Comparing NR Expression among Metabolic Syndrome Risk Factors

(HS-IDA-MD-03-205)

Annelie Jacobsson
(e99annja@student.his.se)

Department of Computer Science
University of Skövde, Box 408
S-54128 Skövde, SWEDEN

Master's dissertation, spring 2003
Study program in Bioinformatics
Supervisor: Kim Laurio
Industry supervisor: Magnus L. Andersson

Comparing NR Expression among Metabolic Syndrome Risk Factors

Submitted by Annelie Jacobsson to University of Skövde as a dissertation for the degree of M.Sc., in the Department of Computer Science.

[03-06-13]

I certify that all material in this dissertation which is not my own work has been identified and that no material is included for which a degree has previously been conferred on me.

Signed: _____

Comparing NR Expression among Metabolic Syndrome Risk Factors

Annelie Jacobsson (e99annja@student.his.se)

Abstract

The metabolic syndrome is a cluster of metabolic risk factors such as diabetes type II, dyslipidemia, hypertension, obesity, microalbuminurea and insulin resistance, which in the recent years has increased greatly in many parts of the world. In this thesis decision trees were applied to the BioExpress™ database, including both clinical data about donors and gene expression data, to investigate nuclear receptors ability to serve as markers for the metabolic syndrome. Decision trees were created and the classification performance for each individual risk factor were then analysed. The rules generated from the risk factor trees were compared in order to search for similarities and dissimilarities. The comparisons of rules were performed in pairs of risk factors, in groups of three and on all risk factors and they resulted in the discovery of a set of genes where the most interesting were the Peroxisome Proliferator – Activated Receptor - Alpha, the Peroxisome Proliferator – Activated Receptor - Gamma and the Glucocorticoid Receptor. These genes existed in pathways associated with the metabolic syndrome and in the recent scientific literature.

Keywords: Metabolic Syndrome, Nuclear Receptors, Data Mining, Decision trees, Gene expression analysis

Acknowledgements

I would like to thank my supervisor Kim Laurio at University of Skövde for providing me with motivation, regular suggestions and timely feedback, which have greatly improved this thesis. I would also like to thank my supervisor at AstraZeneca, Magnus L. Andersson for giving inspiring support and encouragement during the whole project. I must also thank Marcus Bjäreland at AstraZeneca for the original idea of this project and for giving advice and assistance especially on the data mining part of the project. I am also thankful for valuable feedback from Björn Olsson who has been my examiner during this thesis. Finally I would also like to thank my boyfriend Johan, my parents and my dear friends for giving me much support and encouragement throughout this time, without them this work would not have been possible.

Table of Contents

1 Introduction

Undernutrition and starvation have in the past been profound causes of disease and mortality in several countries of the now industrialised world (Björntorp, 1997). Diseases like tuberculosis, which result from undernutrition, were severe problems in the lesser-developed countries even until the 1930s. In contrast to this we are now facing the opposite problem - that of energy excess - in many parts of the world, especially in the industrialised parts. Recent reports on obesity, insulin resistance and other similar affections are indeed very alarming (Björntorp, 1997; Reaven, 2002; Grundy, Abate & Chandalia, 2002; Petersen & Schulman, 2002). Furthermore, there is strong evidence that this epidemic continues to increase at a considerable rate (Grundy et al., 2002) and it seems unlikely that the population will change their habits considerably. The consequences of obesity are well known and they include insulin resistance and elevation of blood pressure, with accompanying complications such as diabetes mellitus type II, dyslipidemia and cardiovascular disease (Grundy et al., 2002; Reaven, 2002).

Furthermore, the clustering of metabolic risk factors like insulin resistance, obesity, dyslipidemia, hypertension, microalbuminurea and diabetes mellitus type II can lead to even more serious complications (Hansen, 1999). This clustering has been named the Metabolic Syndrome, Syndrome X and the Insulin Resistance Syndrome. According to Meigs (2000), the metabolic syndrome can be defined as the co-occurrence of multiple metabolic and physiological risk factors such as overall and central obesity, dyslipidemia, diabetes mellitus type II, hypertension, microalbuminurea and insulin resistance, which together make a contribution to the development of cardiovascular disease. The metabolic syndrome is a worldwide epidemic, which is likely to increase greatly in the coming years (Grundy et al., 2002). Because of this it is necessary to create an awareness of this problem so that the importance of a healthy diet, regular exercise and the negative effects of smoking becomes clear for the patients suffering from metabolic syndrome (Grundy et al., 2002). However, in some cases it will also be necessary to develop possible drug treatments so that quick changes can be made, with the purpose of curing patients with severe effect of the epidemic. Because of the fact of how widespread and common the metabolic syndrome is, it can be seen as a major burden for national health economies, and the future does not seem to be any brighter (Björntorp, 1997).

Drug target discovery has traditionally started with the implication of a biochemical pathway in a pathophysiological process (Debouck & Goodfellow, 1999). The first step usually involved a survey of the enzymatic activity in the pathway and some particular enzyme was characterised and purified. Occasionally, sufficient information was known about the enzyme's structure and the mechanisms of action in the pathway, to be able to identify classes of small molecules to be targeted. Finally, medicinal chemists tried to optimise the compounds and to remove undesired properties of the enzymes. A similar process was applied to identify receptors and their use as drug targets (Debouck & Goodfellow, 1999). Advances in molecular biology and gene cloning techniques have improved the drug discovery field in several ways (Debouck & Goodfellow, 1999). Gene cloning and *in vitro* gene expression

provides human targets when access to human tissues is limited, which is important because even a single amino acid difference can make the drug ineffective against targets. Cross-hybridisation with cloned sequences can be used for identifying related targets. One limiting aspect has been target validation, which means the linking of targets to therapeutic capacity. This step in target discovery is difficult because it requires a detailed understanding of pathophysiological processes. The ideal technique for solving this problem would be to compare normal genes with disease genes in human on a whole genome scale. A potential approach for this could be using a gene expression profiling technique such as DNA microarrays or techniques within the area of proteomics (Debouck & Goodfellow, 1999). By investigating the expression patterns of genes, indirect information can be derived about the protein function. Gene expression analysis provides the opportunity to compare the expression of thousands of genes between 'disease' and 'normal' tissues and cells. The technique further allows the identification of multiple potential targets.

In recent years biological research has become more and more database-driven, which has its ground in experiments of large-scale functional genomics, proteomics and gene expression analysis (Bertone and Gerstein 2001). Gene expression array technologies have simplified the analysis of the expression levels of large sets of genes simultaneously (Tamames et al., 2002). By investigating the expression profiles of genes, microarray technology can provide a wide variety of information regarding the biology of the concerned organism. In exploring the genomic content of organisms, DNA microarrays have been used for e.g. protein function prediction for related genes, genotyping (Tamames et al., 2002) and to correlate the levels of gene expression with subcellular localisation (Bertone & Gerstein, 2001). As mentioned above, the advantages of gene expression analysis have also led to the technique being applied in the field of drug discovery, where it has been successfully used in drug target identification (Debouck & Goodfellow, 1999).

In addition to biological databases containing structural and sequential data, many diverse types of experimental data are also organised into databases (Bertone & Gerstein, 2001). The experimental data often has its primary focus on different aspects of protein function. It is however important that this data can be related to other similar biological data in order to put it in a useful context (Bertone & Gerstein, 2001). It has therefore become a great challenge in bioinformatics to integrate databases so that the information from different biological databases can be put together, which could indeed lead to the fact that large-scale studies can be conducted in order to analyse many different datasets.

Because of the advances in genome research and the large scale of the experiments, there has also been a change in methods of deriving knowledge of genes and proteins (Bertone and Gerstein 2001). Earlier one had to read most of the existing literature to be able to learn about the experimental knowledge about a protein or a gene, but nowadays one can use integrated database analysis and data mining to generate a preliminary set of hypotheses regarding this data. The biological databases contain a possible gold mine of valuable information, but the problem is to analyse the enormous amount of data and to extract meaningful patterns from it (Deogun, Raghavan, Sarkar & Sever, 1997). As databases become larger, it therefore becomes increasingly more difficult to support decision-making (Thuraisingham, 1998). In order to try to improve this decision-making, a technique referred to as data mining can be used. The aim of data mining within bioinformatics is to identify distinguishing

properties in a given dataset. Data mining refers to all techniques that use information to be able to extract knowledge about data (Witten & Frank, 1999).
In this present work we apply data mining techniques to the BioExpress™ database to search for a set of marker genes, which could be associated with metabolic syndrome in humans. BioExpress™ (Gene Logic Inc.) is a database that contains both clinical data and gene expression data from donors in the United States. The basic idea is to examine the possible link between the metabolic syndrome and nuclear receptors (NR) using the BioExpress™ database. Four different classifiers are used to classify each risk factor individually and the results from each classifier are then compared, with the purpose of trying to identify marker genes involved in the metabolic syndrome. The motivation for using NRs is first and foremost the suitability of NRs as drug targets but also the fact that an additional database, a nuclear receptor (NR) database, is accessible at the research and development site at AstraZeneca in Mölndal. The NR database contains information about known nuclear receptors and co-factors. The hypothesis is that by using four different classifiers – one for each individual risk factor – and then comparing the results from the classifications, it will be possible to identify biological marker genes involved in the metabolic syndrome. The results from the data analysis on the gene expression data for the four specified risk factors show that the generated classifiers did not have preferable high cross-validation. The comparisons of rules that were generated from the decision trees resulted in the discovery of a set of genes were the most interesting were the Peroxisome Proliferator – Activated Receptor – Alpha (PPARA), the Peroxisome Proliferator – Activated Receptor – Gamma (PPARG) and the Glucocorticoid Receptor (NR3C1). These genes were found in the assembly of rules that were generated during the comparison and in the Knowledge Bank, which is a collection of genes created by a literature study by Halinen and Norseng (2002). The Knowledge Bank includes genes that are known to be associated with metabolic syndrome risk factors. Furthermore, the three genes also exist in pathways that are associated with the metabolic syndrome and in the recent scientific literature where an association has been suggested between the genes and metabolic syndrome risk factors.

The remainder of this thesis is organised as follows. The second chapter includes background information about the metabolic syndrome and associated risk factors. The chapter also covers information about data mining, decision trees and related work. A fundamental review of NRs and their known functions is also made. The third chapter covers a presentation of the problem, including the problem definition, hypothesis and aims and objectives. The fourth chapter includes a detailed explanation of the steps included in the method used in this thesis. In the fifth chapter the results generated from the experiments performed in the study are presented. Finally, chapter 6 covers a discussion on the results and chapter 7 includes the conclusions drawn from the project. The chapter also covers suggestions of possible future work that we recommend.

2 Background

This chapter gives an overview of the metabolic syndrome and the included risk factors. Furthermore, the general structure and function of nuclear receptors are also described in order to explain their importance as targets for the development of new therapeutics of common metabolic diseases. Data mining is described and the included steps are explained and theory behind the machine learning technique named decision trees is also described. Finally, a review of related work that has been performed on data mining and the metabolic syndrome is included.

2.1 Metabolic Syndrome

Biological risk factors such as hypertension, obesity, dyslipidemia and hyperglycaemia are known to be closely interrelated and to be strong risk factors for both cardiovascular disease and type 2 diabetes (Wamala et al., 1999). In the presence of a combination of these risk factors an even more serious disease has been suggested. The disease has been given various names such as Syndrome X, the Insulin resistance syndrome and the metabolic syndrome (Hansen, 1999). In short one could say that metabolic syndrome is a condition where several different risk factors such as obesity, hypertension, dyslipidemia, diabetes mellitus type II, insulin resistance and microalbuminuria are believed to cooperate to increase the risk of coronary heart disease (Hansen, 1999). Further details regarding the risk factors included in the metabolic syndrome are covered in the following subchapters. Environmental risk factors such as smoking, alcohol and diet can also have an impact on metabolic syndrome patients and it has for example in previous studies been proposed that a possible link exists between cigarette smoking and the metabolic syndrome (Eliasson et al., 1996).

Unfortunately, various definitions of the metabolic syndrome exist and there is no internationally agreed upon definition. In this study a combination of the definition proposed by the World Health Organisation (WHO) and the United States' National Cholesterol Education Programme (NCEP) Adult Treatment Panel III (NCEP, 2001) has been used. The WHO (Alberti & Zimmet, 1998) recommends the following definition of the metabolic syndrome:

> glucose intolerance, impaired glucose tolerance (IGT) or diabetes mellitus and/or insulin resistance together with two or more of the other components listed below:

- Impaired glucose regulation or diabetes mellitus type II
- Insulin resistance
- Raised arterial pressure (140/90 mmHg)

- Raised plasma triglycerides (1.7 mmol l^{-1};150 mg dl^{-1}) and/or low HDL-cholesterol (men: <0.9 mmol l^{-1}, 35 mg dl^{-1}; women: <1.0 mmol l^{-1}, 39 mg dl^{-1})
- Central obesity (males: waist to hip ratio >0.90; females: waist to hip ratio >0.85 and/or BMI>30 kg m^{-2})
- Microalbuminurea (urinary albumin excretion rate 20 µg min^{-1} or albumin:creatinine ratio 30 mg g^{-1})

According to ATP III (NCEP, 2001), the diagnosis of metabolic syndrome can be made when three or more of the risk parameters shown below are present:

- Abdominal obesity (men: >102 cm; women: >88 cm)
- Raised triglycerides (150 mg dl^{-1})
- Low HDL-cholesterol (male: <40 mg dl^{-1}, female: <50 mg dl^{-1})
- High blood pressure (135/85 mmHg)
- Raised fasting glucose (110 mg dl^{-1})

At AstraZeneca´s research and development site in Mölndal, efforts have been made on research on metabolic syndrome and the development of new therapeutic drugs for cardiovascular diseases in general (Halinen & Norseng, 2002). It is of great interest to find possible drug targets for the development of pharmaceuticals for the metabolic syndrome because of the fact that it is a life threatening disorder and also, the number of patients with the syndrome is likely to increase in the future.

2.1.1 Insulin resistance

Insulin resistance is the body's inability to store glucose, especially in skeletal muscle, adipose tissue and/or liver, with the help of insulin (Hellenius et al., 1991). The definition of insulin resistance is a lower than normal response to insulin in cells, tissues or the whole body. Insulin resistance cannot be seen as a disease, but as a physiological change that increases the risk of developing one or more disorders, such as dyslipidemia and high blood pressure (Reaven, 2002). The more resistant an individual is to insulin, the more likely it is that the individual will develop one or more of the abnormalities. In contrast, the more abnormalities present the greater the risk that the patient in addition will develop insulin resistance. Recent studies have also shown that the presence of insulin resistance in an individual is a good predictor for deciding if the patient will develop type II diabetes mellitus (Petersen & Shulman, 2001). Insulin resistance is caused by a number of factors, where the most important are: obesity, physical inactivity and hormones (Grundy, 1999). Additional well-known factors are diet and ageing. Glucose metabolism is closely related to the plasma level of free fatty acids (FFA) (Hellenius et al., 1991). There is always a balance between

insulin levels and the levels of FFA. When the levels of FFA increase, an increase in insulin to the blood can inhibit the release of FFA. If hyperinsulemia cannot be retained, that is if the high levels of insulin cannot be maintained, the levels of FFA will gradually increase which in the end will lead to insulin resistance.

2.1.2 Obesity

Obesity is an increase in body weight beyond the limitation of physical and skeletal requirements, as a result from an excessive accumulation of fat in the body (Andersson, 1994). Body Mass Index (BMI) can be used to determine if an individual is overweight or obese. An overweight person can be defined as a person with a BMI between 25.0-29.9 and an obese person can be defined as a person with a BMI greater or equal to 30 (Hansen, 1999). BMI is calculated as body mass (m) in kilos divided with the square of body length (l^2) in meters. Traditionally BMI has been used as a parameter for diagnosing patients with the obesity risk factor in the metabolic syndrome (Okosun et al., 2000).

Several studies have shown that distribution of fat in the body, especially in the central or abdominal areas, may be an even more important correlate of the metabolic syndrome. Consequently a waist-to-hip ratio has been proposed as a better predictor of the syndrome (Okosun et al., 2000). The waist-to-hip ratio can be used as an indicator of central adiposity and with a ratio of over 0.85 for women and 0.95 for men, the patient can be classified as having central obesity, independent of the value of BMI. Although the waist-to-hip ratio is the most widely used index for accessing abdominal obesity, the parameter has its disadvantages. The waist-to-hip ratio is an imperfect measure of abdominal fat mass, especially in non-obese individuals (Okosun et al., 2000). Insulin resistance is common in individuals with obesity, independent of diabetes, and abdominal obesity has been considered to be dangerous for decreased glucose metabolism (Hellenius et al., 1991). According to the ATP III overweight and obesity are major risk factors for coronary heart disease and weight reduction will reduce all the risk factors in the metabolic syndrome (NCEP, 2001).

2.1.3 Dyslipidemia

Recent analysis shows that elevated triglycerides are additional independent risk factors for the development of coronary heart disease (NCEP, 2001). Higher than normal levels of serum triglycerides can often be seen in persons with the metabolic syndrome diagnosis, although the rise in triglycerides can also depend on genetic factors. Additional factors which contribute to elevated triglycerides are obesity and overweight, physical inactivity, cigarette smoking, excess alcohol intake, high carbohydrate diets, several diseases and certain drugs (NCEP, 2001).

Low high-density lipoprotein cholesterol (low HDL-C) is one additional strong independent risk factor of coronary heart disease (NCEP, 2001). The causes of this abnormality are many, including elevated triglycerides, overweight and obesity, physical inactivity and type II diabetes mellitus.

Features of the metabolic syndrome risk factor dyslipidemia include high triglycerides (hypertriglyceridemia) and low HDL cholesterol values (Hansen, 1999). Hypertriglyceridemia can be defined as fasting triglyceride levels of over 200 mg/dl. Low HDL-C is defined as less than 40 mg/dl for men and less than 50 mg/dl for women (NCEP, 2001). Earlier research has shown that there is a link between the components of dyslipidemia - elevated triglycerides, increased small LDL and decreased HDL cholesterol - and insulin resistance (Hansen, 1999).

2.1.4 Hypertension

High blood pressure or hypertension as it is also named, can be defined as a systolic blood pressure of higher than or equal to 140 mmHg and a diastolic blood pressure of higher than or equal to 90 mmHg (Hansen, 1999). Hypertension promotes atherosclerosis and increases the risk of coronary heart disease and stroke (Campbell, Reece & Mitchell, 1999). Atherosclerosis tends to reduce the diameter of the blood vessels and the elasticity, thereby increasing the blood pressure. Hypertension can be inherited but other factors can also be associated with hypertension, including smoking, reduced physical activity and a high carbohydrate diet. Recent research has further associated hypertension with metabolic abnormalities such as insulin resistance, central obesity and dyslipidemia.

2.1.5 Microalbuminurea

In the last decade, microalbuminurea has been recognised as an individual risk factor of cardiovascular disease (Yip & Trevisan, 1999). It has been associated with all the factors in the metabolic syndrome, including hypertension, raised plasma triglycerides and reduced HDL-cholesterol. Microalbuminurea can be defined as an increase of albumin in the urea (Andersson, 1994). The problem with microalbuminurea is that it is rather difficult to measure with conventional methods. Microalbuminurea is present in about 5-10 percent of the non-diabetic population and the risk of being affected increases with age (Yip & Trevisan, 1999). In patients with hypertension or diabetes, insulin resistance can often be found even in the absence of microalbuminurea, but the opposite is rare. Because of this fact, patients with microalbuminurea should be aware of the possibility of having an increased risk of the metabolic syndrome and the risk factors involved in this syndrome.

2.1.6 Type II Diabetes mellitus

Type II diabetes mellitus can be characterised as elevated glucose levels in the blood and it is either due to a deficiency of insulin or to a reduced responsiveness in target cells because of some change in insulin receptors, or both (Campbell, Reece & Mitchell, 1999). Type II diabetes is a common metabolic disorder and it becomes more likely with increasing age.

The effects of diabetes type II are long-term damage or dysfunction of various organs and people with diabetes also are at increased risk of cardiovascular disease.

Obesity and insulin resistance are common in the early stages of type II diabetes mellitus. The fact is that the majority of patients with diabetes type II are obese and the obesity causes or increases the risk for insulin resistance (Alberti & Zimmet, 1998). WHO has proposed a definition of type 2 diabetes mellitus as a fasting plasma glucose level of 7.0 mmol/l or above or a blood concentration 2 hours after an oral glucose tolerance test of 10.0 mmol/l or above (Alberti & Zimmet, 1998)

2.2 Nuclear Receptors

It has recently been proposed that the dysregulation of nuclear receptors may contribute to the metabolic syndrome (Francis et al., 2002). The fact that nuclear receptors seem to be dysregulated in many common diseases make them good targets for the development of new drugs for the treatment of common metabolic diseases.

The nuclear receptor family includes 49 distinct members in human and is one of the largest groups of transcription factors (Francis et al., 2002). The molecular structure of the nuclear receptors is composed of 5-6 regions that have modular character (Laudet & Gronemeyer, 2002; Chawla et al., 2001). They are characterised by a ligand-independent AF-1 transactivation domain in the NH_2-terminal region and a DNA binding domain, which is composed of two zinc fingers that target the receptor to specific DNA sequences known as hormone responsive elements. Furthermore, they consist of a ligand-binding domain and a AF-2 domain in the COOH-terminal region (see figure 1). In the presence of ligand binding, nuclear receptors undergo a conformational change that dissociates corepressors and enables the recruitment of coactivator proteins to facilitate transcriptional activation (Chawla et al., 2001).

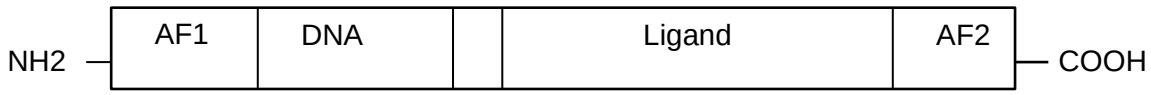

Figure 1. A schematic structure of a typical nuclear receptor.

