing values for the 75 samples, is plotted in the diagram shown in Figure 10.

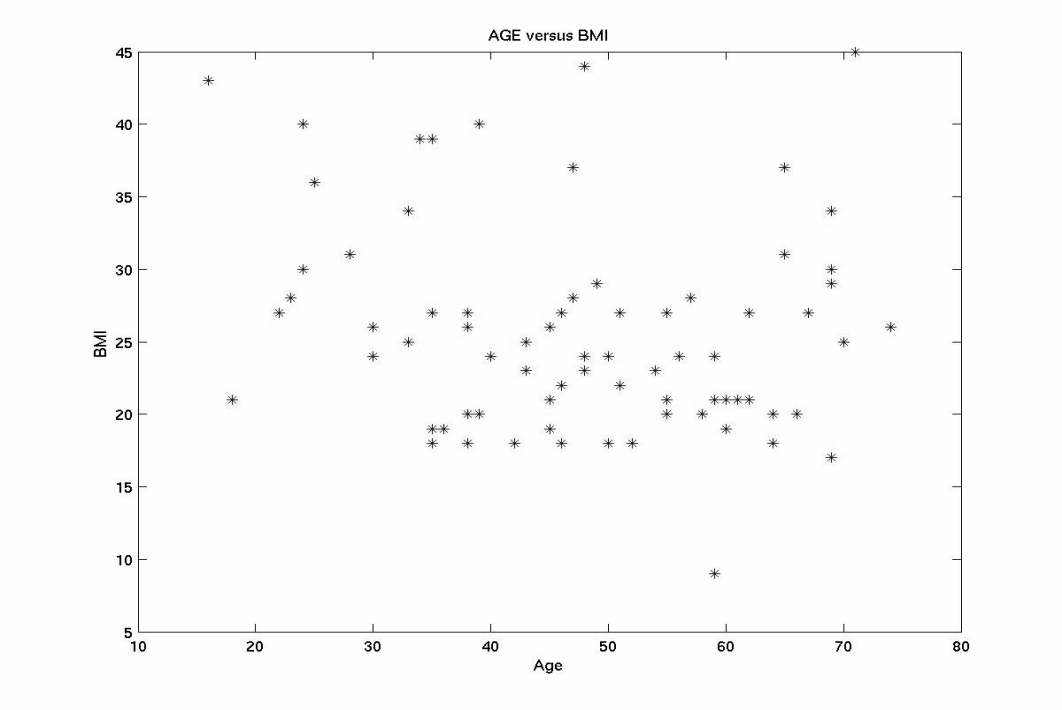


Figure 10: Plot showing the association between the ‘age’ and the ‘BMI’ values for the 75 samples in the data set.

As it can be seen in Figure 10, no association exists between the values of BMI and age. Thus, the values of these two attributes are completely independent of each other.

As mentioned earlier, only the abscall measures of the gene expressions data will be considered when comparing gene expression profiles of cancer and non-cancer patients. This is because, the average difference measures of gene expression are very much time-dependent and also very individual, see Section 2.4. Thus, these measures are very context specific and might therefore vary tremendously depending on the patient and the time of the day that the measures were taken. The abscall measures on the other hand can handle a greater amount of noise and are thus considered to be more appropriate for this study.

5.2 Deriving decision trees and production rules

The C4.5 algorithm is implemented in the C programming language and it runs on a Unix platform. The only parameter used in the C4.5 algorithm is the decision variable. This means that all the performed runs are very similar except for the fact that different decision variables

are used in different runs. The decision variables are chosen depending on which aspect of the C4.5 algorithm is being analyzed. Another difference between the performed runs is that different data sets are used as input for each run. This is done in order to study the classification performance of C4.5 when different kinds of data sets are used as input to the algorithm.

In this project, four decision tree approaches are performed. Different kinds and amounts of data are used as input to the C4.5 algorithm when creating the decision trees. The resulting trees therefore have different decision variables and illustrate different aspects of the data. In this way it is possible to analyze the different perspectives of the application of decision trees on the kind of data found in the Gene Logic database. As mentioned earlier, four different tests were set and examined. These tests and the results gained are presented below.

5.2.1 Expression of 108 breast-related genes as input data

By using the abscall values for the expression of the known breast-related genes as input to the C4.5 algorithm, it may be possible to create a decision tree that can identify the genes important for the classification of breast cancer samples. Thus the decision tree may be able to separate cancer patients from non-cancer patients considering only the expression of a set of breast-related genes.

In this test, the expression profiles of 108 genes, shown in appendix A, represented by abscall values for each patient were used as input. The input data thus contained 75 different sets of abscall values. This input resulted in the pruned decision tree shown in Figure 11.

The resulting decision tree in Figure 11 presents four genes, out of the 108 genes tested, whose expressions are considered to be most relevant in order to separate cancer patients from non-cancer patients in the data set used. The apoptosis-related gene BAX is shown twice in the resulting tree. This is because two different probes of BAX were earlier used on the HG-U95 chipset. Since BAX appears twice as a significant marker in the decision tree, the presence of BAX is identified by C4.5 as a strong marker of breast cancer. However, the greatest marker of breast cancer according to the resulting tree is the expression of a nuclear antigen called MKI67. As it can be seen in Figure 11, 2 out of 47 instances belonging to the path “if MKI67 = P then cancer” have been misclassified. The expression of MKI67 is however still present in 45 out of 53 cancer patients and is therefore considered to be a valuable marker. This is also indicated by the fact that the remaining data set, consisting of 28 patients where the expression of MKI67 is not present, only contains 8 patients suffering from

breast cancer. According to Figure 11, the presence of androgen receptor (AR) and the absence of a gene called synuclein-gamma (SNCG) also seem to contribute to the separation of cancer and non-cancer patients. The final results from this run further showed that the decision tree in Figure 11 correctly classified 53 cancer patients and 20 non-cancer patients. In other words, only 2 patients were misclassified by the below decision tree.

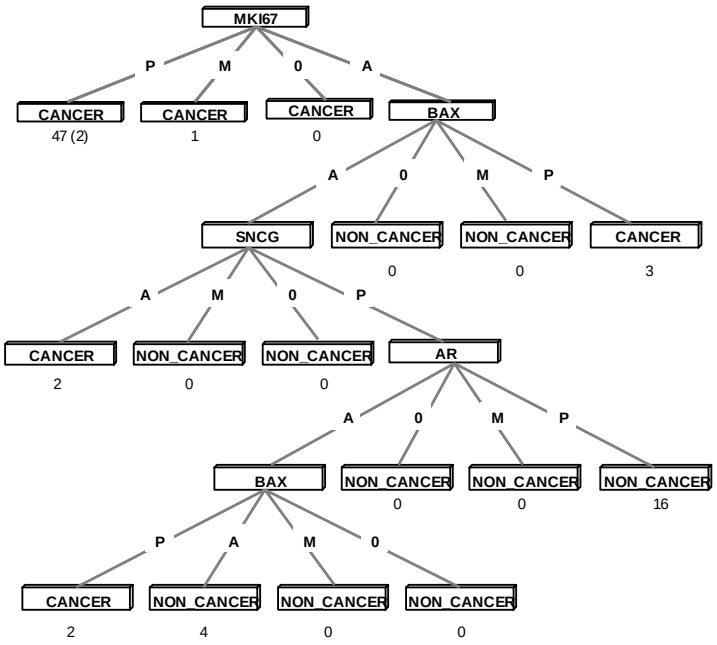


Figure 11. The pruned decision tree gained when the abscall values of 108 breast-related genes were used as input. Four genes (MKI67, BAX, SNCG, and AR) are identified by this tree. The symbol 'P', in the above tree, stands for 'present' and indicates when a certain gene is expressed. 'A' stands for 'absent' and is used when the expression of a gene is not detected. The 'M' symbol denotes an expression between the symbols 'P' and 'A'. Finally, '0' (zero) denotes a situation where the expression of a gene has not been examined for the patient. The first number beneath each node of the tree denotes the number of instances, out of the total number of cases, which belong to that path in the tree. This number may be followed by a second number in a parenthesis, for example, 47 (2). This second value (2) denotes the number of classification errors encountered out of the total number of classifications made from the data in that particular path of the decision tree (47).

The set of production rules induced by this approach, where gene expression data is used as input to the C4.5 algorithm, are presented below. Every enumerated rule is composed of one or several attribute-values and a resulting classification. For instance, the first production rule below indicates that the presence of MKI67 expression is an indicator of patients suffering from breast cancer.

```
1. MKI67 = P
     → class CANCER

2. BAX = P
     → class CANCER

3. AR    = P
   BAX   = A
   SNCG  = P
   MKI67 = A
     → class NON_CANCER

4. AR  = A
   BAX = P
     → class CANCER

5. BAX   = A
   SNCG  = P
   MKI67 = A
     → class NON_CANCER

6. SNCG = A
     → class CANCER
```

The above production rules are analyzed in Section 6.1.

As described in Section 2.5, production rules are generated by writing a rule for each path in the decision tree from the root to each leaf. Some of the conditions on the left-hand side of each rule may then be eliminated and simplified. Therefore usually fewer rules are created compared to the total number of leaves in the trees. The first production rule shown above, derived from the decision tree shown in Figure 11, covers 47 patients where the expression of MKI67 is present. This path can be seen at the top of the decision tree shown in Figure 11. Acoording to Figure 11, the second production rule shown above covers 3 cancer patients where the expression of BAX is present. This rule is gained with the first probe of the BAX gene, which has been identified as the second most relevant attribute in Figure 11. As can further be seen, the third production rule derived from the decision tree in Figure 11 covers 16 non-cancer patients in the data set. This rule is gained by following the decision tree branches starting at the the top of the tree where the expression of MKI67 is absent, corresponding to the fourth condition of the production rule, MKI67 = A. Eventually, by following the second, the third, and finally the first condition of the rule , this rule will correctly classify 16 non-cancer patients. As can further be seen in Figure 11, the fourth derived production rule shown above covers 2 cancer patients. This rule consists of AR and the second probe of the BAX gene, which has been identified as the fourth most relevant attribute in the decision tree.

According to Figure 11, the fifth derived production rule again classifies the 16 non-cancer patients who were earlier covered by the third production rule mentioned above. The fifth production rule is thus very similar to the third production rule. However, accoding to Figure 11, the fifth production rule also covers 4 additional non-cancer patients in the data set. This rule is thus gained by following the path starting at the top of the tree where the expression of MKI67 is absent. By then following the first and then the second condition of the rule, this rule will correctly classify a total of 20 (16+4) non-cancer patients. However, the fifth production rule misclassifies the 2 cancer patients shown at the bottom of the tree in Figure 11, where the second probe of the BAX gene is present. Finally, the sixth production rule shown above covers the 2 cancer patients where the expression of SNCG is absent.

The sum of unique patients covered by each of the above described production rules is however 74, even though the data used as input to the C4.5 algorithm contains 75 patients. This is because the input data contains 1 cancer patient where the expression of MKI67 is 'marginal'. This patient has therefore not been covered by any of the six above production rules derived from Figure 11. However, as earlier mentioned in Section 2.5, a default class is defined during each decision tree run and this class comes into play when none of the derived production rules can cover a case. The default class is set to the class which appears most frequently in the input data. The default class during this approach, where gene expression data was used as input to the C4.5 algorithm, was thus set to 'cancer' since the input data contains 53 cancer patients and only 22 non-cancer patients. In other words, the patient following the path "if MKI67=M then cancer" was correctly classified due to the defined default class.

In order to get an indication of how well this decision tree approach will do when it is asked to classify 'new' patients based on their gene expression data, a 5-fold cross validation was performed. Cross validation is a model evaluation method that is often used in order to prevent overfitting.

In K-fold cross validation the data set is divided into *k* subsets, and the holdout method is repeated *k* times. Each time, one of the *k* subsets is used as the test set and the other *k-1* subsets are put together to form a training set, each time resulting in a new decision tree. Then the average error across all *k* trials is computed. The advantage of this method is that it matters less how the data gets divided (Schneider, 1997). Every data point gets to be in a test set exactly once, whereas it gets to be in a training set *k-1* times. As mentioned earlier, the data set in this project contains 75 different samples (patients) and thus a 5-fold cross validation was performed where the patients in each data set were randomly chosen. Each test

set contained 15 cases whereas each training set contained 60 cases. For each training set a decision tree was built and tested on a corresponding test set. All the decision trees created during the cross validation procedure again identified the four genes presented in Figure 11, that is, MKI67, BAX, SNCG, and AR, even though the exact tree topology differed a bit between the created trees. The cross validation procedure resulted in the classifications shown in Table 2. In other words, the cross validation procedure again indicated the value of these four genes for the separation of cancer respective non-cancer patients.

Data set	**Correctly classified**	**Misclassified**	**Error**	**Correctly classified**	**Misclassified**	**Error**
1	60 (44/16)	0 (0/0)	0.0%	13 (9/4)	2 (0/2)	13.3%
2	58 (38/20)	2 (0/2)	3.3%	13 (13/0)	2 (2/0)	13.3%
3	58 (40/18)	2 (0/2)	3.3%	15 (13/2)	0 (0/0)	0.0%
4	58 (42/16)	2 (0/2)	3.3%	13 (9/4)	2 (2/0)	13.3%
5	56 (48/8)	4 (0/4)	6.7%	13 (5/8)	2 (0/2)	13.3%

Table 2. Columns 2-4 present the classification results for the 5 training sets, each containing 60 samples. Columns 2 and 3 give information about the number of correctly classified respective the number of misclassified samples in each data set. The first number in the parentheses, shown in columns 2 and 3, represents the number of cancer patients in each data set. This number is followed by a second number which represents the number of non-cancer patients in each of the data sets. Column 4 presents the error percentage gained during the classifications. The three last columns in this Table present the same kind of information for the five test sets, each containing 15 samples. Note that the distribution of cancer and non-cancer patients varies in each of the test sets.

Column 5 and 6 in Table 2 show the number of correctly classified and misclassified patients in the five different test sets. In parentheses, also shown in these two columns, is the number of cancer patients in each data set. This number is followed by the number of non-cancer patients in each of the data sets. The same kind of information is shown in columns 2 and 3 for the training sets. The number of classification errors and their corresponding error percentage of the total number of cases are shown in column 3 and 4 for the training sets, and in column 6 and 7 for the test sets.