The major role of nuclear receptors is regulation of genes involved in metabolic control and the binding of small, lipophilic ligands that include hormones and metabolites controls the activity of the receptors. Examples of such ligands are steroid hormones like estrogens and progestines and metabolic ligands such as fatty acids and bile acids (Laudet & Gronemeyer, 2002).

Because of their role in regulation, nuclear receptors have received much interest in the area of finding new therapeutic drugs for common metabolic diseases like diabetes and obesity. Nuclear receptors often form complexes with corepressors and these complexes are responsible for the variability of gene responses to different ligands and metabolic environments (Francis et al., 2002). Since the recognition of nuclear receptors' involvement in the metabolic syndrome, the interest has increased in trying to find out more details regarding this connection (Francis et al., 2002).

Further research has to be performed to be able to establish the truth about the associations and which receptors that are actually involved in the metabolic syndrome.

2.3 Data mining

Data mining is defined as the process of discovering patterns in data (Witten & Frank, 1999). The data set used in data mining is often made up of large quantities of data and it is possibly stored in databases (Thuraisingham, 1998). The amount of biological data in databases has greatly increased, and the data is still increasing all the time. A problem with this increase is that there is a growing gap between the amount of data and our understanding of the data (Witten & Frank, 1999). Data mining is about taking advantage of the potentially useful information, which is hidden in this data. The information is often extracted from large quantities of data, which are possibly stored in databases. For many organisations, the common goals of data mining are to detect abnormal patterns or to predict the future from past experiences and trends (Thuraisingham, 1998). A general problem in bioinformatics is to structure the information into meaningful categories, which is of great importance when trying to establish relationships between different biological data sets (Bertone & Gerstein, 2001).

There are several different steps involved in data mining. In the first step the aim is to organise the data so that it is prepared for mining (Thuraisingham, 1998). Additional steps include determining the desired outcomes to mining, performing the data mining, pruning the results so that only the useful results are taken into further account. Finally, the two last steps involve choosing actions, which should be taken with regard to the mining and to evaluate the actions and determine the benefits from them.

There are various approaches for performing data mining in biological databases and within data mining, a well-known domain exists that includes techniques in machine learning. Machine learning can be used to interpret gene information and it can be used both to split up the data in different categories and to classify earlier unknown examples in an efficient way (Bertone & Gerstein, 2001).

2.3.1 Decision trees

Decision tree learning is an example of a technique in machine learning. It has the ability to both describe available data and to predict the classification of new data (Mitchell, 1997). Decision tree techniques are examples of a form of machine learning called supervised learning (Bertone & Gerstein, 2001). In supervised learning, a priori information is required about the data being classified, in opposite to unsupervised learning where no earlier knowledge is needed. Supervised learning generally includes dividing the data set into two categories: *predictors*, which involve features in the data set that are relevant for learning, and the *response variable*, which is the property to be classified (Bertone & Gerstein, 2001). Supervised learning is conducted in two phases: training and testing, which infer that the data set is split up into two distinct sets. The first set is used to train the model, where the correct classifications are already known

from the beginning, and the instances in the second set are then used to classify according to the partitioning made in the training phase (Bertone & Gerstein, 2001).

A decision tree can be thought of as a tree with labelled nodes (Elomaa, 1996). Decision trees classify instances by sorting them down the tree from the root to some leaf node. A decision tree takes an object as input and outputs a decision, where each decision corresponds to one class. Each internal node in the tree corresponds to a test of the value of one of the properties and the branches from the node are labelled with the possible values of the test. One example of a test could be to investigate if a gene is Present or Absent. If the gene for example is Present, this will correspond to one specific class being assigned and if the gene instead is Absent this will correspond to another class being assigned. This example is however simplified because a decision tree is often constructed by several internal nodes and the tests will in this case lead to another test until a specific leaf node will be reached and the class will be assigned. Each leaf node in the tree corresponds to a value to be returned if that leaf is reached (see figure 2).

The reason why decision trees are successful representations is the fact that the results are easy to interpret in comparison to other learning schemes like neural networks etc. (Elomaa, 1996). The learning process is also fast. The technique is primarily used for classifying which specific category a given case belongs to (Mitchell, 1997).

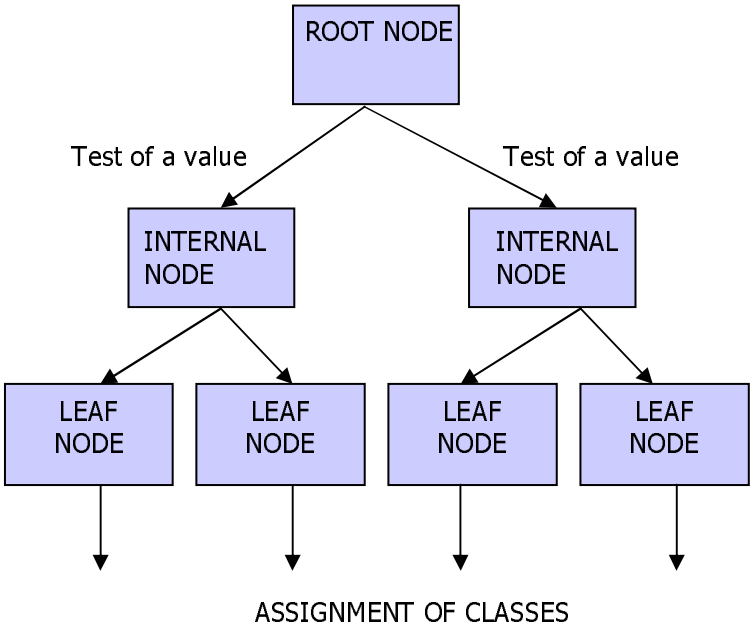

Figure 2. A simplified model of a decision tree. Each node that is not a leaf is connected to a test that splits its set of possible answers into subsets corresponding to different test results. Each branch carries a particular test result's subset to another node. Each leaf node is connected to an assignment of a specified class.

Most algorithms that have been constructed for decision trees are based on variations of a core algorithm that uses a top-down, greedy search through the space of possible decision trees (Mitchell, 1997). The basic decision tree algorithm, ID3 (Quinlan, 1993), can be used as an example to show how decision tree induction is performed. The basic strategy is described below and it is constructed by six steps (Han & Kamber, 2001):

1. A single node that represents the training samples will in the first step be created.
2. If all of the samples are of the same class, then the node will become a leaf, which is then labelled with that class.
3. If this is not true, the algorithm will use an entropy-based measure known as information gain in order to select the attribute that will best separate the samples into individual classes. This specific attribute will become a "test" attribute at that node.
4. A branch is created for each known value of the test attribute and the samples are partitioned correspondingly.
5. The same process is used recursively in order to form a decision tree for the samples in each partition. An attribute that has appeared in one node does not need to be considered anywhere lower in the tree.
6. The partitioning stops only when any of the following conditions are true:

    - All of the samples for a specific node belong to the same class.
    - There are no remaining attributes on which the samples can be further partitioned. In this case, the given node is converted into a leaf and the leaf is then labelled with the class that the majority of the samples belong to.
    - There are no samples for the branch. In this case, a leaf is created that is labelled with the most common class among the samples.

The information gain measure is used to select the best attribute at each node in the growing tree (Han & Chamber, 2001; Mitchell, 1997). The attribute with the highest information gain, i.e. greatest entropy reduction, is chosen as the test attribute for the specific node. The attribute with the highest information gain minimizes the information needed to classify the samples in the resulting partitions. This information gain approach reduces the number of tests that are needed to classify an object and guarantees that a simple tree is found.

When a decision tree has been constructed, many of the branches will reflect inconsistency in the training data due to noise or outliers (Han & Chamber, 2001). The decision tree model has taken too much of the training data into account so that noise has been included in the model as if it was meaningful. Pruning of decision trees can be used to overcome this problem, which is also known as overfitting. The aim with tree pruning is to remove the least reliable branches, which often result in a faster classification and an improved value of correctly classified test data.

2.4 Related work

A study by Rahpeymai (2002) involved using the C4.5 decision tree approach to perform data mining on the Gene Logic database. The decision tree approach was used to try to identify the most relevant genes and risk factors involved in breast cancer, in order to try to separate healthy patients from breast cancer patients in the given data set (Rahpeymai, 2002; Rahpeymai, Olsson & Andersson, 2003). For this purpose four different tests were conducted and for each test a cross validation was performed. In the first test the expression patterns of a set of breast cancer related genes were used as input to the algorithm. The resulting decision tree contained only four genes considered to be the most relevant in order to correctly classify patients. The accuracy in the cross-validation was 89 %. In the second test the risk factors were used as input to the algorithm. The cross validation showed 87% accuracy in classifying the samples. In the third test where both gene expression data and risk factors were used as input, the accuracy of the cross validation was also 87%. In the final test, the algorithm was used to indicate possible signalling pathways involving the four genes identified in the first test. The study demonstrated an application of decision trees for the identification of genes and risk factors relevant for the classification of breast cancer patients.

In a study performed by Halinen and Norseng (2002) the authors investigated the possible relationship between metabolic syndrome and a set of genes using a gene expression database called BioExpress™. The database contains the gene expression profiles for over 7000 diseased and normal human tissue samples. In the first phase of the study, background information about metabolic syndrome and G-Protein coupled receptors (GPCRs) was collected. To be able to classify metabolic syndrome and non-metabolic syndrome patient groups from the BioExpress™ database, the clinical background of metabolic syndrome risk factors was investigated by conducting a literature review. The hypothesis, which was formulated in the project, was that by analysing GPCR expression profiles it would be possible to distinguish tissue samples taken from metabolic syndrome patients from those of non-metabolic syndrome patients. By conducting a literature search and with the help of an expert review panel, the researchers found a total of 21 GPCRs. One critical step in the project was to be able to determine what information was necessary to extract metabolic syndrome patients and non-metabolic syndrome patients and GPCR information from the BioExpress™ database. The solution was to use a diagnostic test for evaluating patients for obesity, hypertension and dyslipidemia according to the WHO and ATPIII metabolic syndrome definitions.

In the next phase, expression data for GPCRs that were differentially expressed in metabolic syndrome and non-metabolic syndrome tissue was analysed. The analysis included both a visualisation of tissue group dissimilarities and an application of the C4.5 decision tree algorithm (Halinen & Norseng, 2002). The results suggest that the expression profile of parathyroid hormone receptor (PTHR1) differs between normal lung tissue samples from metabolic syndrome and non-metabolic syndrome patients taken from the BioExpress™ database. The results also indicate that GPCR expression profiles could not be used to distinguish between cervix and myometrium tissue samples from the same patients group. Furthermore, Halinen and Norseng could not

find any scientific literature reports stating a direct relationship between GPCR gene expression and metabolic syndrome.

Dubitzky et al. (2002) investigated and compared two representatives of two classical machine learning approaches: decision trees and artificial neural networks. The algorithms that were used in the study were the decision tree algorithm C5.0 and the backpropagation algorithm for neural networks. The data set consisted of gene expression data from leukemia patients and the aim of the analysis was to investigate the relative classification performance of the decision tree algorithm C5.0 and the backpropagation artificial neural network and to investigate the methods capability to identify the most relevant genes. The results from the study showed that the best-performing decision tree approach outperformed the best-performing neural network model. The results also showed that the C5.0 decision tree classification model had higher sensitivity and precision. Furthermore, the output provided by the decision tree algorithm is easy to interpret and they are faster to train compared to the neural network model. Through a sensitivity analysis the neural network however was more precise with regard to obtain a ranked list of genes that were interesting, even though this analysis carries a very high computational cost. The conclusion of the study was that the decision tree approach yields more accurate results than the neural network model in terms of classification performance.

3 Presentation of the problem

This chapter includes an explanation of the foundation for this project and a definition of the problem. The aim of the project is discussed and the hypotheses are shown. Finally, the objectives are described and motivated.

3.1 Project foundation

This project is based upon the previous work of Halinen and Norseng (2002) and it concerns a slightly different approach where, instead of using GPCR data as the foundation for the search of drug targets, this project examines the possibility of nuclear receptors as possible targets. The results from the study made by Halinen and Norseng (2002) pointed to the fact that GPCR expression profiles possibly are not useful in distinguishing metabolic syndrome patients from non-metabolic syndrome patients. Their study hinted that GPCR expression did not vary considerably between metabolic syndrome patients and non-metabolic syndrome patients or between different tissues. In this study nuclear receptors and their corresponding co-factors will therefore be used in order to investigate if the NR genes and co-factors could work as possible markers for the metabolic syndrome and the included risk factors.

3.2 Definition of the problem

The aim of this project is to investigate the ability of nuclear receptors to act as biological marker genes for the metabolic syndrome. The motivation for NRs as focus for the project is two-fold. First, it has been observed that NRs seem to be dysregulated in many common metabolic diseases which are thought to make them excellent drug targets for disorders like the metabolic syndrome and other metabolic diseases (Francis et al., 2002). Secondly, at the research and development site at AstraZeneca in Mölndal an available nuclear receptor database exists, which includes information about known target genes and co-factors for NRs.

In the work of Halinen and Norseng (2002) patients were included into the metabolic syndrome patient group if they had disease diagnoses and/or diagnostic test levels indicating the presence of three or four risk factors. This division implies that patients with only two risk factors are not included in the positive patient group, but instead in the negative. Since there is no agreed definition of the metabolic syndrome, the possibility exists that patients, who in the case of Halinen and Norseng ended up in the negative patient group, instead should be in the positive patient group.

The differences in this approach opposed to the work of Halinen and Norseng is that this project aims to classify each of the risk factors individually and then to compare the rules generated from the classifiers. Therefore the problem stated above can be overcome when the different risk factors are analysed individually and the results are later compared. The purpose is to first compare the results generated from

the chosen technique for two of the risk factors and later compare the results for three of the risk factors and so on. By using this approach, it is hypothesised that more information can be extracted about genes that possibly act as gene markers both in the individual risk factors and in the metabolic syndrome.

3.3 Hypotheses

The main hypothesis is that by creating four different classifiers - one for each individual risk factor - and then comparing the results from the classifications, it is possible to identify biological marker genes involved in the metabolic syndrome. If genes can be found in this project that also exist in the Knowledge Bank (Halinen & Norseng, 2002), this could indicate that these genes are associated with the metabolic syndrome and that they possibly could be used as drug targets. The Knowledge Bank includes genes found by a literature review, which are known to be associated with metabolic syndrome risk factors. Three additional hypotheses were formulated for the comparisons of the results from the classifications. First, if one rule can be found to be overlapping in the results for diabetes, dyslipidemia, hypertension and obesity generated by different classifiers, it is hypothesised that the genes in that rule can be used as markers for the metabolic syndrome. Secondly, if individual genes can be found to be overlapping in the rule sets representing the different risk factors, it is hypothesised that the genes can act as gene markers for the metabolic syndrome and the including risk factors. Third, if an assembly of rules are created that contains the rules for each risk factor it is hypothesised that genes can be identified, which can act as genetic markers for the metabolic syndrome.

3.4 Objectives

Specify criteria for the included risk factors. The first step is to determine which criteria should be used to characterise the risk factors. The BioExpress™ database contains both clinical data and gene expression data. The clinical data includes information about for example what diseases that a specific donor has and what test results he or she has. This step includes a literature review to be able to determine the criteria for the individual risk factors. The specified criteria will be based on the WHO and ATP III definitions of the metabolic syndrome and on the work of Halinen and Norseng (2002).

Selection and grouping of donors from the BioExpress™ database. The criteria determined for each of the four risk factors will be used to select clinical data about donors from the BioExpress™ database. The data selected from the database will correspond to five different patient groups, one for each risk factor and one negative group that include patients that do not fulfil any of the criteria.

Selection of strategy for the treatment of missing values. The BioExpress™ database has the disadvantage of having many missing values, which will lead to the number of samples being low. It is necessary to find a method to handle this problem so that the classification can be performed.

Filtering of the NR data. For the nuclear receptors a NR database is available, including both NRs and co-factors. The NR database contains 174 distinct co-factors and 22 distinct target genes for known NRs. It is necessary to search for gene_id and probe set for the nuclear receptors in the NR database, because they will be used to select the relevant gene expression data from the BioExpress™ database. Furthermore, this phase aims to select the best probe sets among a set of multiple probe sets for some specific genes.

Extraction of gene expression data. The different donor groups and the nuclear receptors will be used to extract the relevant gene expression data from the BioExpress™ database.

Selection of a data mining technique(s). In this objective the aim is to choose a data mining technique, which can later be used to classify patients to each of the four different risk factors. The choice of data mining technique(s) will be determined by the chosen gene expression data, depending on how great quantities of data and the quality of the data. The quality of the data represents the existence of errors in the data and the possible missing values. It is necessary that the data mining technique is suitable for the specified problem in this project. One important property is that it must be easy to interpret and compare the results from the algorithm.

Classification of risk factors. The purpose in this objective is to create one classifier for each of the four risk factors. Gene expression data for the donors in the different risk factor groups will be used as input to the classifiers and four different outputs will be generated.

Analysis and comparison of the four classifications. The four created classifiers will be investigated to see if it is possible to find any similarities. The classifiers will be compared in order to identify marker genes for the metabolic syndrome. This objective includes creating a method for comparing the four outcomes from the classifiers and later also to perform the comparison of the results from the classifiers.

Analysis and evaluation of the comparisons. The results from the comparison of the four different classifiers will be analysed. A validation and a comparison with the study by Halinen and Norseng (2002) will be performed in order to investigate if similar genes are found in this project as the genes included in the Knowledge Bank. The Knowledge Bank is a set of genes that Halinen and Norseng (2002) recorded in a literature review on the metabolic syndrome risk factors. The genes included in the Knowledge Bank were found in scientific literature reports and were known to be related to any of the risk factors included in the metabolic syndrome. If identical genes can be found from the projects, this could indicate that these genes are associated with

the metabolic syndrome and that they possibly could be used as drug targets. Furthermore, if genes are found that are not included in the Knowledge Bank, this does not imply that the genes are not interesting, because the Knowledge Bank is not complete. Hence, the possibility exists that genes that until now are not known to be associated with the metabolic syndrome risk factors still can be interesting. Additional information is needed about the possibly found genes and the analysis includes searching the literature for information regarding the genes.

4 Method

The specified method used in the thesis is characterised by the process being very explorative, meaning that the path from the initial specification of the criteria to the final data processing and analysis of the results was not always straightforward. Several problems were encountered, especially in the data cleaning part of the method, where for example missing values and a low number of samples in different risk factor groups occurred. In addition, the cleaning of the NR data also involved problems like choosing the best probe sets for genes that had several probe sets where the expression patterns differed considerably for the same gene. The overall process includes a strong iterative element, especially in the data collection phase of the study. In this project, the intention was to learn more about the metabolic syndrome, by generating one classifier for each risk factor and then compare the results from each risk factor in order to search for genetic markers, an approach which could be referred to as an indirect way for deriving knowledge and new hypotheses and models regarding the metabolic syndrome.

4.1 Overview of the process

The method used in this project includes a series of steps presented in detail in the following subchapters, and a short overview of the whole process is given in figure 3. In the first phase of the project the clinical background on the metabolic syndrome and the included risk factors was investigated to be able to determine the criteria for creating the donor groups for each risk factor. Diagnostic tests were specified from these criteria that could be used for extracting the donors from the BioExpress™ database.

Different methods for the handling of missing values were investigated and two specific methods for this purpose were tried. In the following step data was extracted from the BioExpress™ database in order to construct the different risk factor groups. Donors with extreme lifestyle factors were then removed from the risk factor groups to decrease the risk of having donors that could affect the following data analysis in a negative way. In order to determine which samples would be in the final risk factor groups that would be used as input to the data analysis, research was made with the intent to find the tissues that are best suited for the expression profile analysis.

When the final risk factor groups were created, a data-cleaning phase followed, the aim of which was to investigate the quality of the probe sets. The risk factor data and the NR data were then used as input in order to extract the gene expression data from GeneLogic. A technique for performing the data analysis was chosen and the analysis was made. The results from the data processing were analysed to determine which of the decision trees that would be used in the comparison of the risk factors. The chosen trees were transformed into rules and the rules were then compared. An analysis followed with the aim of evaluating the comparisons.