In order to further evaluate the performance of the decision tree approach, when using gene expression data as input, the mean value for the error percentages, shown in column 7 for the five test sets shown in Table 2, was calculated:

$$\frac{(13.3\% + 13.3\% + 0.0\% + 13.3\% + 13.3\%)}{5} = 10.64\ \%$$

This calculated mean value is a measure of the risk of the decision tree making an incorrect classification. In other words, the above decision tree approach has the capacity of correctly classifying samples approximately 89 % of the time.

5.2.2 Risk factors as input data

Despite the numerous risk factors for the development of breast cancer that have been investigated, only a few demonstrate a clear association with breast cancer development. Four probable risk factors associated with breast cancer are investigated in this project.

By using the risk factor values of age, BMI, alcohol status, and smoking status for the 75 patients as input to the C4.5 algorithm, it may be possible to shed light on the risk factors that are most important in the development of breast cancer. Thus by only considering the values of the four mentioned risk factors, it may be possible to separate non-cancer patients from patients suffering from breast cancer in the data set.

The risk factors that were used as input resulted in the pruned decision tree shown in Figure 12. As can be seen in Figure 12, the values for smoking respective alcohol status of a patient in Gene Logic can be set to one of the six following attribute-values: ‘current use’, ‘no current use’, ‘never used’, ‘previous use’, ‘occasional’, and ‘unknown’.

The distribution of the 75 patients, with respect to the six possible attribute-values describing alcohol and smoking status in Gene Logic, is shown in Table 3.

	Current use	**No current use**	**Never used**	**Previous use**	**Occasional**	**Unknown**
Alcohol status	4	17	5	0	14	35
Smoking status	5	0	22	12	0	36

Table 3. Distribution of patients with respect to the attribute-values used when describing alcohol and smoking status. Since the risk factor data contains information about 75 patients, the sum of each row is 75.

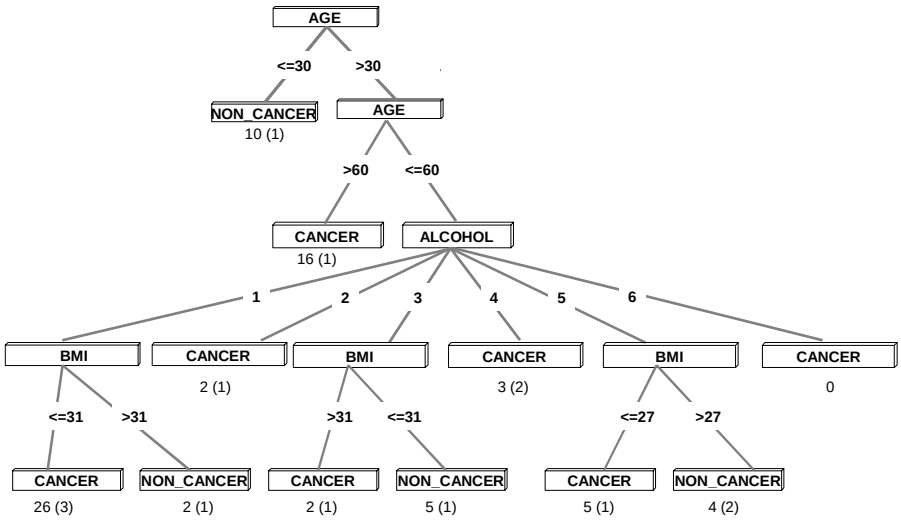


Figure 12: Pruned decision tree created when risk factor data was used as input. As it can be seen, the decision tree has chosen a patient's age as the most significant attribute in order to separate cancer patients from non-cancer patients in the data set. A patient's alcohol status has been chosen as the second most informative attribute whereas the BMI value is considered to be the third most important risk factor for the classification of samples. In the above decision tree, the possible attribute-values for alcohol status are denoted by the numbers 1-6. Number 1 stands for the value 'unknown', 2 stands for 'current use', 3 stands for 'occasional', 4 stands for 'never used', 5 stands for 'no current use' whereas 6 stands for 'previous use'. Similar to Figure 11, the first number beneath each node of the tree denotes the number of instances which belong to that path in the tree. This number may be followed by a second number in a parenthesis. This second value denotes the number of classification errors encountered out of the total number of classifications made from the data in that particular path of the decision tree.

The tree in Figure 12 is consistent with the fact that breast cancer is age-dependent and that the risk of getting breast cancer increases as women get older. According to the decision tree, a patient's age is the most significant risk factor associated with breast cancer. The second most significant risk factor appears to be the alcohol use. The BMI values have also been chosen as rather significant risk factors in the decision tree. However, the decision tree in Figure 12 is the simplified, pruned version of the decision tree created when risk factor data is used as input. It is interesting to also study the un-pruned version of the tree since pruning sometimes results in more complex trees in an attempt to reduce the number of unnecessary nodes. The un-pruned decision tree is shown in Figure 13.

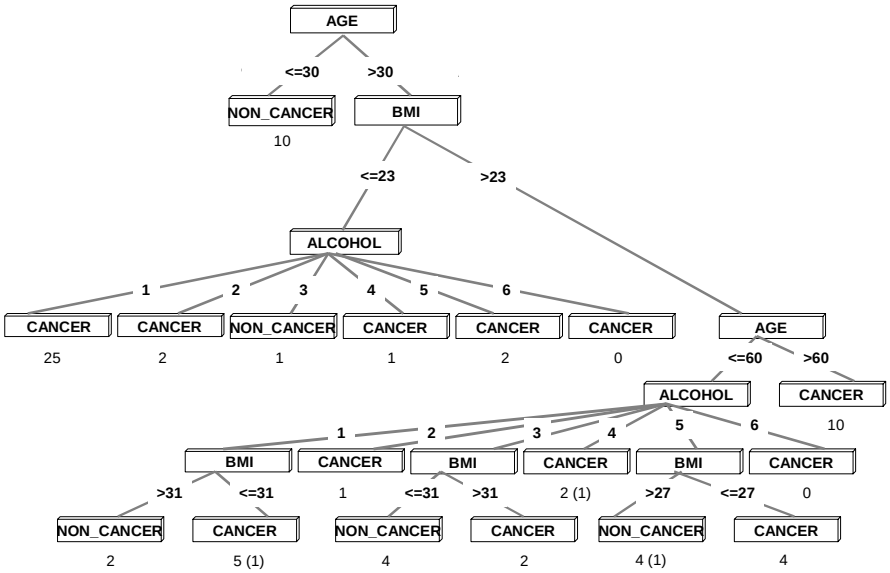


Figure 13. The un-pruned decision created when risk factor data was used as input. This tree has chosen a patient's age as the most significant attribute in order to correctly classify the samples in the data set. In this tree, the BMI value is considered to be more informative compared with the alcohol status, for the separation of samples. Similar to Figure 12, the possible attribute-values for alcohol status are denoted by the numbers 1-6. Number 1 stands for the value 'unknown', 2 stands for 'current use', 3 stands for 'occasional', 4 stands for 'never used', 5 stands for 'no current use' whereas 6 stands for 'previous use'. As described earlier, the first number beneath each node of the tree denotes the number of instances, out of the total number of cases, which belong to that path in the tree. This number may be followed by a second number in a parenthesis. This second value denotes the number of classification errors encountered out of the total number of classifications made from the data in that particular path of the decision tree.

As it can be seen in Figure 13, the un-pruned decision tree contains eight more tree nodes compared with the pruned version of the tree. Similar to the pruned decision tree, the tree in Figure 13 has chosen the attribute 'age' as the most significant risk factor. The un-pruned decision tree has however, instead of alcohol status, chosen the BMI value as the second most significant risk factor involved in breast cancer.

The production rules derived by the C4.5 algorithm, when risk factor data was used as input, are presented below. As was described in Section 2.5, the derived production rules do not always exactly correspond to paths shown in the decision tree. This can for instance be seen in the first production rule below. According to the decision tree in Figure 13, a BMI of 23 is an informative attribute-value in order to correctly classify the 75 patients. This can be seen in Figure 13 where this attribute is used as the second most important attribute for the

classification of 65 out of the 75 patients. In the first production rule below, however, a BMI of 31 has been chosen instead of 23.

```
1. Alcohol_status = UNKNOWN
   Age > 30
   BMI <= 31
      → class CANCER

2. Age > 60
      → class CANCER

3. Age <= 30
     → class NON_CANCER

4. Alcohol_status = UNKNOWN
   BMI > 31
     → class NON_CANCER

5. Alcohol_status = NO_CURRENT_USE
   Age > 30
   BMI <= 27
     → class CANCER

6. Alcohol_status = OCCASIONAL
   Age <= 60
   BMI <= 31
     → class NON_CANCER

7. Alcohol_status = NO_CURRENT_USE
   Age <= 60
   BMI > 27
     → class NON_CANCER
```

The above production rules are derived in the same way as decribed in Section 5.2.1. These rules are further analyzed in Section 6.2.

In the same way as in Section 5.2.1, a 5-fold cross validation was performed in order to get an indication of how well the above decision tree approach will do when it is asked to make new predictions on data it has not already seen. The patients in each data set used during cross validation, were randomly chosen. Each of the five training sets contained 60 patients, whereas the test sets contained 15 patients each. The results are presented in Table 4.

Data set	Correctly classified	Misclassified	Error	Correctly classified	Misclassified	Error
1	56 (47/9)	4 (1/3)	6.7%	6 (4/2)	9 (1/8)	60.0%
2	58 (43/15)	2 (1/1)	3.3%	9 (4/5)	6 (5/1)	40.0%
3	57 (37/20)	3 (1/2)	5.0%	15 (15/0)	0 (0/0)	0.0%
4	57 (39/18)	3 (1/2)	5.0%	14 (13/1)	1 (0/1)	6.7%
5	59 (42/17)	1 (0/1)	1.7%	10 (9/1)	5 (2/3)	33.3%

Table 4: Columns 2-4 present the classification results for the 5 training sets, each containing 60 samples. Columns 2 and 3 give information about the number of correctly classified respective the number of misclassified samples in each data set. The first number in parenthesis, shown in columns 2 and 3, represents the number of cancer patients in each data set. This number is followed by a second number which represents the number of non-cancer patients in each data set. Column 4 presents the error percentage gained during the classifications. The three last columns in this Table present the same kind of information for the five test sets, each containing 15 samples. Note that the distribution of cancer and non-cancer patients varies in each of the test sets.

In order to evaluate the classification performance of the decision tree containing risk factor data, the mean value for the error percentages of the five test sets, shown in column 7 in Table 4, was calculated:

$$\frac{(60.0\% + 40.0\% + 0.0\% + 6.7\% + 33.3\%)}{5} = 28\ \%$$

The above calculated mean value indicates that the decision tree approach has the capacity of classifying approximately 70% of ‘new’ samples correctly.

In order to further investigate the classification performance of the C4.5 decision tree algorithm, when risk factor data is used as input, an additional decision tree was created. This decision tree was created by removing all the patients having the value ‘unknown’ as an attribute-value in the data set. In this way, a data set containing 39 samples, whereas 23 cancer patients and 16 non-cancer patients, was created. The un-pruned decision tree created by this data set is shown in Figure 14.

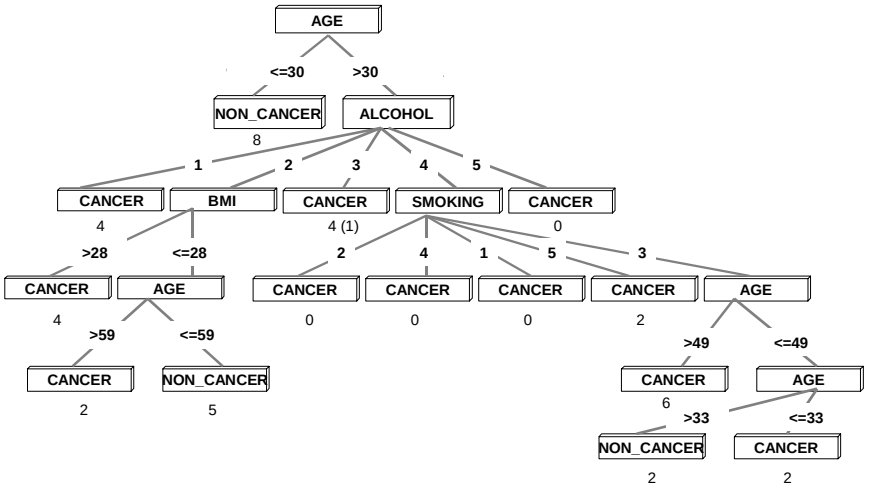


Figure 14: The un-pruned decision tree created when the data set does not contain the value 'unknown' as an attribute value. According to this tree, a patient's age is the most informative attribute in the data set, for the classification of samples. The alcohol status, BMI value and smoking status are chosen as the second, the third and the fourth most informative attributes. In the above decision tree, the possible attribute-values for both alcohol respective smoking status, are denoted by the numbers 1-5. Number 1 stands for 'current use', 2 stands for 'occasional', 3 stands for 'never used', 4 stands for 'no current use' whereas 5 stands for 'previous use'. Similar to Figure 11, the first number beneath each node of the tree denotes the number of instances which belong to that path in the tree. This number may be followed by a second number in a parenthesis. This second value denotes the number of classification errors encountered out of the total number of classifications made from the data in that particular path of the decision tree.

Another interesting result is the appearance of the pruned version of the above decision tree, which is shown in Figure 15.

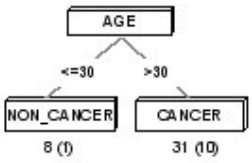


Figure 15: The pruned decision tree created when the data does not contain the value 'unknown' as an attribute value. According to this decision tree, the only attribute necessary for the classification of samples in the data set, is the age. In the same way as the earlier presented decision trees, the first number beneath each node of the above tree denotes the number of instances, out of the total number of cases, which belong to that path in the tree. This number may be followed by a second number in parenthesis. This second value denotes the number of classification errors encountered out of the total number of classifications made from the data in that particular path of the decision tree.