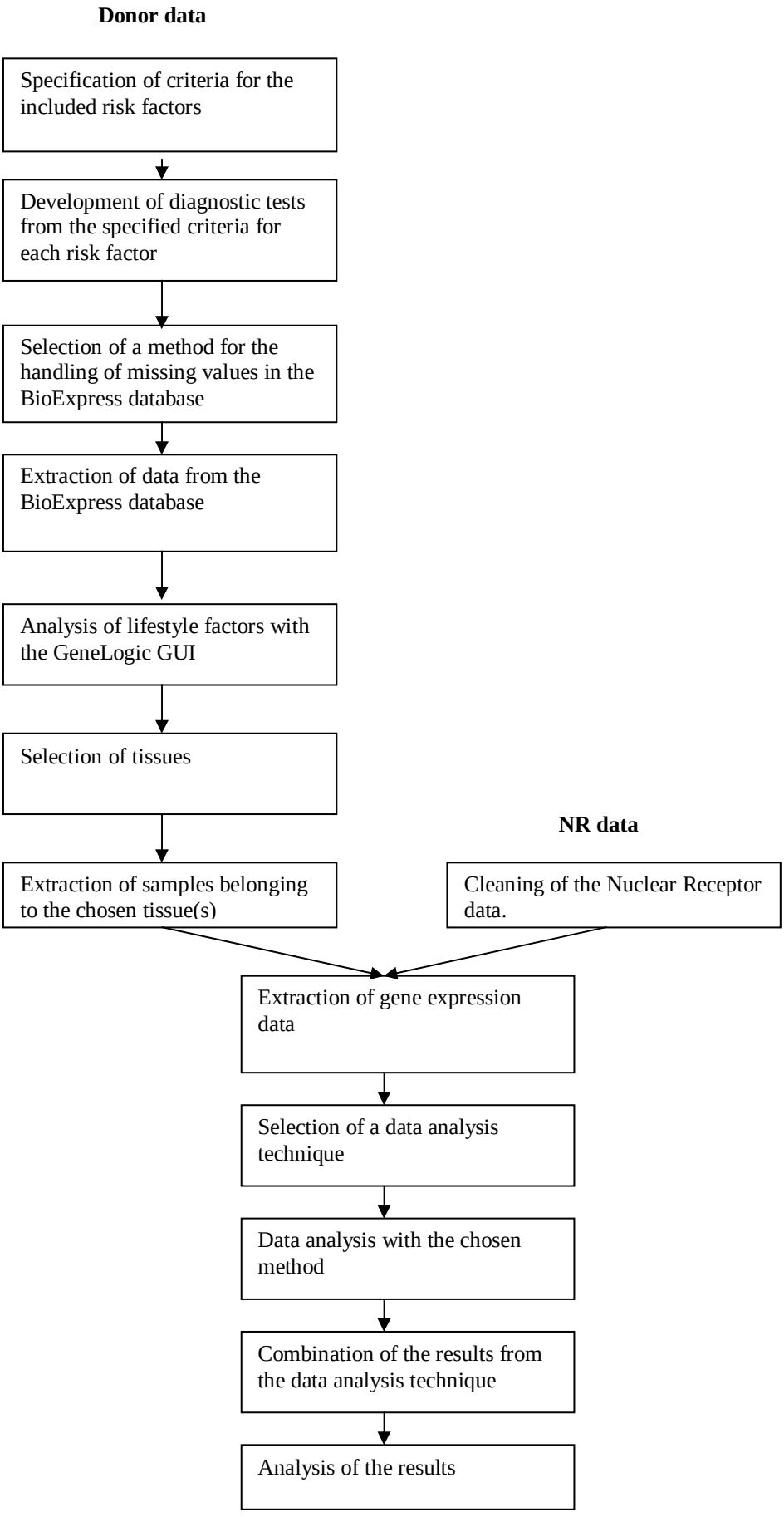

Figure 3. Process model describing the most important steps included in this project.

4.2 Specify criteria for the included risk factors

To be able to define which donors belong to which risk factor group, the first step in the study was to specify criteria for the risk factors, including diabetes mellitus type II, hypertension, dyslipidemia, obesity, microalbuminurea and insulin resistance. In the work made by Halinen and Norseng (2002) criteria were specified for the previously mentioned risk factors to be able to determine which donors could be included in the metabolic syndrome group and in the non-metabolic syndrome group, respectively. The study performed by Halinen and Norseng was used as a starting point when determining the criteria in this work.

A literature review was conducted to investigate if the criteria set up by Halinen and Norseng could be reused without modifications or if any changes would have to be applied. The final criteria specified were based both on the WHO and ATP III definitions of the metabolic syndrome and on the work of Halinen and Norseng (2002). The criteria were used to specify diagnostic tests that later in the project were to be applied to extract donors from the BioExpress™ database. When the BioExpress™ database documentation was investigated it became clear that diagnostic tests were not available for each of the six risk factors. In the case of diabetes mellitus type II, hypertension, dyslipidemia and obesity, diagnostic tests existed in the database but when it comes to microalbuminuria and insulin resistance no test could be applied. Because of this, only diabetes mellitus type II, hypertension, dyslipidemia and obesity could be used in the experiments of classifying the individual risk factors. The diagnostic tests for these risk factors are shown in table 1.

RISK FACTOR	*DIAGNOSTIC TEST*
Diabetes mellitus type II	Fasting/Random blood glucose
Hypertension	Systolic Blood Pressure (SBP)
	Diastolic Blood Pressure (DBP)
Dyslipidemia	High Density Lipoprotein (HDL)
	Triglycerides
Obesity	Body Mass Index (BMI)
	Weight/Length^2 measurement

Table 1. Metabolic Syndrome risk factors and the corresponding diagnostic tests available in the BioExpress database.

Besides the quantitative measurements, qualitative descriptions for the measurements of glucose, high-density cholesterol (HDL) and triglycerides were also considered when determining whether a diagnostic test could indicate a risk factor or not. The qualitative assessment of the measurements has been made by a clinician who has determined if the given test is ´LOW´, ´NORMAL´ or ´HIGH´ when placed in context of that specific donor. It was in this project assumed that the qualitative assessments could give indications of the clinical relevance of the diagnostic test results, which means that a test with both a high measurement value and a ´HIGH´ quality assessment could be considered as more certain than a high measurement value that had a quality assessment that is ´LOW´ or ´NORMAL´. In the case of

hypertension and obesity, no qualitative descriptions were available for their corresponding diagnostic tests in the BioExpress™ database and therefore no qualitative value could be taken into account for these tests.

For each of the four risk factors two different donor groups were created. One donor group was constructed from less stringent criteria and one donor group was constructed from more stringent criteria. The donor group with less stringent criteria had the requirement that a donor had to be guaranteed to fulfil the criteria for the given risk factor to be included in that risk factor group, but there was no constraint against a donor being included in some of the other risk factor groups. The reason for these less stringent requirements was that a more extensive part of the database could be used for the data analysis. The second group was made with a much more stringent requirement, where each donor could only be included in one of the four risk factor groups and no overlap was allowed at all. The motivation for using these groups was that it would be interesting to compare the groups against each other to search for differences and similarities. Table 2 specifies the diagnostic tests used for each risk factor. The final definitions for each risk factor are shown in table 3 for the stringent criteria and table 4 for the less stringent criteria.

RISK FACTOR	DIAGNOSTIC TEST	DESCRIPTION	POSITIVE TEST	NEGATIVE TEST
Diabetes mellitus type II	Glucose	Glucose (mg/dL)	´HIGH´	´NORMAL´
Hypertension	SBP	Systolic Blood Pressure (mmHg)	≥140	<140
	DBP	Diastolic Blood Pressure (mmHg)	≥90	<90
Dyslipidemia	HDL	High Density Lipoprotein (mg/dL)	30≤HDL≤40 or 'LOW´	40<HDL≤80 or ´NORMAL´
	Triglycerides	Triglycerides (mg/dL)	>149 or ´HIGH´	≤149 or ´NORMAL´
Obesity	BMI	Body Mass Index (kg/m^2)	≥30 or $(W/L^2)\geq 30$*	20≤BMI<30 or $20\leq(W/L^2)<30$*

Table 2. Diagnostic tests for the four risk factors in the BioExpress database. The positive tests column includes the diagnostic test for each risk factor that a donor must fulfil in order be included in the specific group. The negative tests correspond to the tests that must be fulfilled for donors included in the negative group, i.e. the donors considered as normal in this thesis.

* W corresponds to the weight of the patient and L corresponds to the length of the patients.

RISK FACTOR GROUP	DONOR INCLUDED
Diabetes mellitus type II	Positive test for Diabetes mellitus (see table 2) or Diagnosis = “DIABETES MELLITUS TYPE II” and **not** Positive test for any other risk factor and **not** Diagnoses for any other risk factor
Hypertension	Positive test for Hypertension (see table 2) or Diagnosis = “HYPERLIPIDEMIA, NOS” or Diagnosis = “HYPERCHOLESTEROLEMIA, NOS” and **not** Positive test for any other risk factor and **not** Diagnosis for any other risk factor
Dyslipidemia	Positive test for Dyslipidemia (see table 2) or Diagnosis = “HYPERTENSION, NOS” or Diagnosis = “HYPERTENSIVE HEART DISEASE, NOS” or Diagnosis = “LIPOIDOSIS, NOS” and **not** Positive test for any other risk factor and **not** Diagnosis for any other risk factor.
Obesity	Positive test for Obesity (see table 2) or Diagnosis = “OBESITY, NOS” and **not** Positive test for any other risk factor and **not** Diagnosis for any other risk factor.
Negative group	Negative test for all risk factors (see table 2) and **not** Diagnosis= “DIABETES MELLITUS TYPE II” and **not** Diagnosis= “HYPERLIPIDEMIA, NOS” and **not** Diagnosis= “HYPERTENSION, NOS” and **not** Diagnosis= “HYPERCHOLESTEROLEMIA, NOS” and **not** Diagnosis= “HYPERTENSIVE HEART DISEASE, NOS” and **not** Diagnosis= “OBESITY, NOS”

Table 3. Final definitions for the risk factor groups created from the stringent criteria

RISK FACTOR GROUP	DONOR INCLUDED
Diabetes mellitus type II	Positive test for Diabetes mellitus (see table 2) or Diagnosis = “DIABETES MELLITUS TYPE II”.
Hypertension	Positive test for Hypertension (see table 2) or Diagnosis = “HYPERLIPIDEMIA, NOS” or Diagnosis = “HYPERCHOLESTEROLEMIA, NOS”
Dyslipidemia	Positive test for Dyslipidemia (see table 2) or Diagnosis = “HYPERTENSION, NOS” or Diagnosis = “HYPERTENSIVE HEART DISEASE, NOS” or Diagnosis = “LIPIDIOSIS, NOS”
Obesity	Positive test for Obesity (see table 2) or Diagnosis = “OBESITY, NOS”
Negative groups	All donors in the GeneLogic database minus the donors in the corresponding risk factor groups specified above

Table 4. Final definitions for the risk factor groups created from the less stringent criteria.

Several changes have been applied compared to the criteria specified by Halinen and Norseng (2002). First of all, in this study each risk factor was investigated individually while Halinen and Norseng (2002) used a combination of risk factors in their method. Furthermore, in the diagnostic test specified for glucose, no quantitative value was used in this project because of the fact that the BioExpress™ database does not distinguish between fasting and random blood glucose. Instead only the quantitative description was used. Regarding the diagnostic test for hypertension, the criteria have become more conservative in the way that the threshold values have been raised to SBP≥140 and DBP≥90 instead of Halinen and Norseng´s criteria of SBP >129 and DBP>84. The motivation for this change is that a more stringent criterion assumedly will give more clear groups with higher confidence in the fact that the donors in the hypertension group are hypertensive. Moreover, according to the WHO a threshold of SBP≥140 and DBP≥90 is recommended. After having a discussion with Dr Gérman Camejo (personal communication, 12 February, 2003) who is a senior principal scientist at AstraZeneca in Mölndal and who has internationally recognised expertise within the field of lipoproteins, it was determined that an interval should be used when defining the criteria for HDL. We did in this thesis not specify different tests for male and female on HDL, which has been made in the previous work by Halinen and Norseng (2002). The thresholds specified for HDL were set on the basis of a discussion with Dr Camejo and from the ATPIII definition of the metabolic

syndrome. An additional modification is that when specifying the qualitative assessment for HDL for the positive test, ´LOW´ was used instead of ´not NORMAL´ because of the fact that an ambiguity occurs when using the latter criteria since ´not NORMAL´ includes both the qualitative assessments ´LOW´ and ´HIGH´. The same motivation is used when changing the qualitative assessment criteria for Triglycerides. ´HIGH´ was used instead of not ´NORMAL´ because of the ambiguity. One additional change in this project was that the diagnosis ”LIPIDOSIS, NOS” has been added to the positive test for dyslipidemia because it can be seen as an additional diagnose for dyslipidemia. In the case of obesity, an interval was used when specifying the criteria for the negative test. According to Dr Camejo (personal communication, 12 February, 2003), a threshold value of 20 could be used as a lower boundary for BMI. Donors with BMI <20 are extreme cases which could affect the following analysis in a negative way.

For the data analysis part of the project, it was necessary to have the four risk factor groups as well as one negative group that contained information about donors that did not belong to the opposite risk factor group. For the negative group made up with stringent criteria, the donors that did not belong to any of the four risk factor groups were included, i.e. the donors that did not have diagnosis or measurements for any of the specified risk factors. The negative groups made up with less stringent criteria instead included all the donors in GeneLogic except for the donors that belonged to the specific risk factor group. For example, the negative group for obesity contains all donors in the GeneLogic database minus the donors in the positive obesity group. In this way an extensive part of the GeneLogic database can be utilised.

4.3 Data extraction

4.3.1. Donor groups

When extracting donor data from the BioExpress™ database, the diagnostic tests discussed in chapter 4.2 (tables 2-4) were used. To extract information from the BioExpress™ database, the Standard Query Language (SQL) was used to query the underlying database tables. The aim of the data extraction was to create groups of donors that were guaranteed to belong to the specified risk factors. However, there was no requirement that the donors could not belong to more that one risk factor group. It was also confirmed that there was no overlap between the different groups.

4.3.2 Missing values

It was clear from the start of the project that the BioExpress™ database contained missing values. Different ways of handling this problem exists: One approach of handling missing values is to discard the units of data that are missing in some variables and to analyse only the complete data (Kotz, Johnson & Read, 1983). In this study this would imply only the use of donors that do not have any missing values in the measurements that are interesting when extracting the data with the diagnostic

tests. A second approach is to use any imputation-based procedure that corresponds to that the missing values are filled in and the resultant complete data are analysed by standard methods (Kotz, Johnson and Read, 1983). Common imputation procedures include substituting the missing values with the mean of the set of recorded values or substituting the missing values with the value that are most common in the set of recorded values. In this study it would be possible to make an estimation of the missing values, either by looking what the mean value are for the specific measurement in the existing values in the database or to look at the most common values in the database. One additional solution could be to investigate the mean values for that specific measurement in the American population and to use that value for imputation. An additional approach could be to set the missing value to the most common occurring value in the database for the specific measurement. Finally a solution is to treat the missing value as a normal value if no other clue is given. The advantage of only including the donors that lack missing values is that the insecurity factor that missing values brings, can be removed. A big disadvantage is however that the number of donors is likely to decrease. Advantages of making estimations of the missing values with respect to the overall population in the United States could be that it better reflects the reality. It is however difficult to make this estimation since it is assumed that the BioExpress™ database includes people that is much more sick than the overall population. The donors included in the database have visited a hospital, assumedly to get help with any problem and they have thereby been recorded in the database. The advantage of using the latter technique, that of treating missing values as normal values, is that the number of samples presumably will increase significantly. Furthermore the technique can possibly be thought of as reasonable, because of the fact that if no value exists, it is probable that a doctor or another person with clinical expertise did not consider it necessary to take the tests from that specific donor. One important disadvantage is the fact that this solution to the problem introduces an insecurity factor, which imply that donors who belong to the specific risk factor group can be falsely excluded and the other way around.

After considering the pros and cons for the different solutions it was decided that donors having a missing value in a column, which result in that a diagnostic test could not be made, should not be included in any risk factor group. This means that only the donors lacking missing values would be used. After trying out this solution to the problem, it became clear that a very low number of samples were left in the hypertension group and in the dyslipidemia group, respectively, to be able to proceed with the data extraction. Possible ways to overcome this problem, with respect to the different techniques for handling missing values reviewed above, were discussed. It was finally decided to use the solution of treating missing values as normal values on the hypertension and the dyslipidemia groups. It was hypothesised that the number of donors would increase greatly when applying the technique, but that it could introduce noise to the tests. Because of this, tables were created later in the process that showed which donors that were used as input to the data processing that had been affected by the technique, i.e. had been treated as normal. The reason was to make it possible to trace the donors when the analysis of the results was made, see Appendix D-F. For the diabetes and the obesity groups, the first solution was instead kept, which was to exclude donors with a missing value from the risk factor groups.

4.4 Analysis of lifestyle factors with the GeneLogic GUI

When the extraction of the donor data was completed, the risk factor group tables could be imported into the GeneLogic database with the GeneLogic GUI, which takes donor-ids as input when importing. When importing the tables into the GeneLogic GUI, it was discovered that samples existed in GeneLogic, which did not exist in the underlying database tables. After examining this fact, it was discovered that it was due to an update, which had not come to our department's knowledge earlier. Because of this, these samples were added manually to the tables where they initially belonged. Furthermore, there were 77 samples in the negative donor group created from stringent criteria, which did not exist on the U133 chip, but on the U95 chip. According to the Affymetrix home page (Affymetrix, 2003) the U133 chip includes the broad number of genes that were featured on its precursors, the U95 chips. However, due to a number of updates in the Affymetrix database the probe sets for the sequences that exist on different chip sets will not be identical. A consequence of this is that data generated from different versions of the human array will not always produce the same results. It is not possible to mix different chipsets, even though they both come from the same organism and because of this, these 77 samples were not taken into account in the future analysis. When importing the risk factor groups made up from the less stringent criteria, no problem occurred.

After importing the donors into GeneLogic, the data sets were examined for donors having extreme lifestyle factors that could have negative effects on the data analysis. Donors in the data sets made up with stringent criteria having a special diet in the form of diabetic diet, were not included in any of the risk factor groups except for the diabetes mellitus type II group. The reason is that even though these donors may not have the diagnosis Diabetes mellitus type II or a ´HIGH´ qualitative assessment of glucose, a diabetic diet could point to the fact that they may be diabetic. In the data sets created by the less stringent criteria the diabetic diet was not taken into account. In a similar way, donors being addicted to alcohol were removed from all groups in both the data set made up from the stringent criteria and the data sets, which originate from the less stringent criteria. The motivation for removing the donors with addiction to alcohol was that it was desired to generate risk factor groups that were representative of the actual population. The control and removal of donors with extreme lifestyle factors was made with the GeneLogic GUI.

To be able to assure that the donors in the risk factor groups made up with the stringent criteria fulfilled all the criteria in the diagnostic tests, these donors were investigated with the GeneLogic GUI. The donors in the negative groups that did not fulfil the negative tests in table 2, were removed from the groups. When this was finished it was assured that all the donors in the risk factor groups with both the strict and the less strict criteria fulfilled all the diagnostic tests specified in the first phase of the project.

4.5 Selection of tissues

To be able to select the tissues that could be used for the data analysis, the GeneLogic GUI could be used. The criterion for using a specific tissue was that there should be more than 5 samples per tissue, because else the samples were considered too few to be able to continue with the data analysis. After consultation with the members of a NR-research group at AstraZeneca in Mölndal, a list of interesting tissues for the nuclear receptors was completed:

I. Adipose tissue
II. Liver
III. Skeletal muscle
IV. Pancreas
V. Small intestine
VI. Kidney

An investigation was made with the purpose of investigating the distribution of gender in the resulting risk factor groups. The motivation for this research was that an overrepresentation of any gender could have a possible impact of the results.

4.6 Filtering of the Nuclear Receptor data

The nuclear receptor database available at AstraZeneca in Mölndal includes two tables. The first table includes data about 22 genes that are known NRs and these genes can therefore be considered as certain, see Appendix A. The second table contains data about 174 co-factors, see Appendix B. Some of the NRs act both as NRs and co-factors and are therefore included in both tables. The genes in the co-factor table are not as certain as the genes in the NR table in a meaning that they are not all acknowledged co-factors to NRs, which could be worth giving notice. The data in the nuclear receptor database contains information about known gene symbol and probe set for the specific gene.

The NRs and co-factors could be imported into GeneLogic with their probe set-ids. The GeneLogic GUI was used in order to examine if any of the probe sets had fragment warnings, which could give evidence about if the given probe set was of good quality. The genes with a corresponding probe set that had a warning were removed from the tables. The fragment warnings corresponded to the fact that the probes are placed too far from 3-prime end or that they matches wrong strand. For the samples that did not have any probe set left after the removal, GeneLogic was queried for a matching probe set, without fragment warning, which could be used instead. For the genes which gene symbols are “SKI” and “RORB” no probe sets were available without fragment warnings in GeneExpress and therefore the genes were not removed from the tables, despite of the fact that they had fragment warnings.

4.7 Extraction of the gene expression data

When the data cleaning of both the donor data and the nuclear receptor data was finished, the extraction of the gene expression data could start. The GeneExpress web interface was used and it takes genomics-ids and gene-ids as inputs. The output from the Web export tool is the gene expression data and the output comes in three different forms: Present/Absent calls, Intensity and P-values. When the extraction of the data was completed, the next step was to look at the data manually in Excel® and Spotfire® to determine if it was possible to find any obvious patterns in the data.

When this analysis was made it was discovered that the gene expression in some cases differed considerable between different probe sets for the same gene. Because of this fact a discussion started with Lisa Öberg (personal communication, 14 March, 2003) who is a member of the bioinformatics group at AstraZeneca in Mölndal and who has experience in working with a method for choosing "good" probe sets for specific genes. Unfortunately, there is no simple procedure for choosing a "good" probe set. The approach that was used for selecting a "good" probe set is described below. All of the probe sets that did not to have fragment warnings were investigated one by one with the help of the E-lab AstraZeneca bioinformatics portal, which contains links and search tools to several functional and structural databases, alignment-tools and additional bioinformatic sites. In the first step, the nucleotide reference sequence (RefSeqN) for the gene was obtained from LocusLink in E-lab. In LocusLink gene-specific information can be accessed for human and RefSeqN gives reference sequence standards for genomes, proteins and standards for human mRNA (Pruitt & Maglott, 2001). The Z-search Multiple Sequence Similarity Search that is provided by E-lab was used to search the Affymetrix Homo sapiens arrays, HG-U133a and HG-U133b to identify probe sets that correspond to the RefSeqN sequence. The result list of probe sets were compared with the non-fragment warning probe sets for the gene. The probe sets that corresponded to the non-fragment warning probe sets were investigated with respect to the percentage identity with the RefSeqN sequence. A 100% identity implies that the non-warning probe set match the RefSeqN at 100%. The higher identity, the higher the chance is that the probe set is correct. One additional test that could give evidence about the quality of the probe set is to look at the alignment to determine whether the probe set is aligned to the RefSeqN sequence in the +/+-direction (correct) or not.

After investigating the probe sets by examining the identity to and their direction with respect to the RefSeqN sequence, it was established that even more research had to be performed to be able to decide which probe set to choose as "good" probe set. Unfortunately, we did not have either the knowledge or the time to move on with these investigations. Because of this, it was decided that multiple probe sets for one and the same gene would be used but that their names would be modified by adding an extra figure after the gene symbol.

4.8 Selection of a data analysis technique

In order to select an appropriate technique for the data analysis phase of the project a literature review was performed and discussions were made with co-workers at AstraZeneca in Mölndal with experience within the field of data mining. The

discussions resulted in a software named Weka-3-2 being found. The Weka-3-2 software consists of a collection of machine learning algorithms and it is implemented in Java (Witten & Frank, 2000). An example of an algorithm that is implemented in Weka-3-2 is the C4.5 decision tree learner (Quinlan, 1993). For background information regarding decision tree algorithms, see subchapter 2.3.1.