The pruned decision tree in Figure 15 is greatly simplified but at the same time it really emphasizes the importance of the 'age' attribute as a significant risk factor of breast cancer. Another difference compared with the earlier results, gained when the input data contained the value 'unknown' as an attribute value, is that the derived production rules are now much easier to both interpret and to evaluate. These derived production rules, gained when risk factor data is used as input to the C4.5 algorithm, are shown below.

```
1. Age > 59
       → class CANCER


2. Alcohol_status = NO_CURRENT_USE
   Age > 49
       → class CANCER

3. Alcohol_status = CURRENT_USE
       → class CANCER

4. Age <= 30
       → class NON_CANCER

5. Alcohol_status = OCCASIONAL
   Age <= 59
   BMI <= 28
       → class NON_CANCER
```

The above production rules are derived in the same way as decribed in Section 5.2.1. These rules are further analyzed in Section 6.2.

However, since the data set now only contained 39 samples, a 3-fold cross validation was performed. This was done in order to evaluate the accuracy of the decision tree approach in predicting 'new' samples, when the input data lacks the value 'unknown'. As mentioned earlier, the patients in each data set used during cross validation, were randomly chosen. Each training set contained 26 samples, whereas the test sets contained 13 samples each. The trees created during the cross validation procedure again identified the 'age' as the most informative attribute for the classification of patients. The classification results are shown in Table 5.

Data set	Correctly classified	Misclassified	Error	Correctly classified	Misclassified	Error
1	25 (17/8)	1 (0/1)	3.8%	13 (6/7)	0 (0/0)	0.0%
2	25 (16/9)	1 (0/1)	3.8%	11 (7/4)	2 (0/2)	15.4%
3	24 (13/11)	2 (0/2)	7.7%	10 (8/2)	3 (2/1)	23.1%

Table 5: Columns 2-4 present the classification results for the 3 training sets, each containing 26 samples. Columns 2 and 3 give information about the number of correctly classified respective the number of misclassified samples in each data set. The first number in parenthesis, shown in columns 2 and 3, represents the number of cancer patients in each data set. This number is followed by a second number which represents the number of non-cancer patients in each of the data sets. Column 4 presents the error percentage gained during the classification. Columns 5-7 in this Table present the same kind of information for the 3 test sets, each containing 13 samples. Note that the distribution of cancer and non-cancer patients varies in each of the test sets.

The mean value for the error percentages of the three test sets, shown in column 7 in Table 5, is:

$$\frac{(0.0\% + 15.4\% + 23.1\%)}{3} = 12.83\ \%$$

The capacity of the decision tree approach in correctly classifying 'new' samples is thus considered to be approximately 87%.

5.2.3 Expression of breast-related genes and risk factors as input data

After creating a decision tree that presented the significant genes involved in breast cancer and also a decision tree identifying the most significant risk factors in the development of breast cancer, it is now interesting to put together the two sets of input data. By using both the expression of breast-related genes *and* the risk factor values for the 75 patients as input, it may be possible to create a decision tree where the nodes can correspond to both genes and risk factors that may be involved in breast cancer. Thus, the decision tree may be able to separate cancer patients from non-cancer patients considering both expression patterns and risk factor data.

Before creating any decision trees, the data set containing risk factor information was divided into two sets. This was done in order to examine the capacity of the C4.5 algorithm in classifying patients with and without the value 'unknown' as a possible attribute-value. The first data set contained risk factor information for all the 75 samples and therefore contained the value 'unknown'. The second data set on the other hand, only contained risk factor information about the patients lacking the value 'unknown'. This latter data set thus consisted

of only 39 samples (see Section 5.2.2). The gene expression data and the first set of risk factor data, containing information about all 75 samples, were put together as one input to the C4.5 algorithm. This input resulted in a decision tree identical to the decision tree shown in Figure 11. The resulting decision tree thus again identified the four genes MKI67, BAX, SNCG, and AR, as the most valuable attributes for classification of the 75 samples. In other words, none of the four risk factors were considered to be relevant for the classification. However, the merge of gene expression data and the second set of risk factor data mentioned above, containing risk factor information for the 39 samples where the attribute-value 'unknown does not exist, resulted in a different decision tree, which is shown in Figure 16.

The resulting decision tree in Figure 16 contains three genes and one risk factor. According to the tree, these four attributes are thus considered to be the most relevant in order to separate cancer patients from non-cancer patients in the data set containing 39 samples.

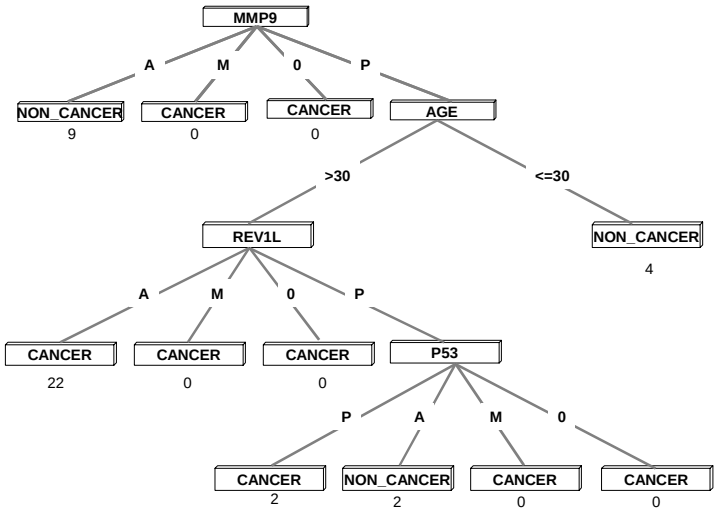


Figure 16: Decision tree created when gene expression data and risk factor data of 39 samples were used as input. This tree identifies three genes and one risk factor. These four attributes are thus considered to be the most relevant in order to correctly classifying samples in the data set. Similar to Figure 11, the symbol 'P' in the above decision tree stands for 'present' and indicates when a certain gene is expressed. 'A' stands for 'absent' and is used when the expression of a gene is not detected. The 'M' symbol denotes an expression between the symbols 'P' and 'A'. Finally, '0' denotes a situation where the expression of a gene has not been examined for the patient. As earlier, the number beneath each node of the above tree denotes the number of instances which belong to that path in the tree. This number may be followed by a second number in a parenthesis. This second value denotes the number of classification errors encountered out of the total number of classifications made from the data in that particular path of the decision tree.

In the above decision tree, the expression of the matrix metalloproteinase-9 (MMP9) has been chosen as the most obvious marker of breast cancer. According to the decision tree, the second most important attribute for the classification of cancer patients is the age attribute. The significance of this risk factor has also been shown by all the created decision trees shown in Section 5.2.2. The decision tree also indicates that the absence of expression of the REV1-like gene is highly associated with cancer patients, in the data set used. The nuclear protein p53 is also considered to be important for the classification of samples.
The set of production rules induced by the C4.5 algorithm, when both gene expression data and risk factor data lacking the value 'unknown' are used as input, is presented below.

```
1. MMP9  = P
   REV1L = A
   Age > 30
          ->  class CANCER

2. MMP9 = A
          ->  class NON_CANCER

3. Age <= 30
          ->  class NON_CANCER
```

The three above production rules are derived in the same way as decribed in Section 5.2.1. These rules are further analyzed in Section 6.3.

Similar to Section 5.2.1 and 5.2.2, a cross validation procedure was performed in order to evaluate the accuracy of the decision tree approach in predicting 'new' samples, when the input data consists of both gene expression data and risk factor data. Since the data set contained only 39 samples, a 3-fold cross validation was performed. As earlier, the patients in each data set during cross validation were randomly chosen. Each training set contained 26 patients, whereas the test sets contained 13 patients each. The classification results are shown in Table 6.

Data set	Correctly classified	Misclassified	Error	Correctly classified	Misclassified	Error
1	24 (18/6)	2 (0/2)	7.7%	10 (4/6)	3 (1/2)	23.1%
2	25 (16/9)	1 (0/1)	3.8%	12 (8/4)	1 (0/1)	7.7%
3	25 (13/12)	1 (0/1)	3.8%	12 (10/2)	1 (0/1)	7.7%

Table 6: Columns 2-4 present the classification results for the 3 training sets, each containing 26 samples. Columns 2 and 3 give information about the number of correctly classified respective the number of misclassified samples in each data set. The first number in parenthesis, shown in columns 2 and 3, represents the number of cancer patients in each data set. This number is followed by a second number which represents the number of non-cancer patients in each of the data sets. Column 4 presents the error percentage gained during the classification. Columns 5-7 in this Table present the same kind of information for the 3 test sets, each containing 13 samples. Note that the distribution of cancer and non-cancer patients varies in each of the test sets.

According to the error percentages of the three test sets, shown in column 7 in Table 6, the mean value is:

$$\frac{(23.1\% + 7.7\% + 7.7\%)}{3} = 12.83\ \%$$

According to the mean value calculated above, the capacity of this decision tree approach in classifying 'new' samples is approximately 87%.

5.2.4 Indication of pathways involving known breast cancer genes

There are also other ways to apply the technique of decision trees on genes involved in breast cancer. By using the expression of a putative breast cancer gene, for example, p53, as a decision-variable in a decision tree, it may be possible to suggest genes occurring in the same pathways as the putative breast cancer gene. In other words, it may be possible to suggest how the expression of certain genes may effect the expression of the putative breast cancer gene, which is used as decision variable. In this way different decision trees can be created for each putative breast cancer gene. However, the approach performed here is not an attempt to illustrate the exact signalling pathways for each putative breast cancer gene. The reconstruction of exact signalling pathways is here regarded as a possible future work.

As described in Section 5.2.1, the tree based on gene expression data identified four genes, out of the 108 genes tested, which were considered to be involved in breast cancer. The expression of these four genes, MKI67, BAX, SNCG, and AR, are therefore chosen as decision variables in different trees. This is done in order to study the possibility of suggesting some of the genes involved in the same signalling pathways as these four genes.

The gene expression data that is used as input to the C4.5 algorithm contains abscall values for at least two different probes for each of the four studied genes. A probe is a sequence of single-stranded DNA that represents a segment of a gene. Different probes of a gene thus represent different sub-sequences of the same gene (see Figure 17).

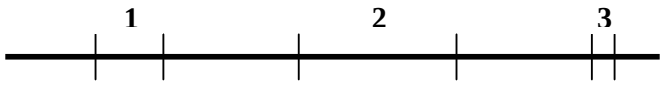


Figure 17: The line represents a gene. The 3 marked segments on this gene represent three different probes. A probe is thus a short sequence of single-stranded DNA that represents a segment of a gene.

In order to identify genes involved in the same signalling pathways, a decision tree and a set of production rules are created for each probe of the four genes studied. In other words, each gene results in more than one decision tree. The data set contains 2 probes of the MKI67 gene, 4 probes of the BAX gene, 2 probe of the AR gene, and finally 3 probes of the SNCG gene. In this way, totally 11 different decision trees and sets of production rules are created. Instead of illustrating each of the 11 trees and rule sets, only some of the production rules created for each of the four genes are presented below. The production rules that are presented below are regarded as relevant by considering how many samples in the data set that were correctly classified by each rule.

5.2.4.1 The MKI67 nuclear antigen

```
1. BCAS1 = A
   RAD51 = A
         ->  class OFF

2. RAD51 = P
         ->  class ON
```

The two production rules shown above indicate the genes that may be involved in the signalling pathways of MKI67, according to the first probe of MKI67. The values ‘OFF’ and ‘ON’ denote the expression of MKI67. In other words, the expression of MKI67 is considered to be ‘absent’ when the value ‘OFF’ is used. Consequently, the expression of MKI67 is considered to be ‘present’ whenever the value ‘ON’ is used. The abscall values, ‘present’ and ‘absent’, have thus been replaced by ‘ON’ respective ‘OFF’.

According to the two above presented rules, the expression of MKI67 is denoted by ‘absent’ whenever the expression of BCAS1 and RAD51 is ‘absent’, whereas the MKI67 gene is considered to be ‘present’ whenever the expression of RAD51 is ‘present’.
The two below production rules show the most significant production rules, considering how many samples in the data set that were correctly classified by each rule, derived by the second probe of MKI67.

```
1. TSG101 = P
   MKI67  = A
        ->  class OFF

2. MKI67 = P
   MYC   = P
        ->  class ON
```

As can be seen, the above production rules present two other genes that may be involved in the signalling pathways as MKI67. These genes are TSG101 and MYC. As can further be seen, the expression of MKI67 gained with the first probe of MKI67 has also been identified in both rules presented above. This is of course because the two probes are very much correlated, since they represent the same gene (see Figure 17). The production rules above were gained when the expression of the second probe of MKI67 was used as decision variable in the C4.5 algorithm.

5.2.4.2 The BCL2-associated X protein (BAX)

```
1. HRAS = P
   BAX  = P
       ->  class ON

2. BAX  = A
   BRAK = A
       ->  class OFF

3. BRAK = P
      ->  class ON
```

According to the above production rules, the genes HRAS and BRAK may be involved in the signalling pathways of BAX. These results were gained when the expression of the first probe of the BAX gene was used as decision variable. As described in Section 5.2.4.1, some of the gained production rules may contain other probes of the gene that is used as the decision variable. In the rules above, the expression of BAX gained with other probes of the gene is shown in the first and the second rule.

The below production rule is the most relevant production rule, that is, the one correctly classifying the largest number of samples in the data set, derived by the second probe of the BAX gene.

```
1. HRAS = P
   BAX  = P
   MMP9 = P
        ->  class ON
```

According to the above rule, the expression of the genes HRAS and MMP-9 may be involved in the signalling pathway of the BAX gene. The expression of BAX gained with other probes of the gene is also shown in the above rule.

In the same way as described above, that is, by considering how many samples in the data set that are correctly classified by each rule, the below production rule is considered to be the most relevant rule derived by the expression of the third probe of the BAX gene.

```
1. NRAS = P
   PHB  = P
      ->  class ON
```

According to the above production rule, the expression of the BAX gene may be present whenever the expression of the PHB and the NRAS gene is present.