The motivations for choosing the Weka-3-2 software including an implementation of the C4.5 algorithm are plentiful. First, decision trees are robust to errors – which includes both errors in classifications of the training examples and errors in the attributes that construct these training examples (Mitchell, 1997). Since the data in the BioExpress™ database have been shown to include errors, this advantageous property of decision trees is very important. In addition, decision tree algorithms can be used even when some training examples have unknown values (Mitchell, 1997), which is also are true in this study where the genes that had Marginal expression were considered as missing because of the insecurity of these abscalls. Furthermore, decision trees are known to generate results that are preferably easy to interpret, which is very important in this study because the results would have to be interpreted in a biological context in order to draw any meaningful conclusions. The advantages of using decision trees in this project compared to for example artificial neural networks are that the neural networks do not provide a theory of what has been learned and they act as black boxes that do not give a clear explanation of the results that have been generated. (Mitchell. 1997).

4.9 Data analysis with Weka-3-2

4.9.1 Weka-3-2

Weka-3-2 contains tools of for example pre-processing, classification, clustering, development of association rules and visualisation of data and all the algorithms are freely available on the World Wide Web (www.cs.waikato.ac.nz/ml/Weka-3-2).

Pre-processing
Since decision trees are supervised learning methods (Bertone & Gerstein, 2001), an extra column must be added in the end of the file, which represents which class the sample belongs to. In this case the column is made up with yes or no's, as a reflection of if the samples belong to the specified risk factor or not. The sample rows in the input matrix were filled with the Present and Absent abscalls for the genes, which means that each gene and sample was assigned one abscall. The Marginal calls were treated as missing values because of the fact that they are hard to handle and that it is not clear how they should be interpreted. The Marginal calls that were treated as missing values did not represent a large part of the total expression patterns for the genes and for most of the genes the percentage of Marginal (missing values) were not higher than 5%. Weka-3-2 handles missing values by splitting the instances, which in this thesis corresponds to donor samples, into pieces (Witten & Frank, 2000). Parts of these instances are then sent down each branch to the leaves of the sub trees involved. The split is achieved by using a numeric weight between 0 and 1. The weights are

chosen to be proportional to the number of instances going down the specific branch. An instance that has been splitted may be further split at a lower node.

In this thesis the instances in Weka-3-2 correspond to the donor samples and the attributes correspond to the genes. The name of the attributes is shown and each attribute is also given a number.

All the attributes were viewed in the pre-processing menu in Weka-3-2 to determine which of them would be used for the processing of the data. For each attribute the number of Absent and Present calls were investigated and the attributes where the number of one specific abscall were less than 10% of the total number of abscalls were removed. Regards were also taken to the percentage of missing values for the attributes. After finishing the filtration of the attributes the user is allowed to apply the filtration to the dataset.

Classification

The method for classification of the different risk factors in the data analysis basically includes four steps, which are described below:

- **ZeroR.** ZeroR is a primitive machine learning scheme, which works by predicting the majority class in the training data if the class is categorical or the average class value if it is numerical (Witten & Frank, 1999). This scheme is not particularly useful for prediction but it is valuable for determining a baseline performance as a benchmark for the other learning schemes. Other learning schemes performing worse than ZeroR indicates serious overfitting. The ZeroR value can be calculated as:

$$ZR = \frac{Number\ of\ \ samples\ in\ the\ largest\ class}{Total\ number\ of\ \ samples}$$

- **OneR.** OneR is slightly more complex than ZeroR and this learning scheme works by producing simple rules based on one attribute only (Witten & Frank, 1999). It takes one single parameter: the minimum number of instances that must be covered by each rule that is generated. By using this learning scheme the best classifying individual attributes can be ranked to give a clue about which attributes are interesting.

- **Initial C4.5 decision tree.** The initial decision tree is a tree that should have as close as possible to 100% correctly classified instances. The C4.5 decision tree learner is called J48 in Weka-3-2 and it has a number of parameters that can be adjusted. The two parameters that we have adjusted are the “minimal number of objects” and the “confidence factor”. When the “minimal number of objects” is set to 1, which means that 1 is the minimal number of objects that reach a leaf node, and the “confidence factor” is set to 0.999 the initial tree is created.

- **Change the parameters of J48.** The result generated in the previous step only guarantees that the generated classifiers are satisfying, but it does not mean that the method is satisfying, as discussed in the quality measurements subchapter (see chapter 4.9.2). The ideal is to have both a high leave-one-out cross-validation and

high percent correctly classified instances on the tree. Leave-one-out cross-validation is a technique where each instance is in turn left out and the learning scheme is trained on the persisting instances. The training is judged based on the correctness of the remaining instances where 1 means success and 0 means failure. The results from the *n* judgements are then averaged and the average represents the final error estimate (Witten & Frank, 1999). The percent correctly classified instances on the tree represents the percentage of the number of instances in the training set that have been classified to the correct class. The percent correctly classified instances should preferably be higher than the cross-validation but as close as possible. Furthermore, the aim is to get a small tree because large trees often imply overfitting and also makes it hard to interpret the results when placed in a biological context. A hypothesis is said to overfit the training examples if some other hypothesis that fits the training examples less well actually performs better over the entire distribution of instances (Mitchell, 1997).

In Weka-3-2 the user can decide to run the algorithm on the full training set or with cross-validation. The choice made was to run with leave-one-out-cross-validation. J48 is the Weka–3-2 implementation of the C4.5 algorithm and in Weka-3-2 different parameters can be set. The parameters that were changed during the runs in this project were confidenceFactor, minNumObj and numFolds. The confidenceFactor sets the confidence threshold for pruning and is a value between 0 and 1. In this project the confidenceFactor was set to 0.999, 0.75, 0.5 and 0.25. The MinNumObj sets the minimum number of instances in any leaf node and this parameter was set to 1, 2, 3, 5, 10 and 20 in this project. The numFolds sets the number of folds in the cross-validation. The value of this parameter must be greater than 1 and in this study the parameter was set to the number of instances minus 1.

4.9.2 Quality measurements

In the ideal case of supervised learning the number of samples in the data set is large and the classes are known for the samples (Witten & Frank, 1999). Generally, the larger number of training samples, the better the classifier. Furthermore, a larger number of test samples generally results in a more accurate error estimate. The dataset can be split up into two parts, where about 1/3 represents the test set and 2/3 represents the training set. Ideally the test set should be representative for the whole dataset. Several different quality measurements can be used to measure quality. Three important quality measurements that are considered in this project are listed below:

1. Validation of the generated classifier.
2. Cross-validation
3. Biological validity

The validation of the generated classifier involves investigating how many errors the classifier has made on the training set. In Weka-3-2 the percentage correctly classified and incorrectly classified instances is given after every run, which makes it easy to understand. The cross-validation implies the validation of how good the decision trees can get on this dataset. There are different kinds of cross-validation and in this thesis a type called leave-one-out cross-validation has been used, which is attractive for two reasons. First, the greatest possible amount of data is used in each training, which is presumed to increase the chance of the classifier being accurate. Secondly, no random sampling is involved. To choose leave-one-out cross-validation in Weka-3-2 the parameter "NumFolds" is set to N-1, where N is the total number of instances. The last element in the quality measurements used in this project is biological validity, which means how to interpret the results in the context of biology and what biological information can be gained from the decision trees. In this project the results generated from the classifiers will be compared to the genes in the specified Knowledge Bank generated from the study made by Halinen and Norseng (2002).

To be able to get both high cross-validation and a high percentage correctly classified instances on the training set the parameters for confidence factor and minimal number of objects were changed and the test options could then be adjusted. Runs were made both on the training set and with cross-validation and the results were then compared.

4.9.3 Data

Since there were enough samples in kidney for both the data sets generated from stringent and less stringent criteria, it was decided that kidney would serve as the tissue used in the project. From the GeneLogic database, gene expression data was extracted both for NRs, co-factors and a combination of both. The overlapping genes that existed in the data that represented a combination of both nuclear receptors and co-factors were removed so that only one copy of that specific gene was left in the data set. The reason why three different data sets were used was that it would be interesting to create decision trees for each of the data set and then compare the results from the data processing. It was hypothesised that the data set that was constructed by a combination of NRs and co-factors should generate the best results because nuclear receptors need their corresponding co-factors to perform their biological activities. The data processing were initially performed on the data containing only NRs on the risk factor groups constructed from the less stringent criteria. Thereafter the processing was made on the comparison of NRs and co-factors and finally, only on the co-factor data. The experiments were then repeated in the same order on the risk factor groups that were created from the stringent criteria.

4.10 Transformation - from trees to rules

When the data analysis with Weka3-2 was finished the next objective was to transform the trees generated from the classifiers into rules. Since the data processing with Weka-3-2 generated multiple trees for each risk factor because of the different parameter adjustments, it was necessary to specify a method for choosing the "best"

trees to transfer into rules. The method used in this thesis was to look at the percentage of leaves/instances in each tree, the cross-validation, the percentage correctly classified instances in the training set, the ConfidenceFactor parameter and the number of leaves. Each tree was analysed manually and the trees that were constructed of only one leaf were initially removed. The next criteria that the trees had to fulfil was that the percentage of the number of leaves divided with the number of instances had to be less than 25%, which is a estimated value, and the reason for this was to reduce the risk of overfitting. The next step after this initial elimination was to rank the existing trees according to their cross-validation. The two trees with best cross-validation were chosen for each risk factor group. If several trees had the same cross-validation, the percent correctly classified instances on the training set were used to select the best trees. If there still existed trees that belonged to the same risk factor group and that had the same cross-validation and percent correctly classified instances on the training set the confidenceFactor was used to differentiate the trees. Finally, the number of leaves in the tree was used to differentiate among the trees so that the trees with the lowest number of leaves were used. The output from using this method was two trees per risk factor group that then would be transformed into rules.

Each tree was transferred into rules manually by creating one string per rule. Each string represented the path from the root to one specific leaf in the tree. Each node in the tree was represented as one element in the string. In each node of the tree one gene is tested if it is Absent (A) or Present (P). All genes that were absent in the node-test got a exclamation mark in front of the gene symbol and each gene that was present were only represented with the gene symbol. The rules that were created from the trees were partitioned according to which class they belonged to, so that all the rules that belonged to class Yes in the specific risk factor group were separated. An example of the transformation process for one tree is shown in figure 4.

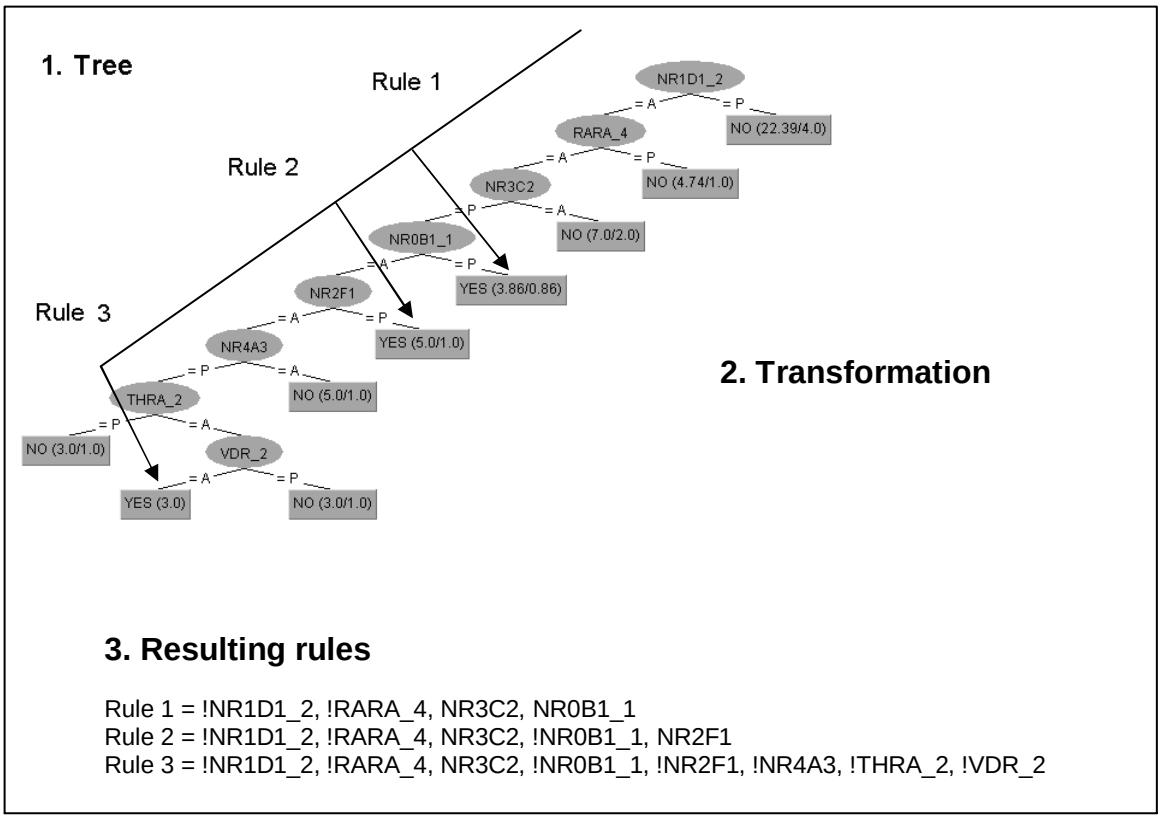

Figure 4. The transformation of a tree into rules. The figure shows the transformation of all paths from the root node to one leaf node where the specified class is yes. Each test in the path from the root to a leaf where the gene is Absent is represented by an exclamation mark in the created rules. In a similar way all genes that are present are represented only by the gene symbol.

4.11 Comparison of rules

4.11.1 Comparison with three different approaches

To be able to compare the rules generated from the trees, an SQL database was created with the motivation that SQL is a language that is very useful in finding similarities and differences between data sets. Additional motivations are that a database gives more ways of combining the data and more information can be used in an easy way. For example information about which pathway the genes belong to can be added. If genes can be identified in the rules that belong to pathways that are known to be associated with metabolic syndrome risk factors, this could indicate that the genes possibly could be used as gene markers for the specific risk factor(s). The database can also be used later if further studies are to be conducted in the future. The database contained three tables, named genes, rules and generules. The two relations genes and rules are connected trough a relation called generules. The relations and their included attributes are shown in figure 5.

```
Genes(gid, pathway_id, pathway_name, db_source)

Rules(rid , name , cf , mno , class)

Generules(rid , gid , gene_id , symbol , abscall)
```

Figure 5. The relations in the created SQL database

The "Genes" relation contains information about the gene_id(gid) for each gene. The gid attribute includes both the old gene_ids extracted from the NR-database, and new gene_ids that correspond to the genes that have been given new gene_symbols because they had several probe sets as discussed in chapter 4.7. The genes included in this relation are the genes from the NR-database available at AstraZeneca in Mölndal. The "Genes" relation also contains attributes about which pathway the genes belong to and the pathways have an id-number and a name. The attribute db_source gives information about which database the pathway originally comes from. The "Rules"-relation is identified by the rid, which corresponds to rule_id. Additional attributes in the relation are the name of the rule, the value of the ConfidenceFactor parameter, the value of the Minimum Number Of Objects parameter and which class the given rule belongs to.

The "Generules" relation includes the attributes "rid" and "gid". The "Generules" relation also includes the attribute gene_id, that is the original gene_id taken from the NR-database, the attribute symbol, which is the gene symbol and the attribute abscall, which explains if the specific gene was Present or Absent in the rule. The SQL-database was then queried in order to compare the rules from different risk factor trees.

Three different approaches were used in order to compare the rules and discover similarities. The first approach was to try to identify complete rules that existed in several risk factor groups. If one rule could be found in for example diabetes, dyslipidemia, hypertension and obesity, it was hypothesised that the rule contained genes that possibly could be used as markers for initially each individual risk factor, but also for the metabolic syndrome. The basic SQL-query that was used in order to investigate which rules that were overlapping between the risk factor groups is shown in figure 6.

```
SELECT GENES.GID
   FROM GENES, GENERULES
       WHERE GENERULES.RID=n AND GENES.GID=GENERULES.GID
INTERSECT
SELECT GENES.GID
   FROM GENES, GENERULES
       WHERE GENERULES.RID=n AND GENES.GID=GENERULES.GID
```

Figure 6. Example of an SQL-query that was used in order to investigate if any rules were found to be overlapping between the risk factors. *n* is the rule_id for the different rules that were changed in the different SQL-queries. The query in figure 6 shows only the comparison of two rules between two risk factors. To perform a comparison of more than two risk factors, the query must be extended additionally so that four intersections will be made.

The second approach was to try to identify genes that existed in several risk factor rule sets and to try to find as many overlapping genes in the rules as possible. For example, when searching for genes that were overlapping between the risk factor rule sets classified as Yes for diabetes and dyslipidemia, the genes with the same symbol were referred to as overlapping. It was also important to assert that the same genes did not exist with the same abscall in the corresponding rule sets classified as No because of the fact that the gene then was not useful in serving as a gene marker for that specific risk factor. Genes that were both Present and Absent in one risk factor rule set, were also thought not to be able to serve as separate markers because of the ambiguity. The basic SQL-query that was used in order to investigate which genes that existed in several risk factor rule sets is shown in figure 7.

```
SELECT GENERULES.SYMBOL
   FROM RULES,GENERULES
      WHERE RULES.RID=GENERULES.RID
        AND RULES.NAME='DIAB2_NR'
        AND RULES.CF=0.5
        AND RULES.MNO=3
        AND RULES:CLASS=YES
INTERSECT
   SELECT GENERULES.SYMBOL
      FROM RULES,GENERULES
        WHERE RULES.RID=GENERULES.RID
           AND RULES.NAME='DYSLIP_NR'
           AND RULES.CF=0.75
           AND RULES.MNO=2
           AND RULES:CLASS=YES
```

Figure7. Example of an SQL-query to investigate if genes existed that were overlapping between the risk factor rule sets. The SQL-query in the figure shows only the comparison of genes that existed in both the diabetes and dyslipidemia rule set. The query was extended to investigate whether the same gene also existed in the rule sets classified as No. To investigate if the overlapping genes existed with the same abscall, an additional SQL-query was used on the tables. The symbols that are shown in italic in the figure represent the ones that were changed in the different SQL-queries.

The third and last approach used for comparing the rules, was to compare the whole rules independently of whether there were rules that were overlapping or not by creating an assembly of the rules. If there for example was a rule that existed in the diabetes rule set that was classified as yes, that rule was assembled with one rule in the dyslipidemia rule set classified as yes and so on so that all possible combinations were generated. A small program that was implemented in C++ performed the generation of the assemblies. This resulted in that donors that fulfilled both the rule included in the diabetes rule set and the rule in the dyslipidemia rule set could be thought of as having both diabetes and dyslipidemia. The same combinations were made for all risk factors to be able to determine the rules for the metabolic syndrome. When combining the rules from different risk factor rule set, the possibility existed that the same genes would occur in several rules in one assembly. If the gene had the same abscall in all of the compared rules in one assembly, the redundancy was removed. If a gene instead existed with different abscalls in one assembly of rules, the whole rule assembly was removed. If genes existed with both abscalls in different assemblies of rules it was not possible to decide which of the abscalls that in fact were true and the genes were not considered to be as interesting as the genes that only had one abscall.

4.11.2 Pathways

In order to investigate if the rules in the different rule sets included genes that were involved in any pathway that was considered as interesting in the context of the metabolic syndrome, two different BioCarta pathways were used. The pathways were named “Basic mechanisms of action of PPARa, PPARb(d) and PPARg and effects on gene expression”, see figure 9, and “Visceral Fat Deposits and the metabolic syndrome”, see figure 10. The genes included in the former mentioned pathway are the Peroxisome Proliferator – Activated Receptor - Alpha (PPARA), the Peroxisome Proliferator – Activated Receptor – Delta (PPARD), the Peroxisome Proliferator – Activated Receptor - Gamma (PPARG), the Retinoic X Receptor – Alpha (RXRA) and the Retinoic X Receptor – Gamma (RXRG). The genes included in the letter pathway are: the Glucocorticoid Receptor (NR3C1), the Retinoic X Receptor – Alpha (RXRA), the Retinoic X Receptor – Beta (RXRB) and the Peroxisome Proliferator – Activated Receptor – Gamma (PPARG). The motivation for looking at these specific pathways was that they both had obvious connections to the metabolic syndrome and the including risk factors. The SQL-database was queried in order to investigate if any of the genes in the rules were also included in the pathways. The basic SQL-query that was used is shown in figure 8. Figure 9 and 10 gives a brief overview of the two chosen pathways.

```
SELECT GENERULES.SYMBOL,GENERULES.ABSCALL
   FROM GENES,RULES,GENERULES
      WHERE GENES.GID=GENERULES.GID
        AND RULES.RID=GENERULES.RID
        AND GENES.PATHWAY_ID=561
        AND RULES.NAME='DIAB2_NR'
        AND RULES.CLASS='YES'
```

Figure 8. The figure shows an example of an SQL-query used to investigate which of the genes in the rule sets that were included in any of the pathways that were used. The symbols shown in italics in the figure represents the symbols that were changed in the different SQL-queries.