The below production rule presents two other genes that may be involved in the signalling pathways of the BAX gene. This rule was gained when the expression of the fourth probe of the BAX gene was used as decision variable in the C4.5 algorithm.

```
1. P53  = A
   SNCG = A
        ->  class OFF
```

According to the above rule, the expression of the BAX gene may be absent whenever the expression of the genes p53 and SNCG is absent.

5.2.4.3 The Androgen receptor (AR)

```
1. AR       = P
   SDBCAG84 = A
   AGR2     = P
   AGR2     = P
   UP       = P
        ->  class ON
```

```
2. AR   = A
   LIG4 = P
        ->  class OFF
```

The two production rules shown above present the genes that may be involved in the signalling pathways of androgen receptor (AR), according to the first probe of AR. As it was described for the production rules shown earlier, both the production rules presented above contain other probes of the gene that is used as the decision variable. In other words, the expression of AR, gained with the other probes of the gene, is shown in both rules presented above. The first production rule above shows three candidate genes that may be involved in the activation of the androgen receptor. These genes include SDBCAG84, AGR2, and UP. According to the second production rule above, the expression of AR may be absent whenever the expression of ATP-dependent DNA ligase IV (LIG4) is present.

The below production rules were gained when the expression of the second probe of the AR gene was used as decision variable.

```
1. AR = P
     ->  class ON

2. AR = A
     ->  class OFF
```

Both the production rules shown above have only identified the first probe of the AR gene as the most relevant factor in the signalling pathway of AR. In other words, the above production rules do not contribute with any new information about genes involved in the signalling pathway of AR since both probes represent the same gene.

5.2.4.4 Synuclein-gamma (SNCG)

```
1. NCOA3 = P
   RAD51 = A
   FGFR1 = P
   IL24  = A
   TFF1  = A
         ->  class OFF
```

When the first probe of the Synuclein-gamma (SNCG) gene was used as decision variable, the five genes presented in the production rule above were identified to be involved in the signalling pathway of SNCG.

According to the second probe of the SNCG gene, two other genes may be involved in the same signalling pathways as SNCG. These genes are shown in the production rule below.

```
1. CDH13 = P
   XLKD1 = P
        ->  class ON
```

The below production rule was derived with the third probe of the SNCG gene.

```
1. UP = P
     ->  class OFF
```

According to the above production rule, the expression of the SNCG gene may be absent whenever the expression of UP is present.

6. Analysis of results

In this Chapter, the gained decision trees and production rules shown in Section 5.2 will be further analyzed.

6.1 Expression of 108 breast-related genes as input data

A total of six production rules were derived from the approach where gene expression data was used as input to the C4.5 algorithm. These rules are shown in Section 5.2.1. According to the first production rule, the presence of the expression of MKI67 (antigen identified by monoclonal antibody Ki-67) is a marker of patients suffering from breast cancer in the data set used. Ki-67 is an antibody that reacts with the nuclear antigen MKI67, which is expressed in proliferating cells but not in quiescent cells. Consequently, the antibody is used in tumor pathology to detect proliferating cells (Schluter et. al 1993). Since tumor cells are characterized by an uncontrolled cell growth, it is only logical for the expression of MKI67, only expressed in proliferating cells, to be a marker of breast cancer patients.

According to the second rule shown in Section 5.2.1, the expression of the BAX gene is also a marker of breast cancer patients in the data set. The protein produced by the BAX gene is a key regulator of apoptosis, that is, programmed cell death. According to Reed (1996), this pro-apoptotic gene is normally expressed in breast tissue. This statement then explains the path identified by the second production rule.

The third production rule contains the expression of all four genes shown in the decision tree in Figure 11. In other words, all four genes are considered to be relevant for the classification of non-cancer patients in the data set. These four genes thus include: AR, BAX, SNCG, and MKI67. According the third production rule, the expression of AR and SNCG are presents in non-cancer patients whereas the expression of BAX and MKI67 are absent. The androgen receptor (AR) is a member of the family of steroid receptors. This receptor binds androgens and acts as a transcription factor. Mutations in the AR gene have been observed in patients with male breast cancer (Lobaccaro et. al. 1993). Synuclein-gamma (SNCG) on the other hand is a member of the human synuclein gene family. The synuclein-gamma gene was recently found to be overexpressed in advanced infiltrating carcinoma of the breast (Lavedan et. al. 1998). Exactly how the expression of these two genes, AR and SNCG, is relevant for the classification of non-cancer patients in the data set is however yet unclear.

As further shown in the third production rule, the expression of MKI67 is considered to be absent in non-cancer patients. As mentioned earlier, MKI67 is only expressed in proliferating cells and thus not in quiescent cells. Since non-cancer cells do not undergo as much proliferation as tumor cells, the path “if MKI67=A then non_cancer” is considered to be correct. The last path identified by the third rule is: “if BAX = A then non_cancer”. This path could also be explained by the fact that non-cancer cells do not undergo as much proliferation as tumor cells and thus the expression of the pro-apoptotic gene BAX does not have to be present in non-cancer cells. In other words, it is not necessary for the expression of BAX to be present in non-cancer cells since there is no need to kill any out of control proliferating tumor cells via the apoptosis machinery.

Similar to the second production rule described above, the fourth production rule shown in Section 5.2.1 again indicates that the presence of the expression of BAX is a marker of breast cancer patients in the data set. According to the fourth rule, the absence of the expression of androgen receptor (AR) is another marker of breast cancer patients.

All the paths shown in the fifth production rule have already been identified by the third production rule which was described earlier in this Section. The expression of the three genes shown in the fifth rule, that is, the expression of BAX, SNCG, and MKI67, as a way of classifying cancer patients has therefore already been analyzed above.

Finally, the last derived production rule shows that the absence of the expression of synuclein-gamma (SNCG) is a valuable marker for patients suffering from breast cancer in the data set. As described earlier, the SNCG gene has been recently found to be overexpressed in advanced infiltrating carcinoma of the breast (Lavedan et. al. 1998). Overexpression of SNCG may indicate breast cancer malignant progression from benign breast or in situ carcinoma to the highly infiltrating carcinoma (Ji et. al. 1997). However, even though the relation shown by the sixth production rule does not agree with the information gained from studied literature, the relation between breast cancer and the expression of the SNCG gene should be further investigated.

As described in Section 5.2.1, a cross validation was performed in order to evaluate the decision tree approach based on gene expression data. The results from the cross validation procedure are shown in Table 2. According to the mean value calculated from the misclassifications done on test data, the decision tree approach has the capacity to correctly classify samples approximately 89% of the time. Is this a good classification result? One way to answer this question is by comparing this result with result from a simple randomized

classification approach. If it is possible with the simple algorithm to correctly classify 89% of the samples, then the gained classification result will be considered as insignificant.

The simple randomized algorithm is one that contains information about the distribution of cancer and non-cancer patients in the data set. As mentioned earlier, there are 53 cancer patients and 22 non-cancer patients in the data set that is used. This means that the data set contains approximately 71% cancer patients and 29% non-cancer patients. Since the data set contains a larger proportion of cancer patients, the randomized algorithm would probably create a default class and thus eventually classify *all* patients as cancer patients. This would then result in the algorithm classifying only 71% of the samples correctly. Comparisons with this randomized algorithm answers the question asked earlier. The decision tree approach based on gene expression data, with 89% chance of correctly classifying a 'new' sample, is thus considered to be useful. In other words, the C4.5 decision tree algorithm is considered to be a useful classification technique in order to find the most important markers of breast cancer and to consequently separate cancer patients from non-cancer patients with the kind of gene expression data used in Gene Logic.

As mentioned earlier, column 6 in Table 2 shows the number of misclassified cancer and non-cancer patients in each of the test sets during the cross validation procedure. By studying these numbers it is possible to examine the tendency of the C4.5 algorithm to misclassify cancer patients. In real life, the consequences are of course much more fatal if a cancer patient is misclassified as a non-cancer patient, than vice versa. It is it is therefore more important for the algorithm to reduce the number of misclassified cancer patients than the number of misclassified non-cancer patients. In other words, it is more important for the first number in parentheses to be low than the second number. Of course, in a perfect classification result the overall number of misclassified patients would be zero. However, as can be seen in Table 2, the tendency of the C4.5 algorithm to misclassify cancer patients compared with non-cancer patients, based on gene expression data, is equal.

As can further be seen in Table 2, the distribution of cancer and non-cancer patients differs in each of the test sets during the cross validation. This unequal distribution of patients may thus have affected the gained classification results. The results shown in Table 2 however, do not indicate any variation in the C4.5 classification performance depending on the distribution of patients in the test sets. For instance, as can be seen in columns 5-7 in Table 2, the third test set contains 13 cancer and 2 non-cancer patients. According to Table 2, this distribution has resulted in an error percentage of 0.0%. In other words, this result indicates that there are no misclassifications done when the test set consists of more cancer

patient than non-cancer patients. However, the second test set in the cross validation procedure shown in Table 2 contains 15 cancer patients and thus no non-cancer patients. Yet, according to Table 2, the error percentage gained with this distribution is 13.3%. The error percentage gained with the second test set is thus higher than the error percentage gained with the third test set, even though the second test set only consists of cancer patients. This result may indicate that the C4.5 classification performance is not affected by the distribution of patients in the test sets.

6.2 Risk factors as input data

Unfortunately, the risk factor data that was used for the creation of the decision tree shown in Figure 12 was quite incomplete and thus contained a large number of attributes where the values were ‘unknown’. Another weakness of the input data was that there were only a few patients in each group that was created with respect to the alcohol and smoking status (see Table 3). These factors complicate the interpretation and evaluation of the decision trees shown in Figure 12 and 13.

As can be seen in Figure 12, the second most significant risk factor of breast cancer appears to be the alcohol use. This claim is however rather uncertain since several patients have the value ‘unknown’ as their alcohol status and therefore the decision tree probably has interpreted this value as an ordinary measure of a person’s alcohol consumption habits. This could be a reason why the decision tree contains more misclassifications (errors) compared with the decision tree shown in Figure 11. The BMI attribute has also been chosen as a rather significant risk factor in Figure 12. It is yet difficult to interpret the association between the value of BMI and the risk of breast cancer through the created decision tree.

The un-pruned version of the decision tree, based on risk factor data, is shown in Figure 13. This un-pruned version is analyzed in an attempt to increase the understanding of the pruned decision tree based on risk factor data, shown in Figure 12. However, when looking further at both the pruned and the un-pruned decision tree, it is obvious that the risk factor data splits the data into small subsets where there is no order. A reason to this disorder could be the ‘unknown’ values in the data set, which makes it difficult to find any regularities in data when trying to construct correct and interpretable decision trees. An interesting factor in both the pruned and un-pruned decision trees is however, that neither of the decision trees has chosen smoking status as a significant risk factor associated with breast cancer. This claim may very well be correct since according to most studies there is presently no strong support

for a relationship between cigarette smoking and the risk of breast cancer (Mesko et. al. 1990).

The seven production rules derived from the approach where risk factor data, containing the value 'unknown', was used as input to the C4.5 algorithm are shown in Section 5.2.2. The second and the third derived production rule emphasize the fact that only a few women under the age of 30 develop breast cancer, whereas older women have an increased risk of developing the disease. This assertion has been pointed out by several other studies (Lidbrink, 2001).

The resulting decision trees and production rules, gained when risk factor data containing the value 'unknown' is used as input to the C4.5 algorithm, indicate a quite indistinct role for the value of BMI as a possible risk factor of breast cancer. This can be due to the fact that there appears to be a difference in the effect of BMI on pre-menopausal women compared with post-menopausal women, see Section 2.3.1.1. According to several studies, a higher BMI is weakly associated with a *decreased* risk of pre-menopausal breast cancer whereas a higher BMI also is associated with an *increased* risk of postmenopausal breast cancer (Trentham-Dietz et. al. 1997). In this project, the patients in the input data had not been separated according to their menopausal status. This probably further complicated the process of finding any associations between the values of BMI and the risk of breast cancer.

The results from the 5-fold cross validation that was performed in order to evaluate the decision tree approach where risk factor data was used as input to the C4.5 algorithm are shown in Table 4. As can be seen in Section 5.2.2, the mean value for the error percentages of the five test sets shown in Table 4 was calculated. The results indicate that the decision tree approach has the capacity to correctly classify samples approximately 70% of the time. However, in Section 6.1 a randomized algorithm was described where the sample distribution of the input data was known and the algorithm was thus predicted to correctly classify 71% of 'new' samples**.** The classification performance of the decision tree approach based on risk factor data, containing the value 'unknown', is thus considered to be just as good as the classification performance of the randomized algorithm.

As further described in Section 6.1, column 6 in Table 4 shows the number of misclassified cancer and non-cancer patients in each of the test sets during the cross validation procedure. As can be seen in Table 4, the distribution of cancer and non-cancer patients differs in each of the test sets during cross validation. According to columns 5 through 7 in Table 4, the tendency of the C4.5 algorithm to misclassify patients, based on risk factor data,

appears to be a bit lower when a test set consists of more cancer patients compared with non-cancer patients.

As further described in Section 5.2.2, an additional decision tree was created where all the samples having the value ‘unknown’ as a measure in the risk factor data set were removed. The final data set thus only contained 39 samples. The resulting un-pruned decision tree is shown in Figure 14, whereas the pruned version is shown in Figure 15.

As can be seen in Section 5.2.2, the decision tree in Figure 14 selects all four risk factors as important for the separation of cancer patients and non-cancer patients in the data set. Figure 14 is again consistent with the fact that breast cancer is very much age-dependent. Similar to the decision tree shown in Figure 12, this tree has chosen the values of alcohol status and BMI as the second respective the third most relevant risk factor involved in breast cancer. The interesting difference compared with two the other decision trees, shown in Figure 12 and 13, is however that this tree has chosen to use the attribute ‘smoking status’ in the resulting tree. In other words, according to the decision tree shown in Figure 14, the values of smoking status is considered to contribute to the separation of cancer patients from non-cancer patients in the data set. According to the un-pruned tree, the smoking status of a patient is however the attribute that the least contributes with the classification of samples, since the attribute is used as a node rather far down in the decision tree.

The five resulting production rules, derived from the input data lacking the value ‘unknown’ as an attribute value, are shown in Section 5.2.2. The derived production rules are considered to agree with recent studies, involving risk factors and the risk of breast cancer. According to the second rule, patients above the age of 49 have breast cancer, even though they have no current use of alcohol. This must be due to the high age of the patients. According to rule 3, most patients with a current use of alcohol suffer from breast cancer. However, according to most studies, the relationship between alcohol and the risk of breast cancer exists only for a relatively high consumption (Ranstam & Olsson, 1995). This rule should therefore be taken with a large grain of salt since it is the *amount* of alcohol that matters the most.

According to the fifth rule presented above, most patients in the data set that are under the age of 59, who are not over-weight and who only have an occasional use of alcohol, do not suffer from breast cancer. The statements in this rule are considered to agree with most of the previous studied literature where the effects of the age, BMI, and alcohol status on the risk of breast cancer are discussed, see Section 2.3.1.1. Also, as further shown in Section 5.2.2,

the first and the fourth production rule again show the significance of a patient's age for the classification of samples in the data set.

In order to evaluate the accuracy of the decision tree approach described in Section 5.2.2, where the input data lacks the value 'unknown' as an attribute value, a cross validation procedure was performed. The results are shown in Table 5. According to the mean value calculated for the misclassifications done on test data, the capacity of the decision tree approach in classifying 'new' samples is approximately 87%. This classification result is thus better than the classification result gained with the data set containing the value 'unknown' as an attribute value. As described earlier in Section 6.1, where comparison with a randomized algorithm was performed, a classification performance of 87% is considered to be useful. In other words, the C4.5 decision tree algorithm, containing risk factor data as input, is considered to be a useful classification technique in order to separate cancer patients from non-cancer patients in the data set lacking the value 'unknown' as a measure of attributes. However, the classification result gained with this approach is slightly worse than the classification result gained with the tree based on gene expression data, which achieved 89% classification accuracy. This difference is, of course, small enough to possibly be a chance effect. However, some possible problems with risk factor data, when used as input to the C4.5 algorithm, are shortly discussed in Chapter 7.

Additionally, as can be seen in column 6 in Table 5, the tendency of the C4.5 algorithm to misclassify cancer patients, based on risk factor data lacking the value 'unknown', is lower than the tendency to misclassify non-cancer patients.

As can further be seen in Table 5, the distribution of cancer and non-cancer patients differs in each of the test sets during cross validation. Looking at the classification results in Table 5, it is difficult to see how this unequal distribution of patients may have affected these results. However, when further analyzing the classification results in Table 5, a small tendency to misclassify patients might be found when the test sets contains more cancer patients compared with non-cancer patients. In other words, the error percentages gained during the classification of test sets, shown in column 7 in Table 5, tend to be a bit higher when the test sets contain more cancer patients than non-cancer patients. Thus, compared with the results shown in Table 4, the results in Table 5 indicate the exact opposite about the effect of the unequal distribution of patients on the classification performance of C4.5.

6.3 Expression of breast-related genes and risk factors as input data

In this experiment, the gene expression data and risk factor data were put together as one input set. As mentioned in Section 5.2.3, the first input data contained both gene expression data and risk factor data about all 75 samples and thus contained the value ‘unknown’ as attribute-value. This input resulted in a decision tree identical with the decision tree shown in Figure 11. This result was however rather expected since the gene expression data is more complete compared with the risk factor data, containing the value ‘unknown’ as an attribute value. This result also further emphasizes the relevance of the four genes identified in Figure 11, that is, MKI67, BAX, SNCG, and AR, for the classification of the samples.

As further described in Section 5.2.3, merging the gene expression data and risk factor data, lacking the value ‘unknown’ as attribute-value, resulted in a different decision tree which is shown in Figure 16. The three derived production rules are also shown in Section 5.2.3.

According to the first production rule, the presence of MMP9 and the absence of the REV1-like gene are important factors for the classification of breast cancer patients in the data set. The enzyme Matrix metalloproteinase-9 (MMP-9) plays important roles in tumor invasion and angiogenesis. Secretion of MMP-9 has been reported in various cancer types including breast cancer (Yao et. al. 2001). This may thus explain why the presence of MMP-9 expression is considered to contribute with the classification of cancer patients in the data set. In a study made by Scorilas et. al. (2001), the results suggest that MMP-9 may be an independent favorable prognostic factor in node-negative breast cancer patients. Node-negative means that the biopsied lymph nodes are free of cancer and the cancer has thus not metastasized. According to the study of Scorilas et. al, the overexpression of MMP-9 in breast cancer may therefore be used as a marker to subdivide node-negative breast cancer patients in order to determine the optimal treatment modality. The expression of MMP-9 is also identified by the second derived production rule shown in Section 5.2.3. According to the second production rule the absence of MMP9 contributes to the classification of the non-cancer patients in the data set.

The REV1-like gene mentioned in the first rule, is the human homolog of the S. cerevisiae mutagenesis protein REV1. Consistent with its role as a fundamental mutagenic protein, the REV1 gene is ubiquitously expressed in various human tissues (Wang et. al. 1999).

Finally, the last production rule derive from the decision tree in Figure 16 again indicates the age-dependency of breast cancer.

An interesting fact is that the p53 gene has not been included in any of the derived production rules, even though this gene is present in the decision tree shown in Figure 16. In other words, the expression of p53 has not been a significant factor for the classification of samples.

The results from the 3-fold cross validation that was performed, on the data set containing the 39 samples, are shown in Table 6. As shown in Section 5.2.3, the mean value of the error percentages, for the three test sets, indicate that the capacity of the decision tree approach in correctly classifying 'new' samples is approximately 87%. This classification performance is considered to be a useful classification result, compared with the performance of the randomized algorithm described in Section 6.1. The C4.5 decision tree algorithm, containing both risk factor data and gene expression data as input, is thus considered to be a useful classification technique in order to separate cancer patients from non-cancer patients in the data set lacking the value 'unknown' as a measure of attributes. Similar to the classification performance presented in Section 5.2.2, the classification capacity of the decision tree approach based on both gene expression data and risk factor data is slightly worse than the approach based on only gene expression data, which achieved 89% classification accuracy. As mentioned earlier, this difference is however small enough to only be a chance effect.

As described in Section 6.1, column 6 in Table 6 shows the number of misclassified cancer and non-cancer patients in each of the test sets during cross validation. When further analyzing the results in Table 6, it is obvious that the distribution of cancer and non-cancer patients differs in each of the test sets. According to Table 6, the tendency of the C4.5 algorithm to misclassify patients, based on both gene expression data and risk factor data, is a bit lower when the test sets contain more cancer patients than non-cancer patients.

6.4 Indication of pathways involving known breast cancer genes

Table 7 summarizes all signalling pathways suggested by production rules in Section 5.2.4, for the four studied genes MKI67, BAX, AR, and SNCG. In this Table, the symbol x is used in order to mark the genes that are suggested, by the C4.5 algorithm, to be involved in the same signallig pathways as the four studied genes.

As can be seen, Table 7 does not show the relation between any probes of the four studied genes MKI67, BAX, AR, and SNCG. In other words, even though some of the probes for the four studied genes are involved in some of the derived production rules shown in

Section 5.2.4, these relations are not shown in Table 7. This is because, as described in Section 5.2.4, different probes of a gene represent sub-sequences of the *same* gene. This means that even if other probes of the gene, used as decision variable, are included in the derived production rules, these probes do not contribute with any new information about the signalling pathway.

MKI67				
BAX				
AR				
SNCG		x		
BCAS1	x			
RAD51	x			x
TSG101	x			
MYC	x			
HRAS		x		
BRAK		x		
MMP9		x		
NRAS		x		
PHB		x		
P53		x		
SDBCAG84			x	
AGR2			x	
UP			x	x
LIG4			x	
NCOA3				x
FGFR1				x
IL24				x
TFF1				x
CDH13				x
XLKD1				x
	MKI67	**BAX**	**AR**	**SNCG**

Table 7: All gene-relations identified by the C4.5 algorithm. These gene-relations are suggested by the derived production rules shown in Section 5.2.4. The symbol x is used to mark the genes that are suggested to be involved in the same signalling pathways as MKI67, BAX, AR, or SNCG. For instance, the last column marks the genes suggested to be involved in the same signalling pathways as SNCG. As can for example be seen, the symbol x is used in this column to mark the UP gene. This means that there is a production rule in Section 5.2.4, which has the expression of UP as a condition on the left-hand side whereas the expression of a probe of SNCG is used for the classification, on the right-hand side of the rule.

It is very difficult to evaluate the signalling pathways presented in Section 5.2.4, since not all pathways involving the four putative breast cancer genes, MKI67, BAX, SNCG, and AR, are yet known. It is however possible to speculate around the reliability of the genes and pathways that are identified by the C4.5 algorithm.

6.4.1 The MKI67 nuclear antigen

According to the both production rules identified by the first probe of the MKI67 gene, shown in Section 5.2.4.1, RAD51 may be involved in the signalling pathway of MKI67. Tsuzuki et al. (1996) concluded that the RAD51 protein may have roles in recombination and double-strand break repair and thus plays an essential role in the proliferation of cells. Elevated levels of RAD51 recombination protein has been detected in tumor cells (Raderschall et. al. 2002). The MKI67 antigen is also expressed in proliferating cells and has not been detected quiescent cells (Schluter et. al. 1993). This could be a reason to why MKI67 and RAD51 are considered, by the production rule, to be involved in the same signalling pathway since they both have a role in proliferating cells. Another gene identified by the first probe of MKI67 is a gene called BCAS1. Although not consistently expressed, BCAS1 (Breast carcinoma amplified sequence 1) is a candidate oncogene. This gene is overexpressed in most breast cancer cell lines. According to the first production rule, shown in Section 5.2.4.1, the absence of this gene is somehow considered to be associated with the absence of MKI67.

The production rules gained by the second probe of MKI67 contain two other genes that may be involved in the signalling pathways of MKI67. These new genes include TSG101 and MYC. However, as can be seen in Section 5.2.4.1, the expression of MKI67 gained with the first probe of MKI67 has also been identified in both gained rules. This is of course because the two probes are very much correlated, since they represent the same gene. The tumor susceptibility gene, TSG101, identified in the first rule, appears to be important for maintenance of genomic stability and cell cycle regulation. The protein encoded by this gene may play a role in cell growth and differentiation and acts as a negative growth regulator (LocusLink, 2002). In other words, when the expression of TSG101 is present, the gene may act as a negative growth regulator and therefore the cell proliferation will be reduced. This could be a reason to why the expression of MKI67, which is only expressed in proliferating cells, is considered to be absent whenever the expression of TSG101 is present. According to the second production rule, gained by the second probe of MKI67, the expression of the transcription factor MYC is often present whenever the expression of MKI67 is present. Induction of the transcription factor MYC promotes cell proliferation and transformation by activating growth-promoting genes. An oncogenic c-Myc promotes cell growth and cancer development partly by inhibiting the growth inhibitory functions of Smad2 and Smad3 (Feng et. al. 2002). As mentioned earlier, the MKI67 gene is only present in proliferating cells and

since MYC promotes cell proliferation, it is only logical that the expression of MKI67 is present whenever the expression of the transcription factor MYC is present.

6.4.2 The BCL2-associated X protein (BAX)

The protein produced by the BAX gene is a key regulator of apoptosis, that is, programmed cell death. This pro-apoptotic gene is normally expressed in breast tissue, but becomes inactive in approximately one-third of invasive breast cancers (Reed, 1996). In this way the breast cancer cells bypass the process of apoptosis and therefore can grow without control. The production rules created with the first probe of BAX are shown in Section 5.2.4.2. As described in Section 6.4.1, some of the gained production rules may contain other probes of the gene that is used as the decision variable. In the rules gained with the first probe of BAX, the expression of BAX gained with other probes of the gene is shown in the first and the second rule.

According to the first production rule shown in Section 5.2.4.2, the BAX gene may be expressed whenever HRAS is expressed. The RAS proteins are thought to be oncogenic. Initially it was shown that RAS activity can drive cells into proliferation, but RAS can also block apoptosis (Adjei, 2001). The data gained in a study made by Schöndorf et. al. (2001), suggested that HRAS expression effects the particular breast tumor cells by induction of apoptosis in breast cancer patients that are in an early stage. The study indicates a possible mechanism, by which HRAS may act protectively in breast cancer patients. The role of HRAS as an inducer of apoptosis explains the first production rule shown in Section 5.2.4.2, since the BAX gene has a major role in the process of programmed cell death.

The BRAK gene is a novel divergent chemokine. Chemokines are a family of related proteins that regulate leukocyte infiltration into inflamed tissue and play important roles in many disease processes (Hromas et. al. 1999). The expression of this gene seems to be highly associated with the expression of the BAX gene, as indicated in both the second and the third production rules gained by the first probe of the BAX gene, shown in Section 5.2.4.2.

According to the production rule created with the second probe of the BAX gene, a gene called MMP9 may be involved in the signalling pathway of BAX. As described in Section 5.2.3, the enzyme, Matrix metalloproteinase-9 (MMP-9), plays important roles in tumor invasion and angiogenesis and the secretion of this enzyme has been reported in various cancer types including breast cancer (Yao et. al. 2001).

As shown in Section 5.2.4.2, the third probe of the BAX gene has identified two other genes that may be involved in the signalling pathway BAX. These genes include PHB and NRAS. Prohibitin (PHB) is an evolutionarily conserved gene that is ubiquitously expressed. It is thought to be a negative regulator of cell proliferation and may be a tumor suppressor (Sato et. al. 1992). NRAS on the other hand, is a member of the mammalian RAS gene family. Human cells contain four very similar RAS proteins, H-RAS, N-RAS, K-RAS 4A and K-RAS 4B, but it is currently unknown whether each or only one of these RAS proteins contributes to breast cancer cell growth (Jackson, 2001). Suppose that the NRAS gene functions in the same way as the earlier described HRAS gene. This would mean that NRAS would function as an apoptosis inducer and therefore activate the pro-apoptotic BAX gene, as it is shown in the production rule derived by the third probe of BAX. It seems also logical that the expression of tumor suppressor genes, like PHB, would activate the BAX gene in order to reduce the uncontrolled cell growth in breast tumors. This statement is also indicated in the rule presented by the third probe of the BAX gene.