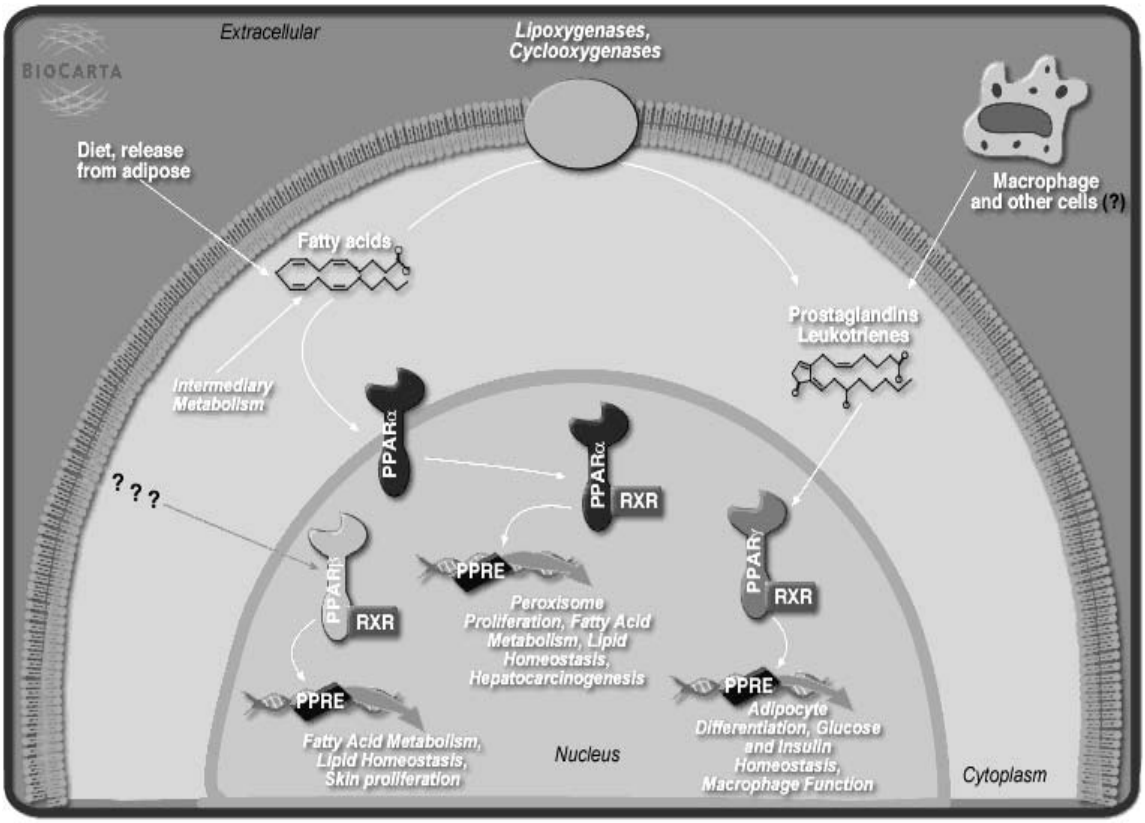

Figure 9. The figure shows an overview of the pathway named "Basic mechanisms of action of PPARa, PPARb(d) and PPARg and effects on gene expression" including the genes with gene symbols PPARA, PPARD, PPARG, RXRA and RXRG. This figure was accessed from the BioCarta home-page (BioCarta, 2003). BioCarta allows the use of any material from their site for a public non-commercial use

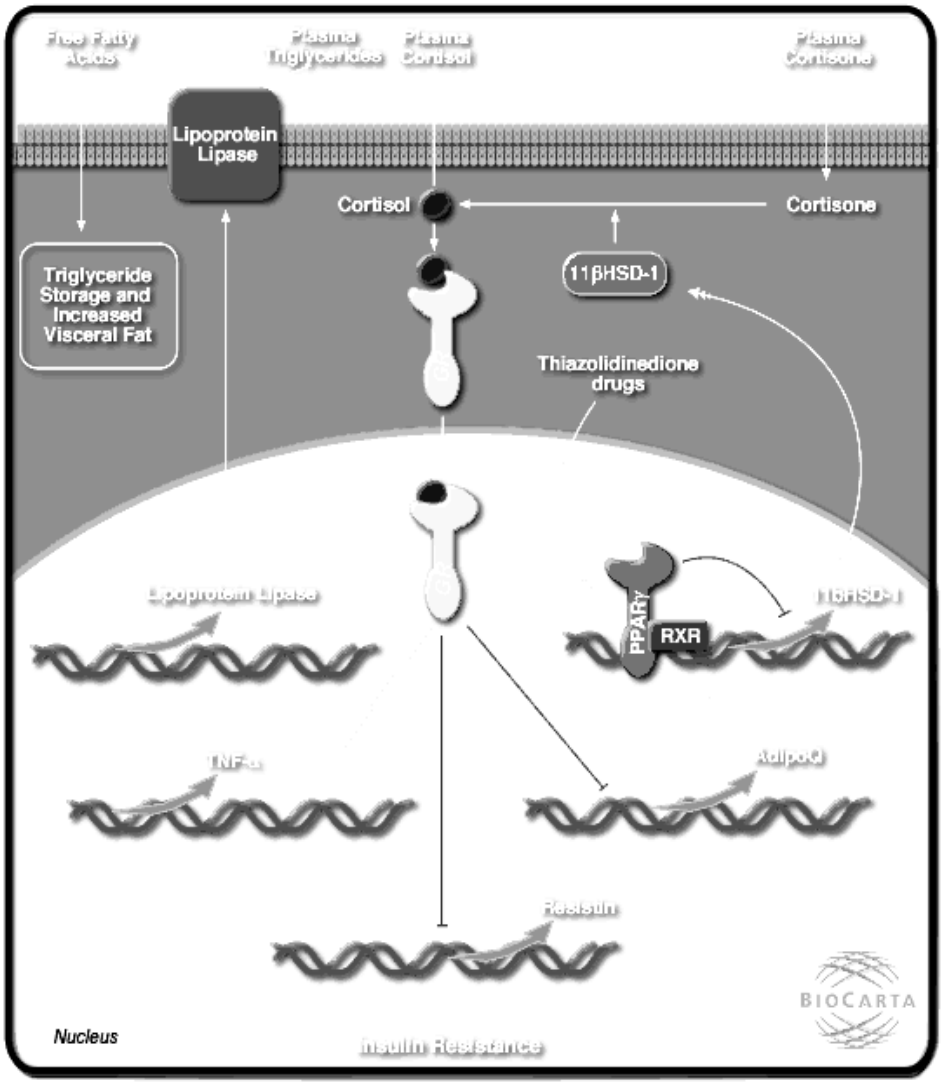

Figure 10. The figure shows an overview of the pathway named "Visceral Fat Deposits and the Metabolic Syndrome" including the genes with gene symbols NR3C1, PPARG, RXRA and RXRB. This figure was accessed from the BioCarta home-page (BioCarta, 2003). BioCarta allows the use of any material from their site for a public non-commercial use

4.12 Data analysis on all risk factors

To be able to evaluate the combinations, new donor groups were created that included donors that fulfilled the criteria of all risk factors included in the metabolic syndrome. The aim was to determine if there were any similarities between the rules generated from the trees with metabolic syndrome patients and the rules that were compared from the individual risk factors included in the metabolic syndrome. It was also interesting to investigate if the quality of the trees with the metabolic syndrome patients would be higher or not compared with the classifications of the individual risk factors. The donor groups for the metabolic syndrome were created from the same

criteria as the separate risk factor groups and one group was created with the stringent criteria and one with the less stringent criteria. It was however noticed that there were only two samples in kidney for the donor groups, which was considered as too few to go on with the study (see table C10 in Appendix C). The donor group that comprised all risk factors except for diabetes did in fact have enough samples in kidney, why it was decided to use the group for further analysis (see table C11 in Appendix C). All the steps described above for cleaning the donor data for the separate risk factor groups were then performed, similarly the removal of donors with extreme lifestyle factors were also conducted with the same method. The gene expression data for the donor groups for the metabolic syndrome except for diabetes were collected for NRs, co-factors and a combination of both.

A data analysis phase was conducted and the same method was used as when classifying the individual risk factors. In Weka-3-2, the same parameters were changed and the trees were evaluated on the basis of cross-validation and the percent correctly classified instances on the training set etc. as described in section 4.9.2. for the evaluation of the single risk factors.

5 Results

This chapter covers the results generated from the study. The result chapter follows the steps in the method chapter and the results are explained and shown as figures or tables.

5.1 Donor data

The diagnostic tests specified in the first phase of the project were used to extract donors from the BioExpress™ database in order to create the different risk factor groups. Initially, when only the donors that did not have any missing values in the measurements in the diagnostic test were used, the results from the data extraction were four different risk factor groups that can be viewed in table 5. The donors included in the positive obesity group are donors that are obese, but that also can be included in some of the other risk factor groups.

Obesity		**Hypertension**		**Dyslipidemia**		**Diabetes type II**	
Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive
2336	968	19	1245	11	267	2149	829

Table 5. Total number of donors fulfilling the negative/positive tests in table 2. The table shows that the number of donors in the negative hypertension and dyslipidemia groups is very low, whereby another technique for the handling of missing values was applied.

There are three possibilities that can explain why a specific donor has been included in the obesity group: the donor either has a historical medical journal that explains the fact that the donor is obese. The donor can also have an event that is obesity, which means that the donor has turned to a hospital to undergo a biopsy because of obesity. The third possibility is that the donor has measurements, i.e. measured data that fulfil the criteria of obesity.

The donors included in the negative obesity group are considered not to be obese from the fact that they do not fulfil any of the criteria expressed above. This is also true for all the other risk factors.

By looking at the risk factor groups in table 5, it can easily be stated that the number of donors in the negative hypertension group and in the negative dyslipidemia group are very low. Because of the fact that donors that were included in the negative groups were not allowed to fulfil any of the criteria explained above, research was done to try to explain why so few donors were included in the negative hypertension group and in the negative dyslipidemia group. The results from the investigation showed that the major bottlenecks in the negative hypertension group were the number of donors that had a measurement of normal systolic and diastolic blood pressure. In the same way the major bottlenecks in the negative dyslipidemia group were the

number of donors with normal measurements of triglycerides and HDL. One probable explanation for this could be that the clinical experts performing the medical examination at the hospital did not consider it necessary to take the specific test from that particular donor.

The technique that involved considering the donors in the negative hypertension and dyslipidemia groups that had missing values in systolic and diastolic blood pressure, triglycerides and HDL, as normal, eliminated the low number of donors. The results from the data extraction with the technique show that the number of donors in the two groups increased greatly, as can be viewed in table 6 below.

Obesity		Hypertension		Dyslipidemia		Diabetes type II	
Negative	Positive	Negative	Positive	Negative	Positive	Negative	Positive
2003	648	3264	1024	4202	159	4389	1086

Table 6. Total number of donors considered negative/positive when the donors having missing values on the measurements of the dyslipidemia and hypertension risk factors and that thereby have been included in the negative dyslipidemia and hypertension group, respectively.

5.2 Creating the Risk factor groups

The donors in table 6 were used as a starting point when creating risk factor groups that fulfilled on one hand the stringent criteria and on the other hand the less stringent criteria, which were determined in the first phase in the study. The results were four risk factor groups constructed from the stringent criteria and that thereby fulfilled the positive test (see table 4) for the four risk factors. It was also clear that no overlap existed between the different groups. The negative group contained donors_ids for the donors that were considered as normal and which thereby were not included in any of the four risk factor groups. However, it is possible that these donors have other diseases that are not considered in this project. The number of donors for each of the groups is shown in table 7.

The groups created from less stringent criteria included donors that were allowed to exist in more than one risk factor groups. The number of donors in each of the groups except for the negative groups is therefore identical with the numbers given in table 6, which was the initial table. The negative groups created from less stringent criteria include all donors in the GeneLogic database minus the donors in the specific risk factor groups. For example the donors in the negative diabetes group include the donors in GeneLogic except for the donors in the positive diabetes group, for exact figures (see table 8).

Risk factor group	Number of donors
Diabetes type II	701
Dyslipidemia	30
Hypertension	505
Obesity	287
Negative group	1221

Table 7. The number of donors in the risk factor groups generated with stringent criteria.

Risk factor group	Number of donors
Diabetes type II	1086
Dyslipidemia	159
Hypertension	1024
Obesity	648
Negative Diabetes type II	169
Negative Dyslipidemia	192
Negative Hypertension	143
Negative Obesity	144

Table 8. The number of donors in the risk factor groups generated with less stringent criteria.

5.2.1 Lifestyle analysis with the GeneLogic GUI

When the risk factor groups constructed from stringent criteria were imported into GeneLogic a total of 77 donors included in the negative group, could not be imported. This resulted in that the 77 donors were excluded from the group. After removing donors with extreme lifestyle factors the numbers decreased additionally. The lifestyle factors that were considered were addiction to alcohol and diabetic diet. The number of donors left after the import and the removal of donors with extreme lifestyle factors, is shown in table 9.

Risk factor group	Number of donors
Diabetes type II	690
Dyslipidemia	30
Hypertension	488
Obesity	282
Negative group	1122

Table 9. The number of distinct donors after removing extreme lifestyles. Stringent criteria.

The import of the risk factor groups with less stringent criteria was performed without any problems. The resulting number of donors in each group can be viewed in table 10.

Risk factor group	Number of donors
Diabetes type II	828
Dyslipidemia	158
Hypertension	1018
Obesity	645
Negative Diabetes type II	166
Negative Dyslipidemia	189
Negative Hypertension	141
Negative Obesity	142

Table 10. Number of distinct donors after the removal of extreme lifestyles. Less stringent criteria.

5.2.3 Selection of interesting tissues

After removing the donors that had extreme values in measurements of the risk factors and exploring which tissues that the samples in the different risk factor groups belonged to, the results were collected in a set of tables. Tables C1-C9 in Appendix C show the number of samples included in the different tissues for all the samples in the risk factor groups constructed from stringent criteria. When comparing the results with the list of interesting tissues made by the NR-group, see chapter 4.5, it was shown that kidney was the only tissue that had more than 5 samples in each risk factor group and that also existed in the list of interesting tissues. The interesting tissues and the corresponding number of samples are marked in Appendix C.

The resulting number of samples after assuring that the negative test was fulfilled in the negative groups with kidney as tissue were the ones that were used in the following data analysis. The exact number of samples is visualised in tables 11-12.

Risk factor group	Number of samples
Obesity	17
Hypertension	31
Diabetes type II	20
Dyslipidemia	8
Negative group	37

Table 11. The number of samples in kidney after removal of donors in the negative group not fulfilling the negative test on groups with stringent criteria.

Risk factor group	Number of samples
Diab2_2_kidney	57
Ob2_kidney	48
Hypert2_kidney	100
Dyslip2_kidney	32
Diab2_2_kidney_neg	220
Dyslip2_kidney_neg	231
Ob2_kidney_neg	180
Hypert2_kidney_neg	247

Table 12. The number of samples in kidney made up with loose criteria.

To be able to trace the donors that were placed in the negative donor group and that had missing values in the measurements on SBP, DBP, HDL or Triglycerides, a table has been created that shows the donor_id for the specific donors. Tables have also been created for the negative groups in dyslipidemia and hypertension constructed from less stringent criteria. The tables are shown in Appendix D, E and F, respectively.

5.2.4 Investigation of the distribution of gender in the data sets

The distribution of males and females in the different risk factor groups in kidney is shown in table 13. The results show that there is a slightly overrepresentation of males in all risk factor groups except for hypertension in the stringent groups and all risk factor group except for dyslipidemia in the less stringent groups.

RISK FACTOR GROUP	MALE	FEMALE
Stringent criteria		
Diab2_kidney	14 (70%)	6 (30%)
Ob_kidney	10 (59%)	7 (41%)
Hypert_kidney	13 (42%)	18 (58%)
Dyslip_kidney	5 (63%)	3 (37%)
Neg_kidney	27 (73%)	10 (27%)
Less stringent criteria		
Diab2_2_kidney	38 (67%)	19 (0.33%)
Ob2_kidney	29 (60%)	19 (40%)
Hypert2_kidney	51 (51%)	49 (49%)
Dyslip2_kidney	14 (44%)	18 (56%)
Diab2_2_kidney_neg	117 (53%)	103 (47%)
Ob2_kidney_neg	126 (55%)	105 (45%)
Hypert2_kidney_neg	105 (58%)	75 (42%)
Dyslip2_kidney_neg	141(57%)	106 (43%)

Table 13. The distribution of gender in the risk factor groups. The table shows both the number of donors that belong to each gender and the corresponding percentage.

When Excel and Spotfire were used to manually check the gene expression data, no obvious patterns could be found. It was however noticed that genes existed that were Present or Absent all the time and that therefore could not be used in the classifications because of the low variance.

5.3 Results from the data analysis with Weka-3-2

The data analysis with Weka-3-2 showed that the majority of the performed classifications had results that did not perform better than the ZeroR classifier, which is the simplest form of learning scheme that exist in Weka-3-2. The results show that the percentage correctly classified instances on the training set seem to decrease when the parameter Minimum number of objects is increased. Similarly, the cross-validation seem to increase when this parameter is increased. The size of the tree and the number of leaves are in the same way decreased. Figure 11 gives an example of output from the Weka-3-2 ZeroR learning scheme. In the run information section of the ZeroR output in figure 11, information is given regarding the number of instances and attributes that was used during the run and a list of the included attributes represented as gene symbols is shown. The test mode represents if the user chose to run with cross-validation or evaluation of the training set. Under the summary part of the output, information is given about the number of (and percentage of) correctly classified instances and incorrectly classified, respectively. The confusion matrix describes the number of instances that are accurately and inaccurately classified by the decision tree.

```
=== Run information ===
Scheme:       weka.classifiers.ZeroR
Relation:     diab2_nr_kidney.csv-weka.filters.AttributeFilter-V-R3-5,11,13,17,20,24,36-40,44-
45,54,58,62-63,70,72-74
Instances:    57
Attributes:   23
              AREG
              ESR1_1
              ESR1_2
              ESRRA_2
              ESRRG
              HNF4A_4
              NR0B1_1
              NR1D1_2
              NR2F1
              NR2F2
              NR2F6
              NR3C1
              NR3C2
              NR4A2_3
              NR4A3
              PPARG
              RARA_4
              RORA_1
              RORA_2
              THRA_2
              VDR_1
              VDR_2
              DIABETES
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: NO

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          37               64.9123 %
Incorrectly Classified Instances        20               35.0877 %
Kappa statistic                          0
Mean absolute error                      0.4574
Root mean squared error                  0.4775
Relative absolute error                100.0905 %
Root relative squared error            100.0415 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   Class
                 0          0          0          0          0        YES
                 1          1          0.649      1          0.787    NO
=== Confusion Matrix ===

   a  b   <-- classified as
   0 20 |  a = YES
   0 37 |  b = NO
```

Figure 11. Example output from Weka-3-2 ZeroR learning scheme.

An example output from Weka-3-2 with the J48 learning scheme, which is the implementation of the C4.5 algorithm in Weka-3-2, when using cross-validation is shown in figure 12. The output from the J48 learning scheme is structured in a similar way compared with the output from the ZeroR learning scheme. The run information section gives information about the number of instances and attributes used in the classification. However, in the Classifier model section a visualisation of the created tree is shown and the number of leaves and the size of the tree are shown. In the Summary section of the J48 output, information about the number and percentage correctly classified instances and incorrectly classified instances are given. The output when evaluating on the full training set is structured in the same way as when using cross-validation, the only difference is that the number of correctly classified instances instead represents the number of correctly classified instances on the training set and that no cross-validation therefore has been used.