According to the production rule derived by the last probe of the BAX gene, two other genes may be involved in the signalling pathway of BAX. These genes include SNCG and p53.

The p53 gene is a tumor suppressor that limits cellular proliferation by inducing cell cycle arrest and apoptosis in response to cellular stresses such as DNA damage and oncogene activation. One of the mechanisms by which p53 induces mitochondria-mediated cell death events is by activating genes that are directly involved in the initiation of mitochondria-induced apoptosis. Among these is the pre-apoptotic BAX gene. This could be a reason to why the two genes, BAX, and p53, may be involved in the same signalling pathways. The synuclein-gamma (SNCG) gene has been shown to be overexpressed in breast carcinomas. The normal function of the synuclein gene family is unknown (Clayton & George, 1998). Details of the properties of any member of the synuclein family may therefore provide useful information for understanding the characteristics and function of other family members. According to the production rule derived by the last probe of BAX, the expression of the BAX gene may be absent whenever the expression of the synuclein-gamma gene is absent.

6.4.3 The Androgen receptor (AR)

The androgen receptor is a member of the family of steroid receptors. This receptor binds androgens and acts as a transcription factor. Breast cancer in men is rare but among the risk factors that have been identified are a family history of breast cancer and evidence of androgen insufficiency (Lobaccaro et. al. 1993). A decrease in androgen action within the breast cells could thus account for the development of male breast cancer.

According to the first production rule derived by the first probe of the AR gene, the expression of AR may be absent whenever the expression of the serologically defined breast cancer antigen 84 (SDBCAG84) is absent. Unfortunately not much is known about SDBCAG84 and it is therefore very difficult to evaluate this statement. However, the expression of AR is further considered to be present whenever the expression of the gene AGR2 is present. AGR2 (anterior gradient 2 homolog) is expressed in estrogen receptor (ER)-positive breast cancer cell lines and has been found to be co-expressed with estrogen receptors in breast cancer cells. ER-negative breast cancers are less well-differentiated and more aggressive than ER-positive tumors (Thompson & Weigel, 1998). Similar to the androgen receptor, the esterogen receptor is a member of the steroid receptor family (Mata de Urquiza, 2001). This could be a reason to why both AR and AGR2, which is found to be co-expressed with estrogen receptors in breast cancer cells, are considered to be involved in the same signalling pathway. According to Thompson & Weigel (1998), the co-expression of AGR2 with esterogen receptors suggests that AGR2 may be involved in the tumor biology specific to the well-differentiated phenotype of hormonally-responsive breast cancers. Furthermore, the AGR2 gene is shown twice in the first production rule shown in Section 5.2.4.3. This additionally emphasizes the central role of AGR2 in the signalling pathways of AR.

As is further shown by the first production rule shown in Section 5.2.4.3, the expression of AR may be present whenever the expression of UP is present. According to Kanzaki et. al. (2002), the expression of UP gene product, Uridine phosphorylase (UPase), is higher in human solid tumors including breast carcinomas, compared with normal tissue. According to the authors, this finding suggests that the expression level of UP gene may be an independent prognostic marker in human breast carcinoma. However, for the moment the role of UP in the signalling pathway of AR is unclear.

According to the second production rule derived by the first probe of AR, the ATP-dependent DNA ligase IV (LIG4) may be involved in the signalling pathway of AR. LIG4 has a central role in the joining of single-strand breaks in double-stranded DNA (Sibanda et. al. 2001). The role of LIG4 in the signalling pathway of AR is yet unclear.

However, the two production rules derived by the second probe of the AR gene have not been able to identify any genes that are considered to be involved in the signalling pathway of AR. As it can be seen in Section 5.2.4.3, the second probe of AR has only been able to identify the first probe of AR mentioned earlier. The second probe of the AR gene has thus unfortunately not been able to contribute with any information about other genes involved the signalling pathway of the androgen receptor.

6.4.4 Synuclein-gamma (SNCG)

SNCG is a member of the human synuclein gene family. The synuclein-gamma gene, also known as BCSG1 (Breast cancer-specific gene 1), was recently found to be overexpressed in advanced infiltrating carcinoma of the breast (Lavedan et. al. 1998). According to the first production rule shown in Section 5.2.4.4, the expression of SNCG may be absent whenever the expression of NCOA3 is present. NCOA3 is a nuclear receptor co-activator that directly binds nuclear receptors and stimulates the transcriptional activities in hormone-dependent fashion. This gene is amplified and overexpressed in breast and ovarian cancer cell lines as well as in breast cancer biopsies, interacts with estrogen receptors in a ligand-dependent fashion, and functions to enhance estrogen-dependent transcription (Anzick et. al. 1997). A reason to why SNCG and NCOA3 are considered to be in the same signalling pathway could thus be the fact that both genes are overexpressed in breast cancer cells.

The expression of SNCG may be absent whenever the expression of RAD51 is absent. This relation is also shown in the first production rule derived by the first probe of SNCG. The protein encoded by the RAD51 gene is a member of the RAD51 protein family. RAD51 family members are highly similar to bacterial RecA and S. cerevisiae Rad51, and are known to be involved in the homologous recombination and repair of DNA (Yang et. al. 2001). The BRCA1 and BRCA2 proteins, implicated in familial breast cancer, form a complex with RAD51, and these genes are thought to participate in a common DNA damage response pathway (Kato et. al. 2000). However, the exact role of RAD51 in the signalling pathway of SNCG is yet unclear.

The first production rule in Section 5.2.4.4 further indicates that the expression of SNCG may be absent whenever the expression of FGFR1 is present. The protein encoded by FGFR1 is a member of the fibroblast growth factor receptor family. The extracellular portion of the protein interacts with fibroblast growth factors, setting in motion a cascade of downstream signals, ultimately influencing mitogenesis and differentiation (LocusLink,

2002). About 25-30% of breast tumors have more receptor protein on the cells than is normal (Schnitt, 2001). Since SNCG also is overexpressed in breast cancer cells, this could be a reason to why FGFR1 is considered to be in the same signalling pathway as SNCG.

The expression of SNCG may be absent when the expression of Interleukin 24 (IL24) is absent. This relation is further shown in the first production rule in Section 5.2.4.4. IL24 is up-regulated in melanoma cells. This gene induces selective anticancer properties in breast carcinoma cells by promoting p53 independent apoptosis (Ellerhorst et. al. 2002). The SNCG gene is presumably also involved in the process of apoptosis (Marzieh Jönsson, personal communication.). This would explain why SNCG and IL24 are considered, by the production rule, to be involved in the same signalling pathway.

The first production rule gained by the first probe of SNCG has further identified a gene, TFF1, which may be involved in the same signalling pathway as SNCG. The expression of SNCG is considered, by the production rule, to be absent whenever the expression of TFF1 is absent. TFF1 is a gene expressed only in human breast cancer. This gene is regulated by estrogen in MCF-7 human breast tumor cells (Moisan et. al. 1985). As mentioned earlier, SNCG may also be expressed in breast cancer cells. This may explain why the TFF1 gene is considered to be in the same signalling pathway as SNCG.

As is shown in Section 5.2.4.4, the production rule derived by the second probe of SNCG identifies to other genes that may be involved in the signalling pathway of the SNCG gene. According to the derived production rule, the expression of SNCG may be present whenever the expression of the genes CDH13 and XLKD1 is present. Cadherin 13 (CDH13) is a member of the cadherin superfamily. This gene may mediate cell-cell interactions and it may act as a tumor suppressor in breast cancer (LocusLink, 2002). CDH13 is considered to be a cadherin with growth inhibitory functions and diminished expression in human breast cancer (Lee, 1996). The gene XLKD1 (extracellular link domain-containing 1) on the other hand is a lymph-specific receptor for the glycosaminoglycan hyaluronan. The extracellular matrix glycosaminoglycan hyaluronan is an abundant component of skin and mesenchymal tissues where it facilitates cell migration during wound healing, inflammation, and embryonic morphogenesis. XLKD1 is thus considered to be a novel, specific lymphatic marker in breast cancer tissue (Cunnick et. al. in 2001). However, the reason to why these two genes, CDH13 and XLKD1, are considered to be involved in the same signalling pathway as SNCG is yet unclear.

According to the rule derived with the third probe of the SNCG gene, the expression of SNCG may be absent whenever the expression of UP is present. As mentioned earlier in

Section 5.2.4.3, the expression level of UP gene may be an independent prognostic marker in human breast cancer since the expression of this gene has been found to be higher in human solid tumors, including breast carcinomas, when compared with normal tissue (Kanzaki et. al. 2002). A reason to why UP and SNCG are considered to be in the same signalling pathway could be that both genes have a higher expression in breast tumors. The exact signalling pathway involving these genes is however yet unknown.

7. Discussion

According to Zweiger (1999), a key aspect of the genomics revolution is the transformation of large amounts of biological information into an electronic format, leading to an information-based approach to biomedical problems. Microarrays, in particular, provide huge amounts of gene-expression data in order to help our understanding of the molecular basis of health and disease. With an enormous amount of data stored in databases and data warehouses, it is increasingly important to use powerful tools for analysis of such data. Data mining approaches, searching for valuable information in volumes of data, may be applied at several stages in the drug-development process and could ultimately have broad applications in disease diagnosis and patient prognosis. According to the results gained in this project, decision tree approaches are considered to be a useful technique for the mining of biological data. The greatest advantage of decision tree models is that the results are very easy to interpret compared with results gained from other data mining techniques, like neural networks.

However, as mentioned in Chapter 3, the hypothesis defined in this project was that decision trees are a useful classification technique in order to extract valuable information about the main genes and risk factors involved in breast cancer. The results gained during the project, with the C4.5 algorithm, are considered to support the hypothesis. In other words, the C4.5 decision tree algorithm is considered to be a useful data mining technique in order to extract valuable information, regarding breast cancer, from the Gene Logic database. However, a question that might come to mind is if the results gained in this project also can be gained with other data sets, found in other databases. Even though it is quite difficult to answer this question, there is no reason to believe otherwise. The same decision tree approaches used in this project can probably produce informative decision trees and production rules also when other data sets are used as inputs. Does this mean that the decision tree algorithm is the best approach for the classification of samples in a given data set? Other algorithms may be just as good in classifying samples but it is unlikely that we would find other classification algorithms where the gained results are as comprehensible as the results gained with the decision tree approach. Even though better classification approaches may exist, the classification improvements can only be marginal. This is shown by the results gained in this project where the capacity of the different decision tree approaches, in correctly classifying ‘new’ samples, has been calculated. As can be seen in Chapter 5, the decision tree approaches used in this project have the capacity of correctly classifying samples approximately 87-89%

of the time, depending on the kind of input data used. The decision tree model is however a useful classification technique mainly when *genetic diseases* are being studied. The reason for this is that, in non-genetic diseases there are often many other factors involved and the relations between these factors are often also more complex. It may thus be difficult to find any regularities in the data. The decision tree model may therefore be an insufficient classification tool in studies involving non-genetic diseases, since the resulting decision trees and production rules may become too complex to interpret.

Looking further at the results gained in this project, there are some facts that have to be considered when the decision trees and production rules, based on risk factor data, are being evaluated. As mentioned in Section 5.2.2, the risk factors that were used in this project included a patient's age, BMI value, alcohol status and smoking status. It was also mentioned that the values for smoking and alcohol status were set to one of the following: 'current use', 'no current use', 'never used', 'previous use', 'occasional', or 'unknown'. The problem is however the difficulty of drawing a clear line between some of these possible attribute-values in the Gene Logic database. For instance, what is the difference between 'previous use' and 'no current use'? How should one interpret the value 'occasional', when it is used as a way of describing a patient's alcohol or smoking status? These kinds of questions make it very difficult to interpret the resulting decision trees and production rules that were created when risk factor data was used as input. It is also very common for patients not to reveal the truth about their alcohol and smoking habits when they are asked by a doctor. This can be due to several different reasons. For example, the patient might feel embarrassed about his or her unhealthy habits and might therefore not be completely honest when answering these kinds of questions. This is probably the most common problem in studies where different risk factors are analyzed in order to find the most relevant risk factors related to a certain disease. In order to reduce this problem it is important to define clear, non-overlapping, attribute-values when studying risk factors.

However, problems can also arise when the data is being stored in the information database. As can be seen in Figure 10, the data set contains a 59 years old patient having a BMI value of 9. This is, of course, extremely unlikely and is thus probably due to a typing error or other similar event. It is therefore important to double-check the information before putting it into the database, in order to avoid these kinds of problems. The incorrect BMI value may have affected the significance of the BMI attribute as a risk factor of breast cancer. In other words, it may have caused the BMI attribute to be placed further down in the created decision trees and thus be treated as a less significant attribute for the classification of samples. Looking

further at the resulting decision trees and production rules based on risk factor data, shown in Section 5.2.2, it is quite difficult to evaluate the significance of the BMI value for the classification of samples. As described in Section 6.2, this may be due the fact that the patients in the input data have not been separated according to their menopausal status. However, this difficulty of interpreting the significance of the BMI value could also be due to the incorrect BMI value, although this is quite unlikely since there is only one BMI value that differs in this way.

As is described in Section 5.2.4, the C4.5 decision tree approach was also tested for the indication of signalling pathways. As mentioned, the input data contained *abscall* values for several probes for each of the four genes that were analyzed. The different probes of the same gene should be very much correlated since they represent the same gene. It is therefore very natural for the C4.5 algorithm to find correlations between the different probes of a gene. A not so powerful algorithm would probably only be able to find these obvious correlations since they are the most natural and therefore easier to find. However, as can be seen in Section 5.2.4, the C4.5 algorithm has been able to identify several other genes which may be correlated with the four studied genes. The C4.5 algorithm may therefore have the capacity to look beyond the most obvious correlations when looking for regularities and patterns in data. This is however an uncertain implication since the correlation between different probes of a gene may also have been weak. In order to strengthen the above implication, further examinations must be performed where the correlation between the different probes are analyzed. However, in a perfect experimental setup the input data would not have contained several probes for each gene. Instead, each gene would have been represented by only one probe. The data used as input to the C4.5 algorithm may then have resulted in improved decision trees and production rules.

The relations and associations described by the production rules, shown in Section 5.2.4, are quite difficult to evaluate. This is because not all signalling pathways involving the four breast-related genes MKI67, BAX, AR, and SNCG are yet known. Several articles have however been found in the literature, which indicate some of the pathways suggested by these production rules, derived by the C4.5 algorithm. These findings indicate that the C4.5 algorithm has the capacity of suggesting signalling pathways by using the *abscall* values found in the Gene Logic database. However, there are also some production rules, shown in Section 5.2.4, that have identified yet unknown putative relations. These relations could very well be correct since the C4.5 algorithm earlier has shown to exhibit the capacity of identifying other known relations between genes in Gene Logic. In other words, the C4.5

algorithm may be used as a tool for the indication of possible signalling pathways, even though the results must be further examined and evaluated in order to increase the credibility of the suggested pathways. Another way to increase the accuracy of the identified production rules, shown in Section 5.2.4, is to compare the rules with rules gained when larger data sets are used, containing more than 75 samples. If the gained production rules still identify the same relations and pathways shown previously with the data set containing the 75 samples, the reliability of the identified relations and pathways would be further increased. It is however important to remember that this is an explorative study, where the C4.5 algorithm has been used in order to find agreements between the resulting decision trees and the existing breast cancer literature. The application of decision trees, on especially gene expression data, is thus a novel research approach and it is therefore beyond this study to make a thorough evaluation, even though some contradictions may exist in the results gained during this study.

As mentioned above, decision tree approaches have been used for quite a long time now but so far the approaches have only been used on other types of data than gene expression data. However, the results gained in this project, with the C4.5 algorithm, indicate that decision tree approaches also can be useful when gene expression data is used as input. It is therefore important for experts in computational science to continue developing the different decision tree algorithms. In this way it might be possible to further improve the performance of these algorithms and consequently improve their capacity of interpreting the huge amounts of gene expression data.

However, one matter that might be questioned in this project is if the concept of ‘data mining’ really has been applied on the Gene Logic database and thus if the C4.5 algorithm actually is a useful data mining technique. As described in Section 2.1, the term data mining refers to using a variety of techniques in order to identify information or decision-making knowledge in bodies of data. In this project however, the data in Gene Logic has been filtered and criterions have been used in order to reduce the amount of input data to the C4.5 algorithm. The concept of ‘data mining’ is usually used without these kinds of pre-processing done during this project. However, the C4.5 algorithm is considered to be a useful classification technique even without any filtering of data. This was revealed in a small test, which was done in order to examine the capacity of the C4.5 algorithm in creating decision trees based on randomly chosen genes as input. The expression of 100 randomly chosen genes from Gene Logic was therefore used as input to the C4.5 algorithm. The pruned version of the resulting decision tree consisted of 6 genes where the putative breast cancer gene MKI67 was included. In fact, MKI67 was chosen as the root node in the resulting decision tree. As can be

seen in Figure 11, MKI67 was also chosen as the root node in the first performed approach where the expression of the 108 breast-related genes was used as input. These results indicate that the C4.5 algorithm might have the ability to identify MKI67 even though the input data contains the expression of randomly chosen genes. The C4.5 algorithm is therefore considered to be a useful data mining technique when gene expression data from the Gene Logic database is used.

As mentioned both in Chapter 5 and 6, cross validation was performed in order to evaluate the classification performance of the four decision tree approaches performed in this study. As also mentioned earlier, the test sets used during the cross validation procedures differed with respect to the distribution of cancer and non-cancer patients. This unequal distribution is however not believed to have an effect on the classification performance of the C4.5 algorithm. This is because, when the gained classification results are analyzed in Chapter 6, it is difficult to find any obvious tendency showing that the C4.5 classification performance has been affected by the unequal distribution of patients. In other words, the existing difference between the classification result of each data set, shown in Tables 2, 4, 5, and 6, might only be due to a chance effect.

8. Conclusions

This research was performed in order to test the feasibility of the application of decision trees, created by the C4.5 algorithm, on the kind of data found in the Gene Logic database (see Section 2.4). Different kinds of data sets were therefore used as input in order to analyze the different aspects of the C4.5 application. Cross validation was performed in order to evaluate the performance of the different decision tree approaches. The results from the cross validation were then compared with results gained from a randomized classification algorithm (see Section 6.1). This was done in order to find out if the classification performance of the C4.5 algorithm is better compared with the classification performance of the randomized algorithm.

The first aspect of the C4.5 algorithm was analyzed when gene expression data, described as *abscall* values explained in Section 2.4, was used in order to separate cancer patients from non-cancer patients in a data set containing 75 patients. In other words, the capacity of the C4.5 algorithm in correctly classifying 75 patients, considering only the expression profiles of 108 earlier chosen breast-related genes, shown in appendix A, was examined. According to the decision tree that was created, the expression of four genes out of the 108 genes tested, were considered to be sufficient for the separation of samples in the data set. These four genes include MKI67, BAX, AR, and SNCG, shown in Figure 11. The resulting production rules are also shown in Section 5.2.1. The performance of the decision tree approach was evaluated through cross validation. The cross validation procedure shows 89% accuracy for the resulting tree to correctly classify ‘new’ samples. Compared with the earlier mentioned randomized algorithm, this result indicates that the C4.5 algorithm is a useful classification technique when considering gene expression data as input to the algorithm.

Another aspect of the application of the C4.5 algorithm was analyzed when risk factor information for the 75 samples was used as input data. This was done in order to identify the most significant risk factors necessary for the separation of cancer patients and the non-cancer patients in the data set. The risk factors that were tested included a patient’s age, BMI value, alcohol status, and smoking status. Since some patients in the data set had the value ‘unknown’ for either the alcohol or the smoking status, the resulting decision trees and production rules were quite difficult to interpret (see Section 6.2 and Figure 12 and 13 in Section 5.2.2). In an attempt to simplify this interpretation and to reduce the disorder in the input data, the patients having the value ‘unknown’ as an attribute value were removed from

the data set. The resulting pruned decision tree is shown in Figure 14 whereas the un-pruned version is shown in Figure 15. The production rules created are also shown in Section 5.2.2. The cross validation performed on the last mentioned input data, lacking the attribute value 'unknown', shows that the accuracy of the resulting decision tree in classifying 'new' samples is approximately 87%. In the same way as the results gained during the earlier mentioned tree based on gene expression data, this result indicates that the C4.5 algorithm is a useful classification technique when considering risk factor data as input. An interesting result is also that all the decision trees based on risk factor data emphasize the fact that a person's age is the most important and informative attribute in order to correctly classify the samples in the data set. However, the classification performance of this decision tree approach is slightly worse compared with the classifications performance of the decision tree approach based on gene expression data, which achieved 89% classification accuracy. This difference is, of course, small enough to possibly be a chance effect. However, possible problems with risk factor data, when used as input to the C4.5 algorithm, are discussed in Chapter 7.

In the third run of the C4.5 algorithm, both gene expression data and risk factor data were used as input. The resulting decision tree and the derived production rules are shown in Section 5.2.3. According to the resulting tree, shown in Figure 16, the expression of the three genes MMP9, REV1L, and p53, and also the age of a patient, are the most relevant attributes in order to classify the 75 samples correctly. The cross validation performed on the input data showed that this decision tree approach had the capacity to correctly classify samples approximately 87% of the time. The classification performance is thus as good as the classification performance of the tree based on only risk factor data. In other words, this classification performance is a bit worse compared with the classification performance of the decision tree approach based on only gene expression data, although the difference is small enough to be a chance effect. However, this result again shows the classification capacity of the C4.5 algorithm on both gene expression data and risk factor data. The C4.5 algorithm is thus considered to be a useful classification technique considering the gene expression data and risk factor data found in the Gene Logic database.

Finally, the performance of the C4.5 algorithm was applied for the identification of genes involved in the same signalling pathways as the four genes identified in the decision tree based on only gene expression data, that is, MKI67, BAX, AR, and SNCG, shown in Figure 11. However, the gene expression data used as input contained *abscall* values for at least two different probes for each of these four genes. Therefore, a decision tree and a set of production rules were created for each of the probes. Instead of illustrating each of the

resulting decision trees and corresponding rules, only some of the production rules created for each of the four genes are presented in Section 5.2.4. The presented production rules are regarded as the most relevant rules by considering how many of the 75 samples that were correctly classified by each rule. These most relevant production rules, indicating genes involved in the same signalling pathways as MKI67, BAX, AR, and SNCG, are thus shown in Section 5.2.4.

Considering the overall results gained during this project, the C4.5 algorithm is regarded as a useful classification algorithm for classification of data found in the Gene Logic database. The results shown in Sections 5.2.1, 5.2.2, and 5.2.3 indicate that the C4.5 algorithm has the capacity to identify attributes that are relevant for the separation of non-cancer patients and the patients suffering from breast cancer in the data sets used. C4.5 is thus considered to be a useful tool for classifying the Gene Logic samples correctly.

The production rules gained during the indication of pathways for the four genes MKI67, BAX, AR, and SNCG are shown in Section 5.2.4. These production rules are however more difficult to evaluate since all pathways involving these four putative breast cancer genes are not yet known. Several articles have however been found to indicate some of the pathways suggested by the C4.5 algorithm. Some speculation regarding the reliability of the other identified pathways are presented in Chapter 6.

9. Future work

As mentioned earlier, it is really worth testing the decision tree approach on other important genetic diseases, like for example Hemophilia or the Sickle-cell disease. It would also be interesting to test the C4.5 decision tree approach on other cancer forms than breast cancer. Another interesting idea would be to use the decision tree approach in order to study the differences in gene expression between the different stages of breast cancer.

Similar to this study, very promising results have been found in a study made by Khan et. al. (2001), where artificial neural networks were used in order to classify cancers to specific diagnostic categories based on their gene expression signatures. The ANNs were trained using small, round blue-cell tumors as a model. According to the authors, the ANNs correctly classified all samples and identified the genes most relevant to the classification. As a future work, it would therefore be very interesting to use the C4.5 decision tree approach on the same data set used by Khan et. al (2001). This would make it possible to get an indication of which of the two approaches, that is, artificial neural networks and decision trees, that gives the best classification results. The two approaches may also give the same classification result. Another interesting point would be to study if the resulting decision trees, gained by the decision tree approach, somehow would be more informative and comprehensible compared with the results gained with the ANNs.

However, future trends in healthcare delivery, in particular, when focusing on preventive medicine, will further require the integration of molecular medicine with computer technology. In this way, molecular medicine will eventually shift from costly intervention and treatment of established diseases to prediction and prevention of disease risks. This approach however requires newer systems that will link large scale biological databases with special programs for data mining (Nakamura, 1999).

References

Adjei, A. A (2001), Ras signaling pathway proteins as therapeutic targets, *Current Pharmaceutical Design*, 7, (16), p 1581-1594

American Cancer Society,
URL: http://www.cancer.org/
(Acc 020301)

Anzick, S. L., Kononen, J., Walker, R. L., Azorsa, D.O., Tanner, M. M., Guan, X. Y., Sauter, G., Kallioniemi, O. P., Trent, J. M. & Meltzer, P. S. (1997), AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer, *Science*, 277, (5328), p 965-968

Apte, C. & Weiss, S. (1997), Data mining with decision trees and decision rules, *Future generation computer systems*, 13, p 197-210

Bohanec, M., Zupan, B. & Rajkovic, V. (2000), Applications of qualitative multi-attribute decision models in health care, *International journal of medical informatics*, 58-59, p 191-205

Bristol-Myers Squibb,
URL: http://www.bms.se/cancer/brost_kvin.htm#inledning
(Acc 020305)

Bränden, H. (1997), Molekylär biologi, Studentlitteratur, Lund

Clayton, D. F. & George, J. M. (1998), The synucleins: a family of proteins involved in synaptic function, plasticity, neurodegeneration and disease, *Trends Neuroscience*, 21, (6), p 249-54

Cunnick, G. H., Jiang, W. G., Gomez, K. F. & Mansel, R. E. (2001), Lymphangiogenesis quantification using quantitative PCR and breast cancer as a model, *Biochemical and biophysical research communications*, 288, (4), p 1043-1046

Dairkee, S. H & Smith, H. S (1996), Genetic Analysis of breast cancer progression, *Journal of mammary gland biology and neoplasia*, 1, (2), p 139-151

Ellerhorst, J. A., Prieto, V. G., Ekmekcioglu, S., Broemeling, L., Yekell, S., Chada, S. & Grimm, E. A. (2002), Loss of MDA-7 expression with progression of melanoma, *Journal of clinical oncology: official journal of the American Society of Clinical Oncology,* 20, (4), p 1069-1074

Feng, X. H., Liang, Y. Y., Liang, M., Zhai, W. & Lin, X. (2002), Direct interaction of c-Myc with Smad2 and Smad3 to inhibit TGF-beta-mediated induction of the CDK inhibitor p15 (Ink4B), *Molecular Cell*, 9, (1), p 133-143

Feunteun, J. (1998), Breast cancer and genetic instability: the molecules behind the scene, *Molecular medicine today*, p 263-267

Frawley, W. J., Piatetsky-Shapiro, G. & Matheus, C.J. (1992), Knowledge discovery in databases: An overview, *AI Magazine*, 13, (3), p 213-228

Han, J. (1999), Data Mining, in J. Urban and P. Dasgupta (eds.), *Encyclopedia of Distributed Computing*, Kluwer Academic Publishers, p 1-7

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A. & Trent, J. (2001), Gene-expression profiles in hereditary breast cancer, *The new England journal of medicine*, 344, 539-548

Hromas, R., Broxmeyer, H. E., Kim, C., Nakshatri, H., Christopherson, K 2nd, Azam, M. & Hou, Y. H. (1999), Cloning of BRAK, a novel divergent CXC chemokine preferentially expressed in normal versus malignant cells, *Biochemical and biophysical research communications*, 255, (3), p 703-6