=== Run information ===

Scheme: weka.classifiers.j48.J48 -C 0.5 -M 3
Relation: diab2_nr_kidney.csv-weka.filters.AttributeFilter-V-R3-5,11,13,17,20,24,36-40,44-45,54,58,62-63,70,72-74
Instances: 57
Attributes: 23
AREG
ESR1_1
ESR1_2
ESRRA_2
ESRRG
HNF4A_4
NR0B1_1
NR1D1_2
NR2F1
NR2F2
NR2F6
NR3C1
NR3C2
NR4A2_3
NR4A3
PPARG
RARA_4
RORA_1
RORA_2
THRA_2
VDR_1
VDR_2
DIABETES
Test mode: 10-fold cross-validation

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

NR1D1_2 = A
|   RARA_4 = A
|   |   NR3C2 = P
|   |   |   NR0B1_1 = A
|   |   |   |   NR2F1 = A
|   |   |   |   |   NR4A3 = P
|   |   |   |   |   |   THRA_2 = P: NO (3.0/1.0)
|   |   |   |   |   |   THRA_2 = A
|   |   |   |   |   |   |   VDR_2 = A: YES (3.0)
|   |   |   |   |   |   |   VDR_2 = P: NO (3.0/1.0)
|   |   |   |   |   NR4A3 = A: NO (5.0/1.0)
|   |   |   |   NR2F1 = P: YES (5.0/1.0)
|   |   |   NR0B1_1 = P: YES (3.86/0.86)
|   |   NR3C2 = A: NO (7.0/2.0)
|   RARA_4 = P: NO (4.74/1.0)
NR1D1_2 = P: NO (22.39/4.0)

Number of Leaves  :  9
Size of the tree :  17
Time taken to build model: 0.14 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          32               56.1404 %
Incorrectly Classified Instances        25               43.8596 %
Kappa statistic                          -0.0215
Mean absolute error                       0.4889
Root mean squared error                   0.5804
Relative absolute error                 106.9731 %
Root relative squared error             121.6072 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   Class
                 0.25      0.27       0.333     0.25      0.286    YES
                 0.73      0.75       0.643     0.73      0.684    NO

=== Confusion Matrix ===

  a  b   <-- classified as
  5 15 |  a = YES
 10 27 |  b = NO
```

Figure 12. Example output from Weka-3-2 ZeroR learning scheme.

One additional visualisation of the generated tree is shown in figure 13. Each ellipse represents a node in the tree that includes one attribute and each branch is a test for that specific attribute. For example the root node in the tree shown in figure 13 is represented with the attribute named NR1D1_2. The corresponding test for that specific attribute is that instances where the tested attribute NR1D1_2 is Present will be classified as No and instances that instead have an Absent call for the attribute are instead facing an additional test for the attribute named RARA_4. The leaf nodes, i.e. the rectangles in the figure represent the class that the instance fulfilling the specific test will be classified to. The numbers in the parentheses in the rectangles show the number of instances that have been correctly classified and the number of incorrectly classified instances, respectively.

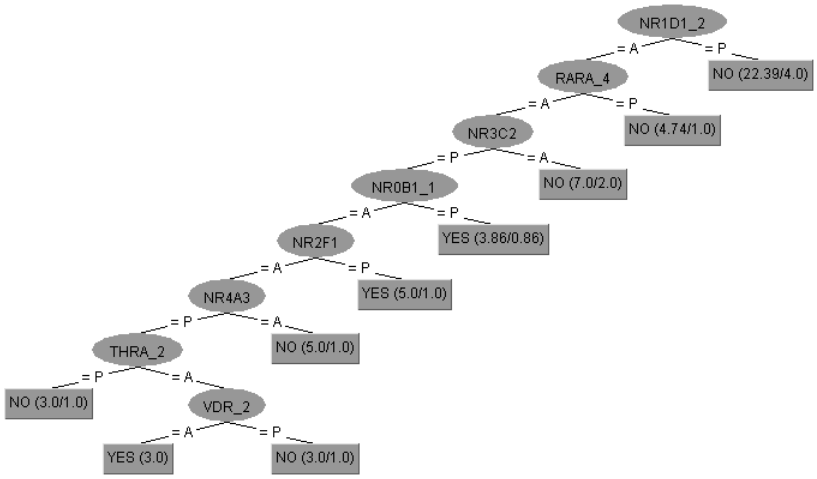

Figure 13. The results for the stringent dyslipidemia donor group and NR with the parameter settings confidence factor: 0.5 and minimum number of objects: 3 visualised as a tree.

The results from the data analysis with Weka-3-2 have been collected into tables that show the parameter settings for each run and the given percent correct classified instances, the cross-validation results, the number of leaves in the tree, the size of the tree and finally the value of the ZeroR classifier. Since there will always be a trade-off between the cross-validation and the percentage correctly classified instances on the training set it is important to consider both of the results when determining the quality of the data processing. Table 14 shows the results from the data analysis with C4.5 in Weka-3-2 on the gene expression data selected for the diabetes group constructed from stringent criteria and NR data. The first column in table 14 includes the parameter settings for the confidenceFactor, which are set to the same values in all the runs performed in this project. The second column includes the

parameter settings for the minimum number of objects used during the data processing in the different runs. This parameter is adjusted in the same way for all the runs performed in this project. The third column contains the results from the runs in the form of the percent correctly classified instances on the training set. The fourth column contains the results from the cross-validation and the fifth and sixth columns include the number of leaves in the tree and the size of the tree, respectively.

The results in table 14, show that no trees exist that have a higher cross-validation compared with the ZeroR classifier. Only four trees exists with the same value of cross-validation compared with ZeroR, but the size of these trees is only one leaf, which means that the trees are too small to use the trees in the following comparisons. It is not however only the cross-validation that is interesting, but also the percentage correctly classified instances on the training set. In table 14 it is shown that the tree with the highest percentage correctly classified instances (i.e. 100 percent) resulted in a low value of the corresponding cross-validation (i.e. 50.9 percent). In table 15 the results from the data processing on the gene expression data from the stringent dyslipidemia group and NR are shown. The table is structured in the same way as table 14 and the same structure is also applied for all the following tables representing the runs with C4.5 in Weka-3-2. The results in table 15 shows that no run with the C4.5 algorithm in Weka-3-2 performed better than the ZeroR classifier, which had a cross-validation of 82.2 percent correctly classified instances. The tree with the highest percentage correctly classified instances (i.e. 100 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 71.1 percent). Ten trees had a cross-validation that equals the ZeroR cross-validation. Table 16 shows the results from the data processing on the gene expression data from the stringent hypertension group and NR. The ZeroR learning scheme had a cross-validation of 54.4 percent and there was no three that performed better than ZeroR. The tree with best cross-validation had 51.5 percent. The tree with the best percentage of correctly classified instances on the training data (98.5) resulted in a corresponding cross-validation of 47.1 percent. Table 17 shows the results from the data processing on the gene expression data from the stringent obesity group and NR. The ZeroR learning scheme had a cross-validation of 68.5 percent. The result from the C4.5 runs shows that a total of 6 trees have cross-validations that are higher than ZeroR, with the value of 81.5 percent. The corresponding percentage of correctly classified instances on the training set were the same as the value of cross-validation, i.e. 81.5 percent. The tree with the highest percentage correctly classified instances (i.e. 100 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 61.1 percent). The results further show that a decrease in the percentage correctly classified instances on the training data often implies an increase in cross-validation. A comparison of the results collected in tables 14-17 shows that table 17, which is the data analysis on the stringent obesity group, is the only table that includes trees that perform better than ZeroR. The three resulting groups: diabetes, dyslipidemia and hypertension all include trees that have cross-validations that are less or equal to their corresponding ZeroR value.

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	50.9	22	43
0.75	1	100.0	52.6	22	43
0.5	1	91.2	54.4	14	27
0.25	1	87.7	52.6	12	23
0.999	2	91.2	52.6	15	29
0.75	2	91.2	50.8	15	29
0.5	2	89.5	47.4	13	25
0.25	2	86.0	54.4	11	21
0.999	3	82.5	54.4	12	23
0.75	3	82.5	54.4	12	23
0.5	3	78.9	56.1	9	17
0.25	3	64.9	54.4	1	1
0.999	5	73.7	52.6	4	7
0.75	5	73.7	52.6	4	7
0.5	5	73.7	52.6	4	7
0.25	5	73.7	52.6	4	7
0.999	10	68.4	52.6	3	5
0.75	10	68.4	52.6	3	5
0.5	10	68.4	50.9	3	5
0.25	10	64.9	56.1	1	1
0.999	20	64.9	64.9	1	1
0.75	20	64.9	64.9	1	1
0.5	20	64.9	64.9	1	1
0.25	20	64.9	64.9	1	1

Table 14. Diabetes – NR results. The table shows the results from the data processing with the gene expression data chosen for the stringent diabetes donor group and NRs as input. ZeroR: 64.9. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	71.1	12	23
0.75	1	100.0	71.1	12	23
0.5	1	97.8	71.1	10	19
0.25	1	82.2	77.8	1	1
0.999	2	97.8	73.3	9	17
0.75	2	97.8	73.3	9	17
0.5	2	97.8	73.3	9	17
0.25	2	82.2	80.0	1	1
0.999	3	88.9	68.9	7	13
0.75	3	88.9	68.9	7	13
0.5	3	82.2	68.9	1	1
0.25	3	82.2	82.2	1	1
0.999	5	88.9	73.3	6	11
0.75	5	88.9	73.3	5	9
0.5	5	82.2	73.3	1	1
0.25	5	82.2	82.2	1	1
0.999	10	82.2	82.2	1	1
0.75	10	82.2	82.2	1	1
0.5	10	82.2	82.2	1	1
0.25	10	82.2	82.2	1	1
0.999	20	82.2	82.2	1	1
0.75	20	82.2	82.2	1	1
0.5	20	82.2	82.2	1	1
0.25	20	82.2	82.2	1	1

Table 15. Dyslipidemia – NR results. The table shows the results from the data processing with the gene expression data chosen for the stringent dyslipidemia donor group and NRs as input. ZeroR: 82.2. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	98.5	47.1	26	51
0.75	1	98.5	47.1	26	51
0.5	1	95.6	50.0	22	43
0.25	1	88.2	41.2	14	27
0.999	2	91.2	47.1	17	33
0.75	2	91.2	47.1	17	33
0.5	2	91.2	47.1	17	33
0.25	2	86.8	44.1	13	25
0.999	3	83.8	47.1	12	23
0.75	3	83.8	47.1	12	23
0.5	3	82.4	45.6	9	17
0.25	3	82.4	42.6	9	17
0.999	5	76.5	51.5	7	13
0.75	5	76.5	51.5	7	13
0.5	5	76.5	47.1	7	13
0.25	5	76.5	41.2	7	13
0.999	10	66.2	35.3	3	5
0.75	10	66.2	35.3	3	5
0.5	10	66.2	35.3	3	5
0.25	10	66.2	35.3	3	5
0.999	20	60.3	41.2	3	5
0.75	20	60.3	41.2	3	5
0.5	20	60.3	41.2	3	5
0.25	20	58.8	41.2	2	3

Table 16. Hypertension – NR results. The table shows the results from the data processing with the gene expression data chosen for the stringent hypertension donor group and NR as input. ZeroR: 54.4. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	61.1	20	39
0.75	1	100.0	61.1	20	39
0.5	1	85.2	70.4	5	9
0.25	1	83.3	75.9	3	5
0.999	2	83.3	55.6	4	7
0.75	2	83.3	63.0	4	7
0.5	2	81.5	74.1	2	3
0.25	2	81.5	77.8	2	3
0.999	3	83.3	68.5	4	7
0.75	3	83.3	79.6	4	7
0.5	3	81.5	81.5	2	3
0.25	3	81.5	81.5	2	3
0.999	5	81.5	81.5	2	3
0.75	5	81.5	81.5	2	3
0.5	5	81.5	81.5	2	3
0.25	5	81.5	81.5	2	3
0.999	10	81.5	66.7	2	3
0.75	10	81.5	66.7	2	3
0.5	10	81.5	66.7	2	3
0.25	10	81.5	70.4	2	3
0.999	20	68.5	68.5	1	1
0.75	20	68.5	68.5	1	1
0.5	20	68.5	68.5	1	1
0.25	20	68.5	68.5	1	1

Table 17. Obesity – NR results. The table shows the results from the data processing with the gene expression data chosen for the stringent obesity donor group and NR as input. ZeroR: 68.5. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

Tables 18-21 show the results from the data processing with the gene expression data chosen for the stringent donor group and a combination of NRs and co-factors as input. Table 18 represents a summary of the results generated from the runs with the gene expression data on the stringent diabetes group and a combination of NRs and co-factors. There are four trees that have better cross-validation results compared with ZeroR, which had a cross-validation of 64.9 percent correctly classified instances and the best tree have a cross-validation of 66.7 The corresponding value of percentage

classified instances on the training set was 87.7 percent. The tree with the highest percentage correctly classified instances (i.e. 98.2 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 59.7 percent). For the dyslipidemia results shown in table 19, the ZeroR cross-validation is 82.2 percent and three trees with a cross-validation of 86.7 percent and which thereby performs better than ZeroR have been created. The corresponding percentage correctly classified instances on the training set are 88.9 percent. The tree with the highest percentage correctly classified instances (i.e. 100 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 66.7 percent).Table 20 includes results for the stringent hypertension group and a combination of NR and co-factors. The ZeroR cross-validation is 54.4 percent and eight trees exist that have a higher value of cross-validation compared with ZeroR. The tree with the highest cross-validation has a value of 57.4 percent correctly classified instances and a corresponding percentage correctly classified instances on training data of 72.1 percent. The tree with the highest percentage correctly classified instances (i.e. 100 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 48.5 percent). Table 21 shows the results for the obesity group and a combination of NR and co-factors. ZeroR is 68.5 percent and 5 trees exist that have a cross-validation that is higher than ZeroR. The tree with the highest value of cross-validation has a value of 77.8 percent correctly classified instances with a corresponding percentage correctly classified instances on the training set of 92.6 percent. The tree with the highest percentage correctly classified instances (i.e. 100 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 57.4 percent).

CF	MNO	TS	CV	Leaves	Size
0.999	1	98.2	57.9	12	23
0.75	1	98.2	57.9	12	23
0.5	1	96.5	56.1	10	19
0.25	1	96.5	59.6	10	19
0.999	2	94.7	63.2	8	15
0.75	2	94.7	63.2	8	15
0.5	2	94.7	63.2	8	15
0.25	2	94.7	63.2	8	15
0.999	3	91.2	63.2	8	15
0.75	3	91.2	63.2	8	15
0.5	3	91.2	63.2	8	15
0.25	3	91.2	63.2	8	15
0.999	5	87.7	64.9	6	11
0.75	5	87.7	64.9	6	11
0.5	5	87.7	66.7	6	11
0.25	5	87.7	66.7	6	11
0.999	10	78.9	59.6	4	7
0.75	10	78.9	59.6	4	7
0.5	10	77.2	59.6	3	5
0.25	10	77.2	57.9	3	5
0.999	20	64.9	56.1	1	1
0.75	20	64.9	56.1	1	1
0.5	20	64.9	56.1	1	1
0.25	20	64.9	57.9	1	1

Table 18. Diabetes – NRs & co-factors results. The table shows the results from the data processing with the gene expression data chosen for the stringent diabetes donor group and a combination of NRs and co-factors as input. ZeroR: 64.9. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	66.7	8	15
0.75	1	100.0	66.7	8	15
0.5	1	100.0	68.9	8	15
0.25	1	95.6	77.8	5	9
0.999	2	97.8	66.7	6	11
0.75	2	97.8	66.7	6	11
0.5	2	97.8	73.3	6	11
0.25	2	93.3	77.8	4	7
0.999	3	95.6	66.7	6	11
0.75	3	95.6	66.7	6	11
0.5	3	93.3	82.2	4	7
0.25	3	93.3	82.2	4	7
0.999	5	88.9	86.7	3	5
0.75	5	88.9	86.7	3	5
0.5	5	88.9	86.7	3	5
0.25	5	86.7	77.8	2	3
0.999	10	86.7	80.0	2	3
0.75	10	86.7	80.0	2	3
0.5	10	86.7	80.0	2	3
0.25	10	86.7	82.2	2	3
0.999	20	82.2	82.2	1	1
0.75	20	82.2	82.2	1	1
0.5	20	82.2	82.2	1	1
0.25	20	82.2	82.2	1	1

Table 19. Dyslipidemia – NRs & co-factors results. The table shows the results from the data processing with the gene expression data chosen for the stringent dyslipidemia donor group and a combination of NRs and co-factors as input. ZeroR: 82.2. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	48.5	16	31
0.75	1	100.0	48.5	16	31
0.5	1	100.0	48.5	16	31
0.25	1	97.1	50.0	14	27
0.999	2	95.6	47.1	14	27
0.75	2	95.6	44.1	14	27
0.5	2	95.6	44.1	14	27
0.25	2	92.6	44.1	12	23
0.999	3	94.1	42.6	13	25
0.75	3	94.1	42.6	13	25
0.5	3	94.1	42.6	13	25
0.25	3	92.6	42.6	12	23
0.999	5	82.4	50.0	7	13
0.75	5	82.4	50.0	7	13
0.5	5	82.4	47.1	7	13
0.25	5	82.4	47.1	7	13
0.999	10	72.1	57.4	4	7
0.75	10	72.1	57.4	4	7
0.5	10	72.1	57.4	4	7
0.25	10	70.6	57.4	3	5
0.999	20	64.7	55.9	2	3
0.75	20	64.7	55.9	2	3
0.5	20	64.7	55.9	2	3
0.25	20	64.7	55.9	2	3

Table 20. Hypertension – NRs & co-factors results. The table shows the results from the data processing with the gene expression data chosen for the stringent hypertension donor group and a combination of NRs and co-factors as input. ZeroR: 54.4. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	57.4	15	29
0.75	1	100.0	57.4	15	29
0.5	1	96.3	57.4	11	21
0.25	1	92.6	61.1	6	11
0.999	2	94.4	64.8	8	15
0.75	2	92.6	66.7	6	11
0.5	2	92.6	63.0	6	11
0.25	2	92.6	75.9	6	11
0.999	3	94.4	64.8	8	15
0.75	3	92.6	66.7	6	11
0.5	3	92.6	68.5	6	11
0.25	3	92.6	77.8	6	11
0.999	5	87.0	68.5	7	13
0.75	5	87.0	70.4	7	13
0.5	5	81.5	72.2	2	3
0.25	5	81.5	77.8	2	3
0.999	10	81.5	64.8	2	3
0.75	10	81.5	64.8	2	3
0.5	10	81.5	64.8	2	3
0.25	10	81.5	68.5	2	3
0.999	20	70.4	68.5	2	3
0.75	20	70.4	68.5	2	3
0.5	20	70.4	68.5	2	3
0.25	20	70.4	68.5	2	3

Table 21. Obesity – NRs & co-factors results. The table shows the results from the data processing with the gene expression data chosen for the stringent obesity donor group and a combination of NRs and co-factors as input. ZeroR: 68.5. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

Tables 22-25 show the results from the data processing with the gene expression data chosen for the stringent donor groups and only the co-factors as input. Table 22 shows the results from the runs with Weka-3-2 for the gene expression data chosen for the stringent diabetes group and the co-factors. The cross-validation for the ZeroR learning scheme is 64.9. No tree exist that have a cross-validation that is higher

than the ZeroR cross-validation but two trees have a cross-validation that is equal to the cross-validation for the ZeroR learning scheme. The corresponding value of correctly classified instances on the training set is 87.7 percent. The tree with the highest percentage correctly classified instances on the trainings set (i.e. 98.2 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 56.1 percent).Table 23 instead shows the results for the gene expression data chosen for the dyslipidemia group generated from stringent criteria and co-factors. The ZeroR cross-validation is 82.2 percent and three trees exist that have a cross-validation of 86.7 percent and which thereby are higher that the ZeroR cross-validation. The corresponding value of correctly classified instances on the training set is 88.9 percent. In table 24 the results from the data processing with the gene expression data chosen for the hypertension group generated with strict criteria and co-factors. The cross-validation for the ZeroR learning scheme is 54.4 percent and eight trees exist that have a cross-validation that is higher than the ZeroR cross-validation. The trees with the highest cross-validation have a cross-validation of 57.4 percent correctly classified instances and the corresponding percentage correctly classified instances on the training data is 72.1 percent. The tree with the highest percentage correctly classified instances on the training data (i.e. 98.5 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 52.9 percent). Table 25 shows the results from the data processing for the gene expression data for the obesity group and co-factors as input. The ZeroR cross-validation is 68.5 percent. No tree exist that have a cross-validation that is higher than the ZeroR cross-validation. However, 5 trees exist that have a cross-validation that equals the cross-validation for ZeroR. The best corresponding value of correctly classified instances on the training set is 70.4 percent.

CF	MNO	TS	CV	Leaves	Size
0.999	1	98.2	56.1	12	23
0.75	1	98.2	56.1	12	23
0.5	1	96.5	46.1	10	19
0.25	1	96.5	56.6	10	19
0.999	2	94.7	63.2	8	15
0.75	2	94.7	63.2	8	15
0.5	2	94.7	63.2	8	15
0.25	2	94.7	61.4	8	15
0.999	3	91.2	61.4	8	15
0.75	3	91.2	61.4	8	15
0.5	3	91.2	61.4	8	15
0.25	3	91.2	61.4	8	15
0.999	5	87.7	63.2	6	11
0.75	5	87.7	63.2	6	11
0.5	5	87.7	64.9	6	11
0.25	5	87.7	64.9	6	11
0.999	10	78.9	54.4	4	7
0.75	10	78.9	54.4	4	7
0.5	10	77.2	54.4	3	5
0.25	10	77.2	52.6	3	5
0.999	20	64.9	56.1	1	1
0.75	20	64.9	56.1	1	1
0.5	20	64.9	56.1	1	1
0.25	20	64.9	57.8	1	1

Table 22. Diabetes – co-factors results. The table shows the results from the data processing with the gene expression data chosen for the stringent diabetes donor group and co-factors as input. ZeroR: 64.9. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	66.7	8	15
0.75	1	100.0	66.7	8	15
0.5	1	100.0	68.9	8	15
0.25	1	95.6	77.8	5	9
0.999	2	97.8	66.7	6	11
0.75	2	97.8	66.7	6	11
0.5	2	97.8	73.3	6	11
0.25	2	93.3	77.8	4	7
0.999	3	95.6	66.7	6	11
0.75	3	95.6	66.7	6	11
0.5	3	93.3	82.2	4	7
0.25	3	93.3	82.2	4	7
0.999	5	88.9	86.7	3	5
0.75	5	88.9	86.7	3	5
0.5	5	88.9	86.7	3	5
0.25	5	86.7	77.8	2	3
0.999	10	86.7	80.0	2	3
0.75	10	86.7	80.0	2	3
0.5	10	86.7	80.0	2	3
0.25	10	86.7	82.2	2	3
0.999	20	82.2	82.2	1	1
0.75	20	82.2	82.2	1	1
0.5	20	82.2	82.2	1	1
0.25	20	82.2	82.2	1	1

Table 23. Dyslipidemia – co-factors results. The table shows the results from the data processing with the gene expression data chosen for the stringent dyslipidemia donor group and co-factors as input. ZeroR: 82.2. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	98.5	52.9	16	31
0.75	1	98.5	52.9	16	31
0.5	1	98.5	52.9	16	31
0.25	1	95.6	52.9	13	25
0.999	2	95.6	48.5	14	27
0.75	2	95.6	47.1	14	27
0.5	2	95.6	47.1	14	27
0.25	2	92.6	48.5	12	23
0.999	3	94.1	47.1	14	27
0.75	3	94.1	47.1	14	27
0.5	3	94.1	47.1	14	27
0.25	3	92.6	48.5	13	25
0.999	5	83.8	54.4	8	15
0.75	5	83.8	54.4	8	15
0.5	5	83.8	52.9	8	15
0.25	5	80.9	48.5	6	11
0.999	10	72.1	57.4	4	7
0.75	10	72.1	57.4	4	7
0.5	10	72.1	57.4	4	7
0.25	10	70.6	57.4	3	5
0.999	20	64.7	55.9	2	3
0.75	20	64.7	55.9	2	3
0.5	20	64.7	55.9	2	3
0.25	20	64.7	55.9	2	3

Table 24. Hypertension – co-factors results. The table shows the results from the data processing with the gene expression data chosen for the stringent hypertension donor group and co-factors as input. ZeroR: 54.4. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CV	MNO	TS	CV	Leaves	Size
0.999	1	96.3	59.3	11	21
0.75	1	96.3	59.3	11	21
0.5	1	96.3	59.3	11	21
0.25	1	96.3	59.3	11	21
0.999	2	92.6	61.1	10	19
0.75	2	92.6	61.1	10	19
0.5	2	90.7	61.1	9	17
0.25	2	79.6	59.3	3	5
0.999	3	90.7	66.7	8	15
0.75	3	90.7	66.7	8	15
0.5	3	90.7	66.7	8	15
0.25	3	90.7	64.8	8	15
0.999	5	85.2	63.0	6	11
0.75	5	85.2	63.0	6	11
0.5	5	85.2	63.0	6	11
0.25	5	75.9	64.8	2	3
0.999	10	75.9	61.1	2	3
0.75	10	75.9	61.1	2	3
0.5	10	75.9	61.1	2	3
0.25	10	75.9	64.8	2	3
0.999	20	70.4	68.5	2	3
0.75	20	70.4	68.5	2	3
0.5	20	70.4	68.5	2	3
0.25	20	68.5	68.5	1	1

Table 25. Obesity– co-factors results. The table shows the results from the data processing with the gene expression data chosen for the stringent obesity donor group and co-factors as input. ZeroR: 68.5. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

Similar runs with the same parameter settings have been generated on the gene expression data selected for the risk factor groups constructed from less stringent criteria. The results from runs with the C4.5 algorithm in Weka-3-2 are shown in tables 26-37. Tables 26-29 shows the results for the classification of the gene expression data chosen for the risk factor groups generated from less strict criteria and NR data. Table 26 shows the results for the gene expression data generated by less strict criteria for the diabetes group. The ZeroR is 79.4 percent correctly classified instances and no tree exists that have a cross-validation that is higher than the ZeroR. Eight trees exist that has a cross-validation that is equal to the ZeroR and the corresponding percentage correctly classified instances on the training set are also 79.4. The tree with the highest percentage correctly classified instances on the training data (i.e. 97.5 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 59.9 percent). Table 27 shows the results for the data processing on the gene expression data generated for the dyslipidemia group generated from less stringent criteria and a NR data. The ZeroR cross-validation is 88.5 percent and no single tree exist that have a higher cross-validation compared with ZeroR. Yet, nine trees exist that have a cross-validation that equals the cross-validation for ZeroR. The corresponding percentage correctly classified instances on the training set are also 88.5 percent. The tree with the highest percentage correctly classified instances on the training set (i.e. 98.9 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 55.0 percent). In table 28 the results for the data processing on the gene expression data chosen for the less stringent hypertension group and NR data are collected. The cross-validation for ZeroR is 64.3 percent correctly classified instances and no tree exist that have a cross-validation that is higher or even equal to the ZeroR cross-validation. The tree with the highest cross-validation has a percentage of 63.2 percent correctly classified instances and the corresponding percentage correctly classified instances on the training set are 64.3 percent. The tree with the highest percentage correctly classified instances on the training set (i.e. 97.1 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 55.0 percent). Table 29 shows the results from the classifications performed on the gene expression data chosen for the less stringent obesity group and NR-data. ZeroR has a cross-validation of 82.8 percent and no tree exist that have a higher cross-validation compared with ZeroR. However, four trees exist that have a cross-validation that is equal to ZeroR and the corresponding percentage correctly classified instances on the training set were also 82.8 percent. The tree with the highest percentage correctly classified instances on the training set (i.e. 98.9 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 72.0 percent).

CF	MNO	TS	CV	Leaves	Size
0.999	1	97.5	59.9	77	153
0.75	1	96.8	59.6	72	143
0.5	1	92.1	64.6	44	87
0.25	1	79.4	78.0	1	1
0.999	2	92.4	62.5	51	101
0.75	2	91.3	64.6	45	89
0.5	2	87.0	69.0	24	47
0.25	2	79.4	78.7	1	1
0.999	3	87.4	66.1	35	69
0.75	3	86.3	68.2	24	47
0.5	3	84.8	70.0	18	35
0.25	3	79.4	78.7	1	1
0.999	5	88.5	74.0	17	33
0.75	5	82.3	74.4	11	21
0.5	5	82.3	77.3	11	21
0.25	5	79.4	78.7	1	1
0.999	10	79.4	79.4	1	1
0.75	10	79.4	79.4	1	1
0.5	10	79.4	79.4	1	1
0.25	10	79.4	79.4	1	1
0.999	20	79.4	79.4	1	1
0.75	20	79.4	79.4	1	1
0.5	20	79.4	79.4	1	1
0.25	20	79.4	79.4	1	1

Table 26. Diabetes – NR results. The table shows the results from the data processing with the gene expression data chosen for the less stringent diabetes donor group and NRs as input. ZeroR: 79.4. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	98.9	77.8	55	109
0.75	1	97.1	77.8	37	73
0.5	1	94.6	82.4	21	41
0.25	1	88.5	88.5	1	1
0.999	2	92.5	83.2	18	35
0.75	2	91.4	87.1	10	19
0.5	2	91.0	87.1	9	17
0.25	2	88.5	88.5	1	1
0.999	3	91.0	84.9	13	25
0.75	3	91.0	88.2	13	25
0.5	3	88.5	88.2	1	1
0.25	3	88.5	88.5	1	1
0.999	5	88.5	88.2	1	1
0.75	5	88.5	88.2	1	1
0.5	5	88.5	88.2	1	1
0.25	5	88.5	88.5	1	1
0.999	10	88.5	88.5	1	1
0.75	10	88.5	88.5	1	1
0.5	10	88.5	88.5	1	1
0.25	10	88.5	88.5	1	1
0.999	20	88.5	88.5	1	1
0.75	20	88.5	88.5	1	1
0.5	20	88.5	88.5	1	1
0.25	20	88.5	88.5	1	1

Table 27. Dyslipidemia – NR results. The table shows the results from the data processing with the gene expression data chosen for the less stringent dyslipidemia donor group and NRs as input. ZeroR: 88.5. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	97.1	55.0	97	193
0.75	1	96.8	55.4	90	179
0.5	1	93.6	55.4	75	149
0.25	1	83.6	55.7	42	83
0.999	2	87.5	53.6	57	113
0.75	2	87.5	54.3	55	109
0.5	2	86.1	52.5	46	91
0.25	2	77.5	56.1	25	49
0.999	3	81.1	50.7	35	69
0.75	3	80.7	50.4	34	67
0.5	3	78.9	50.7	27	53
0.25	3	78.2	53.2	25	49
0.999	5	75.0	55.0	21	41
0.75	5	75.0	55.0	20	39
0.5	5	74.3	53.9	19	37
0.25	5	72.9	57.1	15	29
0.999	10	71.1	53.2	12	23
0.75	10	71.1	53.9	12	23
0.5	10	71.1	53.9	12	23
0.25	10	70.7	57.5	9	17
0.999	20	66.8	58.2	4	7
0.75	20	66.8	58.2	4	7
0.5	20	66.8	59.3	4	7
0.25	20	64.3	63.2	1	1

Table 28. Hypertension – NR results. The table shows the results from the data processing with the gene expression data chosen for the less stringent hypertension donor group and NRs as input. ZeroR: 64.3. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	98.9	72.0	59	117
0.75	1	98.9	72.8	59	117
0.5	1	93.9	75.0	33	65
0.25	1	87.1	82.1	9	17
0.999	2	92.8	73.1	40	79
0.75	2	91.4	73.8	27	53
0.5	2	89.6	75.6	18	35
0.25	2	85.3	82.4	6	11
0.999	3	88.2	77.4	19	37
0.75	3	87.8	78.1	15	29
0.5	3	86.0	79.6	7	13
0.25	3	85.3	82.1	5	9
0.999	5	84.9	80.3	7	13
0.75	5	84.2	80.3	4	7
0.5	5	84.2	80.6	3	5
0.25	5	84.2	81.4	3	5
0.999	10	84.2	81.7	3	5
0.75	10	84.2	81.7	3	5
0.5	10	84.2	81.4	3	5
0.25	10	84.2	80.6	3	5
0.999	20	82.8	82.8	1	1
0.75	20	82.8	82.8	1	1
05	20	82.8	82.8	1	1
0.,25	20	82.8	82.8	1	1

Table 29. Obesity – NR results. The table shows the results from the data processing with the gene expression data chosen for the less stringent obesity donor group and NRs as input. ZeroR: 82.8. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

Tables 30-33 shows the results from the data processing on the gene expression data chosen for the risk factor groups generated from less stringent criteria and a combination of NRs and co-factors. Table 30 show the collection of results for the data processing of the gene expression data chosen for the less stringent diabetes group and a combination of NRs and co-factors. The ZeroR cross-validation is 79.4 percent and no tree exist that have a cross-validation that is higher than ZeroR. One

tree exists that has a cross-validation that equals ZeroR and the corresponding percentage correctly classified instances on the training data is also 79.4 percent. The tree with the highest percentage correctly classified instances in the training set (i.e. 100 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 72.2 percent). Table 31 shows the results for the data processing for the gene expression data chosen for the less stringent dyslipidemia group and a combination of NRs and co-factors. ZeroR has a cross-validation of 88.5 percent correctly classified instances and no trees exist that have a higher cross-validation compared with the ZeroR cross-validation. Six trees exist that have a cross-validation that is equal to the ZeroR cross-validation and the corresponding percentage correctly classified instances on the training data is also the same. The tree with the highest percentage correctly classified instances (i.e. 99.6 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 78.9 percent). In table 32 the results from the data processing with the gene expression data chosen for the less stringent hypertension group and a combination of NRs and co-factors. The cross-validation for the ZeroR learning scheme are 64.3 percent and no tree exists that have a cross-validation that is higher or equal to the cross-validation for ZeroR. The trees with the highest cross-validation have a percentage of 63.2 percent correctly classified instances and the corresponding percentage correctly classified instances is 73.6 percent. The tree with the highest percentage correctly classified instances (i.e. 99.3 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 52.9 percent). Table 33 shows the results for the data processing on the gene expression data chosen for the less stringent obesity group and a combination of NRs and co-factors. ZeroR has a cross-validation of 82.8 percent and no tree exist that have a cross-validation that is higher than the ZeroR cross-validation. However, four trees have been generated that have a cross-validation that is equal to the ZeroR cross-validation. The corresponding percentage correctly classified instances is also the same. The tree with the highest percentage correctly classified instances (i.e. 100 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 71.7 percent).

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	72.2	51	101
0.75	1	100.0	72.2	51	101
0.5	1	98.6	72.2	44	87
0.25	1	98.2	72.9	42	83
0.999	2	95.7	66.1	37	73
0.75	2	95.3	66.1	35	69
0.5	2	95.3	66.5	34	67
0.25	2	93.1	69.7	24	47
0.999	3	93.1	67.1	28	55
0.75	3	93.1	67.5	27	53
0.5	3	93.1	68.2	27	53
0.25	3	91.7	70.4	24	47
0.999	5	88.4	66.8	19	37
075	5	88.4	66.8	19	37
0.5	5	88.4	67.9	19	37
0.25	5	87.0	72.9	13	25
0.999	10	83.8	73.3	10	19
0.75	10	83.8	73.6	9	17
0.5	10	83.8	74.4	9	17
0.25	10	79.4	76.2	1	1
0.999	20	79.4	78.3	1	1
0.75	20	79.4	78.3	1	1
0.5	20	79.4	78.3	1	1
0.25	20	79.4	79.4	1	1

Table 30. Diabetes – NRs and co-factors results. The table shows the results from the data processing with the gene expression data chosen for the less stringent diabetes donor group and a combination of NRs and co-factors as input. ZeroR: 79.4. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	99.6	78.9	35	69
0.75	1	99.6	78.9	33	65
0.5	1	98.9	78.9	29	57
0.25	1	95.7	83.2	18	35
0.999	2	97.0	80.3	23	45
0.75	2	97.5	80.3	23	45
0.5	2	97.5	80.3	23	45
0.25	2	96.1	85.3	18	35
0.999	3	96.8	82.4	20	39
0.75	3	96.8	82.8	20	39
0.5	3	96.1	82.4	17	33
0.25	3	95.0	86.7	14	27
0.999	5	92.1	79.9	12	23
0.75	5	91.4	81.7	8	15
0.5	5	91.4	83.2	5	15
0.25	5	88.5	86.0	1	1
0.999	10	89.6	85.7	4	7
0.75	10	89.6	85.7	4	7
0.5	10	89.6	88.5	4	7
0.25	10	88.5	88.5	1	1
0.999	20	88.5	88.5	1	1
0.75	20	88.5	88.5	1	1
0.5	20	88.5	88.5	1	1
0.25	20	88.5	88.5	1	1

Table 31. Dyslipidemia – NRs and co-factors results. The table shows the results from the data processing with the gene expression data chosen for the less stringent dyslipidemia donor group and a combination of NRs and co-factors as input. ZeroR: 88.5. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	99.3	52.9	69	137
0.75	1	99.3	52.1	69	137