Hu, Y. J. (2001), An integrated approach for genome-wide gene expression analysis, *Computer methods and programs in biomedicine*, 65, p 163-174

Hynes, N. & Dickson, R. B. (1996), Molecular aspects of breast cancer, *Journal of mammary gland biology and neoplasia*, 1, p 137-138

Ingvarsson, S. (2000), Molecular biology of breast cancer, *Oncology Reports*, 7, p 1163-1170

Jackson, J. H. (2001), Specificity of Ras Signaling in Breast Cancer, *The Scripps Research Institute*,
URL: http://www.ucop.edu/srphome/bcrp/progressreport/patho.html
(Acc 020428)

Ji, H., Liu, Y. E., Jia, T., Wang, M., Liu , J. & Xiao, G. (1997), Identification of a breast cancer-specific gene, BCSG1, by direct differential cDNA sequencing, *Cancer Research*, 57, (4), 759-764

Jönsson, M. (2000), Wnt-5a signaling in human mammary cells: Implications for the development of breast cancer, Wallin & Dalholm boktryckeri AB

Kamber, M., Winstone, L., Gong, W., Cheng, S. & Han, J. (1997), Generalization and Decision Tree Induction: Efficient Classification in Data Mining, *Database Systems Research Laboratory School of Computing Science*, p 1-10

Kanzaki, A., Takebayashi, Y., Bando, H., Eliason, J. F., Watanabe, Si. S., Miyashita, H., Fukumoto, M., Toi, M. & Uchida, T. (2002), Expression of uridine and thymidine phosphorylase genes in human breast carcinoma, *International journal of cancer*, 97, (5), p 631-635

Kato, M., Yano, K., Matsuo, F., Saito, H., Katagiri, T., Kurumizaka, H., Yoshimoto, M., Kasumi, F., Akiyama, F., Sakamoto, G., Nagawa, H., Nakamura, Y. & Miki, Y. (2000), Identification of Rad51 alteration in patients with bilateral breast cancer, *Journal of human genetics*, 45, (3), p 133-137

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, , F., Schwab, M., Antonescu, C. R., Peterson, C. & Meltzer, P. S. (2001), Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature medicine* , 7, (6), p 673-679

Kononenko, I. (2001), Machine learning for medical diagnosis: history, state of the art and perspective, *Artificial Intelligence in Medicine*, 23, p 89-109

Kuo, W. J., Chang, R. F., Chen, D. R. & Lee, C. C. (2001), Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images, *Breast cancer research and treatment*, 66, p 51-57

Lake Michigan College,
URL: http://www.lmc.cc.mi.us/liberal/bio/bio111/mitosis.html
(Acc 020310)

Lavedan, C., Leroy, E., Dehejia, A., Buchholtz, S., Dutra, A., Nussbaum, R. L. & Polymeropoulos, M. H. (1998), Identification, localization and characterization of the human gamma-synuclein gene, *Human Genetics*, 103, (1), p 106-112

Lavrac, N. (1999), Selected techniques for data mining in medicine, *Artificial Intelligence in Medicine*, 16, p 3-23

Lee, S. W. (1996), H-cadherin, a novel cadherin with growth inhibitory functions and diminished expression in human breast cancer, *Nature medicine*, 2, (7), p 776-782

Lidbrink, E. (2001), Bröst cancer – prognosen god om tumören upptäcks tidigt, *Medicinskt Forum*, 7, p 2-6

Lobaccaro, J. M., Lumbroso, S., Belon, C., Galtier-Dereure, F., Bringer, J., Lesimple, T., Namer, M., Cutuli, B. F., Pujol, H. & Sultan, C. (1993), Androgen receptor gene mutation in male breast cancer, *Human molecular genetics*, 2, (11), p 1799-1802

LocusLink at the NBCI homepage,
URL: http://www.ncbi.nlm.nih.gov/LocusLink/
(Acc 0200425)

Lodish, H., Baltimore, D., Berk, A., Zipurski, S. L., Matsudaira, P. & Darnell, J. (1995), Molecular cell biology 3:rd edition, Scientific American Books, W. H. Freedman & Co.

Mata de Urquiza, A. (2001), FINDing the ligand: Retinoid receptor activation in the CNS, *Ludwig institute of cancer research and department of cellular and molecular biology*, Karolinska university press

Mendelsohn, J. & Baselga, J. (2000), The EGF receptor family as target for cancer therapy, *Oncogene*, 19, p 6550-6565

Mesko, T. W, Dunlap, J. N. & Sutherland, C. M (1990), Risk factor of breast cancer, *Comprehensive therapy*, 16, p 3-9

Moisan, J. P., Mattei, M. G., Baeteman-Volkel, M. A., Mattei, J. F., Brown, A. M. C., Garnier, J. M., Jeltsch, J. M., Masiakowsky, P., Roberts, M. & Mandel, J. L. (1985), A gene expressed in human mammary tumor cells under estrogen control (BCEI) is located in 21q223 and defines an RFLP, *Cell Genetics*, p 701-702

Mort, M. (2000), Modern Drug Discovery,
URL:http://pubs.acs.org/hotartcl/mdd/00/jan/mort.html#box2
(Acc. 020215)

Nakamura, R. M (1999), Technology that will initiate future revolutionary changes in healthcare and the clinical laboratory, *Journal of clinical laboratory analysis*, 13, (2), p 49-52

National Heart, Lung and Blood institute,
URL: http://www.nhlbisupport.com/bmi/bmicalc.htm
(Acc. 020417)

Nilsson, N.J (1996), Introduction to machine learning,
URL: http://robotics.stanford.edu/people/nilsson/mlbook.html
(Acc 020217)

Pavelic, K. & Gall-Troselj, K. (2001), Recent advances in molecular genetics of breast cancer, *Journal of molecular medicine*, 79, p 566-573

Pendharkar, P. C., Rodger, J. A., Yaverbaum, G. J., Herman, H. & Benner, M. (1999), Association, statistical, mathematical and neural approaches for mining breast cancer patterns, *Expert systems with applications,* 17, p 223-232

Pujol, P., Galtier-Dereure, F. & Bringer, J. (1997), Obesity and breast cancer, *Human Reproduction,* 12, p 116-125

Quinlan, J. R. (1993), C4.5: Programs for machine learning, Morgan Kaufmann Publishers

Quinlan, J. R (1996), Bagging, boosting, and C4.5, *Proceedings of the 13th National Conference on Artificial Intelligence,* p 725-730

Raderschall, E., Stout, K., Freier, S., Suckow, V., Schweiger, S. & Haaf, T. (2002), Elevated levels of Rad51 recombination protein in tumor cells, *Cancer research,* 62, (1), p 219-25

Ranstam, J. & Olsson, H. (1995), Alcohol, cigarette smoking, and the risk of breast cancer, *Cancer detection and prevention,* 19, (6), p 487-493

Reed, J. C. (1996), Balancing cell life and death: BAX, apoptosis, and breast cancer, *The Journal of clinical investigation,* 97, (11), p 2403-4

Sato, T., Saito, H., Swensen, J., Olifant, A., Wood, C., Danner, D., Sakamoto, T., Takita, K., Kasumi, F., Miki, Y., Skolnick M. & Nakamura, Y. (1992), The human prohibitin gene located on chromosome 17q21 is mutated in sporadic breast cancer, *Cancer Research,* 52, (6), p 1643-6

Schluter, C., Duchrow, M., Wohlenberg, C., Becker, M. H., Key, G., Flad, H. D. & Gerdes, J. (1993), The cell proliferation-associated antigen of antibody Ki-67: a very large, ubiquitous nuclear protein with numerous repeated elements, representing a new kind of cell cycle-maintaining proteins, *The Journal of cell biology,* 123, (3), p 513-22

Schneider, J. & Moore, A. W. (1997), A locally weighted learning tutorial using Vizier 1.0, URL: http://www-2.cs.cmu.edu/~schneide/tut5/tut5.html
(Acc 020425)

Schnitt, S. J (2001), Breast cancer in the 21st century: Neu opportunities and neu challenges, *Modern Pathology*, 14, p 213-218

Schondorf, T., Rutzel, S., Andrack, A., Becker, M., Hoopmann, M., Breidenbach, M. & Gohring, U. J. (2001), Immunohistochemical analysis reveals a protective effect of H-ras expression mediated via apoptosis in node-negative breast cancer patients, *International journal of oncology*, p 273-277

Scorilas, A., Karameris, A., Arnogiannaki, N., Ardavanis, A., Bassilopoulos, P., Trangas, T. & Talieri, M. (2001), Overexpression of matrix-metalloproteinase-9 in human breast cancer: a potential favourable indicator in node-negative patients, *British journal of cancer*, 84, (11), p 1488-96

Sibanda, B. L., Critchlow, S. E., Begun, J., Pei, X. Y., Jackson, S. P., Blundell, T. L. & Pellegrini, L. (2001), Crystal structure of an Xrcc4-DNA ligase IV complex, *Nature structural biology*, 8, (12), p 1015-1019

Trentham-Dietz, A., Newcomb, P. A., Storer, B. E., Longnecker, M. P., Baron, J., Greenberg, E. R. & Willett, W. C. (1997), Body size and risk of breast cancer, *American journal of Epidemiology*, 145, (11), p 1011-1019

Tsuzuki, T., Fujii, Y., Sakumi, K., Tominaga, Y., Nakao, K., Sekiguchi, M., Matsushiro, A., Yoshimura, Y. & Morita, T. (1996), Targeted disruption of the Rad51 gene leads to lethality in embryonic mice, *Proceedings of the National Academy of Sciences of the United States of America*, 13, p 6236

Yang, S., VanLoock, M. S., Yu, X. & Egelman, E. H. (2001), Comparison of bacteriophage T4 UvsX and human Rad51 filaments suggests that RecA-like polymers may have evolved independently, *Journal of molecular biology*, 312, (5), p 999-1009

Yao, J., Xiong, S., Klos, K., Nguyen, N., Grijalva, R., Li, P. & Yu, D. (2001), Multiple signaling pathways involved in activation of matrix metalloproteinase-9 (MMP-9) by heregulin-beta1 in human breast cancer cells, *Oncogene*, 20, p 8066-74

Wang, S. C. & Hung, M. C. (2001), HER2 overexpression and cancer targeting, *Seminars in oncology*, 28, p 115-124

Zupan, B., Lavrac, N. & Keravnou, E. (1998), Data mining techniques and applications in medicine, *Artificial Intelligence in Medicine*, 16, p 1-2

Zweiger, G. (1999), Knowledge discovery in gene-expression-microarray data: mining the information output of the genome, *Trends in Biotechnology*, 17, p 429-436

Appendix A

This appendix shows the name and the ID of the108 breast-related genes, from LocusLink, whose expression patterns were used in this project.

GeneName	LocusID	GeneName	LocusID	GeneName	LocusID
KLK10	5655	CDH13	1012	PRKCDBP	8990
ABCG2	9429	CHEK2	11200	RAD51	5888
AGR2	10551	CTDP1	9150	RAD54L	8438
AMPH	273	CTGF	1490	REV1L	51455
AR	367	CYR61	3491	RFC1	5981
ASC	29108	DD96	10158	RPL13	6137
ATM	472	DNTT	1791	SCYB14	9547
BAP1	8314	DSS1	7979	SDBCAG84	51614
BARD1	580	EBAG9	9166	SEL1L	6400
ARHC	389	EPSTI1	94240	SLC22A1L	5002
BAX	581	ERBB2	2064	SNCG	6623
BC-2	27243	ESR1	2099	ST7	7982
BCAA	51742	GA	27165	STAT3	6774
BCAR1	9564	H11	26353	TFF1	7031
BCAR2	9565	HSF1	3297	TP53	7157
BCAR3	8412	IL24	11009	TP53BP1	7158
BCAS1	8537	IL6	3569	TRAF4	9618
BCAS2	10286	LDOC1	23641	TSG101	7251
BCAS3	89751	LIG4	3981	TSP50	29122
BCAS4	55653	MAGED2	10916	UP	7378
BCCIP	56647	MAP2K4	6416	VHL	7428
BCPR	8142	MGC4809	91860	WISP1	8840
BIN1	274	MKI67	4288	WNT2	7472
BIN2	51411	MLN51	22794	XLKD1	10894
BPHL	670	MMP9	4318	XRCC1	7515
BRCA1	672	MRPS26	64949	PGR	5241
BRCA2	675	NBS1	4683	HRAS	3265
BRCA3	8068	NCOA3	8202	KRAS2	3845
BRCA4	60500	NY-BR-1	91074	NRAS	4893
MUC1	4582	OBTP	29964	CCND1	595
BRCAX	57345	PAXIP1L	22976	FGFR1	2260
BRCD1	8105	PES1	23481	PTEN	5728
BRCD2	7797	PHB	5245	MYC	4609
BRIP1	83990	PIP	5304	RB1	5925
BRMS1	25855	POLM	27434	MUC1	4582
CDH1	999	PPP1R14C	81706	LOC118430	118430

April 11, 2002

Neda Rahpeymai

Re: Use of ACS Image

Dear Ms. Rahpeymai:

We received your request to use the American Cancer Society's ("Society") breast artwork located on the Society's Internet web page http://www.cancer.org/eprise/main/docroot/CRI/content/CRI_2_2_1X_What_is_breast_cancer_5?sitearea=CRI.

The Society has no objections to your request. Please use the following credit line:

"Source: American Cancer Society's website www.cancer.org. Reprinted with permission."

The American Cancer Society grants Neda Rahpeymai the limited, nonexclusive right to use the above referenced illustration solely in connection with Neda Rahpeymai's thesis.

Please note that modification or alteration of such illustration is strictly prohibited. Any other use of this or other American Cancer Society copyrighted material without the express written consent of the American Cancer Society is prohibited.

Please feel free to contact me if you have any questions. I can be reached by fax at 404-327-6550 or by e-mail at asaffiot@cancer.org.

Sincerely,
Anthony Saffioti
Corporate Paralegal