0.5	1	98.2	52.9	60	119
0.25	1	97.5	52.1	58	115
0.999	2	97.1	56.4	52	103
0.75	2	97.1	56.1	52	103
0.5	2	97.1	56.1	51	101
0.25	2	95.0	56.4	45	89
0.999	3	93.9	55.7	42	83
0.75	3	93.9	56.1	42	83
0.5	3	93.6	55.7	40	79
0.25	3	92.9	55.4	37	73
0.999	5	86.4	59.3	28	55
0.75	5	85.4	59.3	25	49
0.5	5	85.4	58.9	25	49
0.25	5	85.0	59.6	24	47
0.999	10	78.2	58.6	14	27
0.75	10	78.2	58.6	14	27
0.5	10	78.2	58.9	14	27
0.25	10	77.5	57.9	13	25
0.999	20	73.6	63.2	8	15
0.75	20	73.6	63.2	8	15
0.5	20	73.6	63.2	8	15
0.25	20	73.6	63.2	8	15

Table 32. Hypertension – NRs and co-factors results. The table shows the results from the data processing with the gene expression data chosen for the less stringent hypertension donor group and a combination of NRs and co-factors as input. ZeroR: 64.3. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	71.7	45	89
0.75	1	100.0	71.3	45	89
0.5	1	99.3	71.0	41	81
0.25	1	88.5	79.9	9	17
0.999	2	97.8	73.1	37	73
0.75	2	97.8	72.8	37	73
0.5	2	97.8	72.4	36	71
0.25	2	88.2	80.6	9	17
0.999	3	96.4	71.3	32	63
0.75	3	96.4	71.3	32	63
0.5	3	95.7	71.3	28	55
0.25	3	88.5	79.6	9	17
0.999	5	90.7	73.8	22	43
0.75	5	90.7	73.8	22	43
0.5	5	87.8	74.2	9	17
0.25	5	86.4	80.3	5	9
0.999	10	85.3	74.9	4	7
0.75	10	85.3	77.1	4	7
0.5	10	85.3	78.9	4	7
0.25	10	85.3	79.2	4	7
0.999	20	82.8	82.8	1	1
0.75	20	82.8	82.8	1	1
0.5	20	82.8	82.8	1	1
0.25	20	82.8	82.8	1	1

Table 33. Obesity – NRs and co-factors results. The table shows the results from the data processing with the gene expression data chosen for the less stringent obesity donor group and a combination of NRs and co-factors as input. ZeroR: 82.8. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

Tables 34-37 show the results from the data processing on the gene expression data chosen for the risk factor groups generated from the less stringent criteria and co-factor data. In table 34 a collection of results from the data processing performed on the gene expression data chosen from the less stringent diabetes group and co-factor data is shown. The ZeroR learning scheme has a cross-validation of 79.4 percent

correctly classified instances. No tree exist that have a cross-validation that is higher compared with the ZeroR cross-validation. One tree exist that have a cross-validation that is equal to the cross-validation of ZeroR. The corresponding percentage of correctly classified instances on the training set is also the same. The tree with the highest percentage correctly classified instances (i.e. 98.9 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 66.6 percent). In table 35 the results from the Weka-3-2 classifications are shown for the gene expression data chosen from the less stringent dyslipidemia group and co-factor data. ZeroR have a cross-validation of 88.5 percent and no tree exists that have a cross-validation that is higher than the ZeroR cross-validation. Six trees exist that have a value of cross-validation that is equal to the cross-validation for the ZeroR learning scheme. The corresponding percentage correctly classified instances on the training set is also the same. The tree with the highest percentage correctly classified instances on the training set (i.e. 99.6 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 77.8 percent). Table 36 shows the results generated from the data processing generated for the gene expression data chosen for the less stringent hypertension group and co-factor data. ZeroR has a cross-validation of 64.3 percent correctly classified instances. No trees exist that have a higher of equal value of cross-validation compared with the ZeroR cross-validation. The trees with the highest cross-validation have a percentage of 62.9 correctly classified instances. The corresponding percentage correctly classified instances on the training set are 73.6. The tree with the highest percentage correctly classified instances (i.e. 99.6 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 53.2 percent). Table 37 shows the collection of results generated from the data processing on the gene expression data chosen for the less stringent obesity group and co-factor data. ZeroR has a cross-validation of 82.8 percent. No trees exist that have a higher or equal cross-validation compared with ZeroR. The trees with the highest value of cross-validation have a percentage of 82.4 percent correctly classified instances. The corresponding percentage correctly classified instances on the training set is 82.8 percent. The tree with the highest percentage correctly classified instances (i.e. 99.3 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 72.0 percent).

CF	MNO	TS	CV	Leaves	Size
0.999	1	98.9	66.6	46	91
0.75	1	98.6	66.4	44	87
0.5	1	98.6	66.8	44	87
0.25	1	97.1	68.2	40	79
0.999	2	95.7	69.0	36	71
0.75	2	95.7	67.9	35	69
0.5	2	95.7	67.9	25	49
0.25	2	93.1	69.0	31	61
0.999	3	93.1	66.4	30	59
0.75	3	93.1	67.1	25	49
0.5	3	91.7	67.9	25	49
0.25	3	90.3	68.6	22	43
0.999	5	89.2	66.1	23	45
0.75	5	88.8	66.4	18	35
0.5	5	88.1	67.1	16	31
0.25	5	87.4	71.8	14	27
0.999	10	81.2	73.6	5	9
0.75	10	81.2	73.3	4	7
0.5	10	81.2	73.3	4	7
0.25	10	79.4	75.5	1	1
0.999	20	79.4	78.3	1	1
0.75	20	79.4	78.3	1	1
0.5	20	79.4	78.3	1	1
0.25	20	79.4	79.4	1	1

Table 34. Diabetes – co-factors results. The table shows the results from the data processing with the gene expression data chosen for the less stringent diabetes donor group and co-factors as input. ZeroR: 79.4. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	99.6	77.8	34	67
0.75	1	99.6	77.8	34	67
0.5	1	98.6	78.1	28	55
0.25	1	94.3	84.9	15	29
0.999	2	97.1	78.5	24	47
0.75	2	97.1	78.1	24	47
0.5	2	97.1	78.5	24	47
0.25	2	93.5	83.5	12	23
0.999	3	96.8	81.7	21	41
0.75	3	96.4	81.7	20	39
0.5	3	96.4	83.2	20	39
0.25	3	93.9	87.1	12	23
0.999	5	91.4	79.9	8	15
0.75	5	91.4	81.0	8	15
0.5	5	91.4	82.1	8	15
0.25	5	88.5	86.4	1	1
0.999	10	89.6	85.7	4	7
0.75	10	89.6	85.7	4	7
0.5	10	89.6	88.5	4	7
0.25	10	88.5	88.5	1	1
0.999	20	88.5	88.5	1	1
0.75	20	88.5	88.5	1	1
0.5	20	88.5	88.5	1	1
0.25	20	88.5	88.5	1	1

Table 35. Dyslipidemia – co-factors results. The table shows the results from the data processing with the gene expression data chosen for the less stringent dyslipidemia donor group and co-factors as input. ZeroR: 88.5. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	99.6	53.2	67	133
0.75	1	99.6	53.2	65	129
0.5	1	99.3	54.3	63	125
0.25	1	96.8	55.0	52	103
0.999	2	96.8	52.9	54	107
0.75	2	96.8	52.5	52	103
0.5	2	96.0	52.5	49	97
0.25	2	95.4	54.3	47	93
0.999	3	93.6	57.5	47	93
0.75	3	93.6	56.4	46	91
0.5	3	92.5	56.8	44	87
0.25	3	87.9	58.2	32	63
0.999	5	86.4	58.6	28	55
0.75	5	86.4	58.2	28	55
0.5	5	86.4	57.9	28	55
0.25	5	85,0	58.6	27	53
0.999	10	78.2	58.9	14	27
0.75	10	78.2	59.3	14	27
0.5	10	78.2	59.3	14	27
0.25	10	77.5	61.4	13	25
0.999	20	73.6	62.9	8	15
0.75	20	73.6	62.9	8	15
0.5	20	73.6	62.9	8	15
0.25	20	73.6	62.9	8	15

Table 36. Hypertension – co-factors results. The table shows the results from the data processing with the gene expression data chosen for the less stringent hypertension donor group and co-factors as input. ZeroR: 64.3. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	99.3	72.0	42	83
0.75	1	98.9	72.4	41	81
0.5	1	98.2	74.2	38	75
0.25	1	98.2	78.9	38	75
0.999	2	96.8	74.9	33	65
0.75	2	96.8	74.6	32	63
0.5	2	96.4	74.9	30	59
0.25	2	88.5	80.3	11	21
0.999	3	95.7	72.8	27	53
0.75	3	95.7	72.4	27	53
0.5	3	95.7	72.4	27	53
0.25	3	88.2	80.3	9	17
0.999	5	92.8	77.1	22	43
0.75	5	928	77.4	22	43
0.5	5	92.8	79.2	22	43
0.25	5	86.7	81.0	6	11
0.999	10	83.9	78.1	3	5
0.75	10	83.9	78.9	3	5
0.5	10	83.9	81.0	3	5
0.25	10	82.8	80.6	1	1
0.999	20	82.8	82.4	1	1
0.75	20	82.8	82.4	1	1
0.5	20	82.8	82.4	1	1
0.25	20	82.8	82.4	1	1

Table 37. Obesity – co-factors results. The table shows the results from the data processing with the gene expression data chosen for the less stringent obesity donor group and co-factors as input. ZeroR: 82.8. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

The results from the data processing with Weka-3-2 can be summarised by the fact that no trees generated from the gene expression data chosen for the risk factors created by the less stringent criteria did have a cross-validation that was higher that the cross-validation for the ZeroR learning scheme. The results from the data processing performed on the gene expression data chosen for the stringent risk factor groups showed that the classifications performed on NR-data resulted in that nine trees generated from the stringent obesity group did have a higher percent cross-validation compared with the corresponding ZeroR cross-validation. In the diabetes,

dyslipidemia and hypertension group constructed from stringent criteria no such tree existed. The results from the data processing performed on the gene expression data chosen for the combination of NRs and co-factors trees existed in all of the four risk factor groups that had a cross-validation that were higher compared to the corresponding ZeroR cross-validation. Furthermore, the results from the data processing performed on the gene expression data chosen for the co-factors alone showed that there existed trees in the dyslipidemia and hypertension group that had a higher percentage of correctly classified instances with cross-validation compared with the corresponding ZeroR cross-validation.

5.4 The generation of rules from the trees

Two trees were chosen from each risk factor run in Weka-3-2 and rules were generated from the chosen trees. The results from the generation of the rules are shown in Appendix G. Each rule consists of a number of genes that are either Present or Absent. Each rule also has a class that can be either Yes or No. The rules belonging to different classes have been separated. For the resulting rules generated from the chosen trees, see Appendix G.

5.5 Results from the combinations

5.5.1 First approach – searching for overlapping rules

When searching for entire rules that existed in more than one risk factor rule set, using the created SQL-database, it was shown that there was no overlap between any of the rule sets classified as Yes. This was confirmed for the risk factor rule sets constructed by NRs, co-factors and the combination of NRs and co-factors where each rule in one risk factor rule set was compared with all the rules in the other risk factor rule sets and so on.

5.5.2 Second approach – searching for rules with overlapping genes

When searching for rules in different risk factor groups that had overlapping genes, the SQL-database was used and the results were again not very promising. When looking for rules that had overlapping genes in all of the four risk factor rule sets, not even one gene was found to be overlapping in the groups. This was true for all rule sets, both the ones made up of only NRs, the ones made up of only co-factors and the rule sets that were constructed from a combination of NRs and co-factors.

The combination of rules that existed in three different risk factor groups on the rule sets generated from the loose criteria on the NR rule sets, resulted in the finding of the vitamin D receptor gene, which had the gene symbols VDR_1 and VDR_2. The symbols existed in the diabetes, dyslipidemia and hypertension rule sets classified as Yes and were Absent. VDR exists in two forms in this analysis because of the fact that the gene had two probe sets, as described earlier. For VDR_2 the major drawback was that there also existed rules in the same rule sets that included the gene

VDR_2 that was instead Present. These results show that it is not possible to discriminate if the gene should be Present or Absent in order to serve as a gene marker for the three risk factors. The same gene also existed in the corresponding rule sets classified as No and the gene existed both as Present and Absent. Finally the gene also existed in the obesity rule set classified as No. The results points to that the VDR_2 gene cannot be used as a gene marker alone because of the fact that it exists with the same abscall both in the rule sets classified as Yes and the ones classified as No. And the fact that the gene also existed with the same abscall in the obesity rule set classified as No. For VDR_1 exactly the same problem occurred.

For the rule sets made up with stringent criteria on NR, the gene named Peroxisome Proliferator – Activated Receptor – Gamma with the gene symbol PPARG was overlapping between the rule sets of diabetes, dyslipidemia and hypertension classified as Yes. PPARG existed as Absent in all of the rule sets, but there also existed rules that included PPARG as Present. PPARG also existed in the corresponding rule sets classified as No and the gene was there both Present and Absent. Between the rules in the diabetes, hypertension and obesity rule set classified as No, The Amphiregulin gene with the gene symbol AREG was overlapping, but the same problem as described above for PPARG and VDR existed. In a similar way a gene, which is an orphan nuclear receptor with gene symbol NR1I3 that was overlapping in dyslipidemia, hypertension and obesity, was not considered as a promising result.

For the rule sets constructed from the combination of NR and co-factors and the rule sets made up of only co-factors, no overlap existed whatsoever from the combination of four risk factors and three risk factors.

The combinations of two risk factors resulted in the overlap of a few genes between the different risk factor groups. However, when investigating the overlap in more detail, one of the following two alternatives was always true. Either the overlapping genes also existed in the corresponding rule sets classified as No or the genes had the same abscall as the gene in the rule set classified as Yes, or the overlapping gene existed with the same abscall in the two other risk factor rule sets classified as No. The genes also often existed both as Present and as Absent in the rules that included the overlapping genes. In summary, it can be stated that the genes that were found to be overlapping between the different risk factor groups, were not able to alone discriminate patients belonging to the metabolic syndrome from patients classified as Non-metabolic syndrome patients.

5.5.3 Third approach – combining rules from different risk factors

The combinations of whole rules from the different risk factors resulted in a large number of rules. The combinations were made on the rule sets made up of stringent and less stringent criteria and for the rules constructed from NRs, co-factors and a combination of NRs and co-factors. After removing the rules where a gene existed several times, but with different abscall and removing duplicates with the same abscall, the resulting genes were listed. The resulting lists included rules that could be interpreted as if a donor fulfilled any of the rules in the list, the donor would be classified to the metabolic syndrome. Some of the genes in the lists were either Present or Absent in every rule in the specified list while some genes were Present in some

rules and Absent in some rules. Furthermore, some genes only existed in a few rules while others existed in all rules in the list. Figures containing the compared rules for the rule sets belonging to the diabetes, dyslipidemia, hypertension and obesity group constructed from stringent criteria are shown in Appendix D and serves as an example of how the other lists are constructed. The rules in Appendix D include some genes that have an abscall that is the same in all of the rules that the specific rule exists in. The results from the assembly of the rules in the diabetes, dyslipidemia, hypertension and obesity rule set created from stringent criteria and from NR-data are shown in tables 38-39. The genes that only existed with one abscall are shown in table 38. The genes that existed in the rules that had both the abscall Present and Absent are shown in table 39. The union of the tables 38 and 39 includes all the genes found in the combination of rules between the diabetes, dyslipidemia, hypertension and obesity groups constructed by stringent criteria and NR data.

Gene	Abscall	Number of rules
NR2F1	A	220
NR2F6	P	29
PPARG	A	220
THRA_2	A	23
ESRRG	P	135
HNF4A_4	A	16
RORA_1	A	220
NR3C1	A	96
ESRRA	P	78
NR4A2_3	A	55
RORA_2	A	64
NR1H4	A	20

Table 38. The genes included in the rules that only had one abscall in the assembly of the rules for all risk factors constructed from stringent criteria and NR data.

Gene	Abscall	Number of rules
NR1D1_2	P	32
NR1D1_2	A	188
RARA_4	P	26
RARA_4	A	65
NR3C2	P	143
NR3C2	A	57
NR4A3	P	55
NR4A3	A	65
ESR1_2	P	57
ESR1_2	A	92
NR0B1_1	P	32
NR0B1_1	A	23
VDR_1	P	48
VDR_1	A	39
VDR_2	P	44
VDR_2	A	23
AREG	P	42
AREG	A	88
NR1I3	P	63
NR1I3	A	126
PPARA	P	87
PPARA	A	33
NR4A2_3	P	68
NR4A2_3	A	64

Table 39. The genes included in the rules that only had both abscalls in the assembly of the rules for all risk factors constructed from stringent criteria and NR data.

The results from the assembly of the rules in the diabetes, dyslipidemia, hypertension and obesity rule set created from less stringent criteria and from NR data are shown in tables 40-41. The genes that only existed with one abscall are shown in table 40. The genes that existed in the rules that had both the abscall Present and Absent are shown in table 41.

Gene	Abscall	Number of rules
VDR_1	P	91
NR1I3	A	49
NR2F2	P	42
AR	A	49
RARB	P	21
NR1D1_2	A	38
ESRRG	A	21
THRA_2	P	39
RARA_4	A	91
PPARA	P	91
AREG	A	7
NR2F1	A	91
NR3C1	A	91
RORA_1	A	91
NR0B2	A	91
NR2F6	P	14
ESRRA_2	P	26

Table 40. The genes included in the rules that only had one abscall in the assembly of the rules for all risk factors constructed from less stringent criteria and NR data

Gene	Abscall	Number of rules
NR4A1	P	49
NR4A1	A	42
HNF4A_1	P	41
HNF4A_1	A	21
RXRG	P	49
RXRG	A	14
VDR_2	P	26
VDR_2	A	66
ESR2_1	P	28
ESR2_1	A	63
NR4A2_3	P	25
NR4A2_3	A	19
RORA_2	P	14
RORA_2	A	14
NR4A2_1	P	52
NR4A2_1	A	14
NR1H4	P	27
NR1H4	A	22
NR4A1	P	49
NR4A1	A	42

Table 41. The genes included in the rules that had both abscalls in the assembly of the rules for all risk factors constructed from less stringent criteria and NR data

The results from the assembly of rules from diabetes, dyslipidemia, hypertension and obesity created from stringent criteria and a combination of NRs and co-factors are shown in tables 42-43. Table 42 shows the genes that only existed with one abscall in the rules while table 43 shows the genes that existed with both abscalls.

Gene	Abscall	Number of rules
HTATIP_2	A	108
CRY1	P	108
RFX5	A	108
TCFL4_1	P	36
MSC	P	36
EPAS1_2	A	36
MED6_2	P	36
ZFP95_1	A	36
SMARCA4_5	A	36
SP140	A	36
HDAC7A	A	27

Table 42. The genes included in the rules that only had one abscall in the assembly of the rules for all risk factors constructed from stringent criteria and a combination of NRs and co-factors.

Gene	Abscall	Number of rules
MEF2D_1	P	54
MEF2D_1	A	54
ELF3_1	P	27
ELF3_1	A	27
BLZF1_1	P	27
BLZF1_1	A	27
NFYA_1	P	36
NFYA_1	A	36
CBFA2T1_1	P	36
CBFA2T1_1	A	72
RFXAP	P	36
RFXAP	A	36
PPARA	P	72
PPARA	A	36

Table 43. The genes included in the rules that had both abscalls in the assembly of the rules for all risk factors constructed from stringent criteria and a combination of NRs and co-factors.

The results from the assembly of rules from diabetes, dyslipidemia, hypertension and obesity created from less stringent criteria and a combination of NRs and co-factors are shown in tables 44-45. Table 44 shows the genes that only existed with one abscall in the rules while table 45 shows the genes that existed with both abscalls

Gene	Abscall	Number of rules
DDX17_3	A	1695
SMARCA2_2	P	1695
SUP3TH	A	1695
PPARA	P	1437
PAX8_4	P	1289
CALR_1	P	929
ARNT2	P	760
HMGCS2	P	760
ELF4_1	A	590
SAP30_2	A	590
HTATIP2_1	P	558
VDR_1	P	558
EGR2	P	520
TLE1_2	A	520
NFYC_4	P	392
TCF2_2	A	392
BCL3_2	A	377
SMARCA4_3	P	325
ZNF145	P	324
BTG2	P	262
GLI2	A	262
TRIP11	P	262
GADD45B_3	P	260
NR5A2_1	P	260
ZFP95_1	A	243
VDR_2	P	232
GPS2_2	P	216
NFYA_1	P	130
NOTCH2_2	P	130
NT5C	A	130
RYBP_3	P	130
TNRC11_2	A	130
PPARG	A	116
ZFP95_2	P	115
TFDP2_1	P	77

Table 44. The genes included in the rules that had only one abscall in the assembly of the rules for all risk factors constructed from less stringent criteria and a combination of NRs and co-factors.

Gene	Abscall	Number of rules
CREB1_1	P	785
CREB1_1	A	130
DEDD_1	P	60
DEDD_1	A	78
ELF3_1	P	783
ELF3_1	A	262
HNF4A_1	P	377
HNF4A_1	A	377
MED6_1	P	260
MED6_1	A	1305
MEF2A	P	131
MEF2A	A	131
MSC	P	116
MSC	A	116
MTF1_2	P	313
MTF1_2	A	174
NR2F1	P	130
NR2F1	A	130
NR3C2	P	392
NR3C2	A	131
PAX8_6	P	520
PAX8_6	A	260
PMF1	P	558
PMF1	A	1133
RING1_2	P	260
RING1_2	A	260
RIPK3	P	181
RIPK3	A	962
SMARCD2	P	608
SMARCD2	A	216
SMARCE1	P	780
SMARCE1	A	915
SUV39H1	P	130
SUV39H1	A	262
TTF2	P	266
TTF2	A	357
ZNF226_1	P	1565
ZNF226_1	A	130

Table 45. The genes included in the rules that had both abscalls in the assembly of the rules for all risk factors constructed from less stringent criteria and a combination of NRs and co-factors.

The results from the assembly of rules from diabetes, dyslipidemia, hypertension and obesity created from stringent criteria and co-factors are shown in tables 46-47. Table 46 shows the genes that only existed with one abscall in the rules while table 47 shows the genes that existed with both abscalls.

Gene symbol	Abscall	Number of rules
CRY1	P	149
HTATIP_2	A	149
RFX5	A	149
NFKB2	A	84
NT5C	P	84
TCFL4_1	P	51
MED6_2	P	48
EPAS1_2	A	46
BCL3_2	A	43
DEDD_2	P	43
NCOA2_2	P	40
HDAC7A	A	36

Table 46. The genes included in the rules that only had one abscall in the assembly of the rules for all risk factors constructed from stringent criteria and co-factor data.

Gene symbol	Abscall	Number of rules
MEF2D_1	P	72
MEF2D_1	A	77
BLZF1_1	P	27
BLZF1_1	A	50
ELF3_1	P	72
ELF3_1	A	27
NFYA_1	P	49
NFYA_1	A	49
CBFA2T1_1	P	48
CBFA2T1_1	A	97
RFXAP	P	51
RFXAP	A	49
HNF4A_1	P	18
HNF4A_1	A	82
LHX3	P	43
LHX3	A	43

Table 47. The genes included in the rules that had both abscalls in the assembly of the rules for all risk factors constructed from stringent criteria and co-factor data.

The results from the assembly of rules from diabetes, dyslipidemia, hypertension and obesity created from less stringent criteria and co-factors are shown in tables 48-49. Table 48 shows the genes that only existed with one abscall in the rules while table 49 shows the genes that existed with both abscalls.

Gene	Absca	Number of rules
DDX17_3	A	111
HNF4A_1	P	111
IRF1	P	111
MED6_1	A	111
PMF1	A	111
SMARCA2_2	P	111
ZNF226_1	P	111
HMGCS2	P	56
ELF4_1	A	47
SAP30_2	A	47
ARNT2	A	46
PAX8_6	P	43
RIPK3	A	43
TLE1_2	A	43
SMARCA4_3	P	23
EGR2	P	22
GLI2	A	22
ZNF145	P	22
E4F1	A	20
ESR2_1	A	17
FOXF2	A	11
SIAH2	P	11

Table 48. The genes included in the rules that only had one abscall in the assembly of the rules for all risk factors constructed from less stringent criteria and co-factor data.

Gene	Abscall	Number of rules
SMARCE1	P	43
SMARCE1	A	69
TTF2	P	26
TTF2	A	22
RING1_2	P	22
RING1_2	A	22
ELF3_1	P	78
ELF3_1	A	34
TBL1X_2	P	22
TBL1Z_2	A	21
TNRC11_2	P	11
TNRC11_2	A	22
TCFL4_1	P	18
TCFL4_1	A	17
DKFZp761F01	P	71
DKFZp761F01	A	41

Table 49. The genes included in the rules that both abscalls in the assembly of the rules for all risk factors constructed from less stringent criteria and co-factor data.

It was shown that a total of five genes were found that existed in the rules representing an assembly of risk factor rules classified as Yes that also existed in the Knowledge Bank. These genes were the Peroxisome Proliferator – Activated Receptor – Gamma (PPARG), the Peroxisome Proliferator – Activated Receptor – Alpha (PPARA), Musculin (MSC), the Glucocorticoid Receptor (NR3C1) and the Retinoic Acid Receptor Alpha (RARA). PPARA were Present in both the assembly of risk factor rules generated by less stringent criteria and NR data and the assembly of risk factor rules generated by less stringent criteria and a combination of NRs and co-factors. In addition the PPARA were also found in the assembly of risk factor rules generated by stringent criteria and NR and less stringent criteria and a combination of NRs and co-factors data but PPARA was then found to be both Present and Absent. However, the gene was mostly Present. PPARG existed in the assembly of risk factor rules generated by stringent criteria and NRs and in the assembly of risk factor rules generated by less stringent criteria and a combination of NRs and co-factors. PPARG was Absent in all rules. The MSC gene existed in the assembly of risk factor rules generated by stringent criteria and a combination of NRs and risk factors where the gene was Present. MSC also existed in the assembly of risk factor rules generated by less stringent criteria and a combination of NRs and co-factors but the gene existed both as Present and Absent at equal quantities. The Glucocorticoid Receptor (NR3C1)

existed as absent in the rules constructed from both stringent and less criteria and NRs and RARA_4 existed as Absent in the rules generated from less stringent criteria and NRs and as both Present and Absent in the rules generated by stringent criteria and NR data. The gene however existed mostly as Absent. Additional genes that existed in the assembly of rules for the risk factors that also existed in the Knowledge Bank, were MEF2A that existed in the assembly of rules generated for the less stringent criteria and a combination of NRs and co-factors. The gene existed equally as Present and Absent. Finally the gene CBFA2T1_1 existed in the assembly of genes that was made up from stringent criteria and only co-factors. The gene existed mostly as Absent but also as Present in smaller quantities.

5.5.4 Pathways

The SQL-database was queried in order to investigate which pathways the genes in the rules belonged to and two interesting pathways were used for this purpose. The first pathway is named "Basic mechanisms of action of PPARa, PPARb(d) and PPARg and effects on gene expression". The genes included in the pathways are PPARA, PPARD, PPARG, RXRA and RXRG. After querying the SQL-database, it was found that PPARA and PPARG existed in the assembly rules generated from the trees in Weka-3-2 and also in the pathway. The other interesting pathway is named "Visceral Fat Deposits and the metabolic syndrome" and the included genes in the pathway are: NR3C1, RXRA, RXRB and PPARG. The genes belonging to this pathway and that also are included in the assembly of rules are NR3C1 and PPARG.

5.6 Results from data analysis on all risk factors except for diabetes

5.6.1 Investigation of the distribution of gender in the data sets

The investigation regarding the distribution of gender in the donor group representing all risk factor groups except for diabetes showed that there was an overrepresentation of males in the group. 71 percent of the donors in the group were males as can be seen in table 50. The corresponding negative groups also showed an overrepresentation of males, see table 50.

RISK FACTOR GROUP	MALE	FEMALE
All risk factors except for diabetes mellitus type II	5 (71%)	2 (29%)
Negative kidney stringent	27 (73%)	10 (27%)
Negative kidney less stringent	86 (57%)	64 (43%)

Table 50. The distribution of gender in the donor group representing all risk factors except for the diabetes group and the negative control groups. The table shows both the number of donors belonging to each donor group and the corresponding percentage. The percentage of males and females is the same both for the stringent donor group and the less stringent donor group representing all risk factors except for diabetes. The negative kidney stringent group represents the negative group made up with stringent criteria while negative kidney less stringent group represents the negative group made up with less stringent criteria.

5.6.2 Results from the data analysis with Weka-3-2

The results from the data analysis for the gene expression data chosen for the donor groups representing all risk factors except for diabetes mellitus type II were collected in similar tables as the results from the individual risk factor classifications. The results are shown in table 51-53 for the donor group made up with stringent criteria and in tables 54-56 for the donor group made up with less stringent criteria. Table 51 shows the results from the data processing performed on the gene expression data chosen for the donor group generated from stringent criteria and NR data. The cross-validation of the ZeroR learning scheme are 87.0 percent and no single tree exist that have a higher cross-validation compared with ZeroR. Twelve trees exist that have a cross-validation that is equal to the cross-validation of ZeroR. The corresponding percentage correctly classified instances on the training data was also the same as ZeroR. The tree with the highest percentage correctly classified instances on the training set (i.e. 100 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 85.2 percent). Table 52 shows the results from the data processing of the gene expression data chosen for the stringent donor group and a combination of NRs and co-factors. The ZeroR cross-validation is 87.0 percent and no trees exist that have a higher percent correctly classified instances with cross-validation compared with ZeroR cross-validation. Eight trees exist that have an equal value of cross-validation compared with ZeroR. The corresponding percentage of correctly classified instances on the training set was also the same as ZeroR. The tree with the highest percentage correctly classified instances (i.e. 98.1 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 74.0 percent). In table 53 the results from the data processing of the gene expression data chosen for the donor group generated with stringent criteria and co-factor data are shown. The ZeroR learning scheme has a cross-validation of 87.0 percent correctly classified instances and no single trees exist that have a higher value compared with ZeroR. Eight trees exist that have a cross-validation that is equal to the ZeroR cross-validation. The corresponding percentage of correctly classified instances on the training set was also the same as ZeroR. The tree with the highest percentage correctly classified instances (i.e. 100 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 70.4 percent).

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	85.2	12	23
0.75	1	100.0	85.2	12	23
0.5	1	100.0	83.3	12	23
0.25	1	87.0	85.2	1	1
0.999	2	94.4	75.9	7	13
0.75	2	94.4	75.9	6	11
0.5	2	94.4	81.5	6	11
0.25	2	87.0	85.2	1	1
0.999	3	90.7	79.6	7	13
0.75	3	87.0	81.5	1	1
0.5	3	87.0	81.5	1	1
0.25	3	87.0	87.0	1	1
0.999	5	87.0	85.2	1	1
0.75	5	87.0	87.0	1	1
0.5	5	87.0	87.0	1	1
0.25	5	87.0	87.0	1	1
0.999	10	87.0	87.0	1	1
0.75	10	87.0	87.0	1	1
0.5	10	87.0	87.0	1	1
0.25	10	87.0	87.0	1	1
0.999	20	87.0	87.0	1	1
0.75	20	87.0	87.0	1	1
0.5	20	87.0	87.0	1	1
0.25	20	87.0	87.0	1	1

Table 51. All risk factors except for diabetes – NR results. The table shows the results from the data processing with the gene expression data chosen for the all risk factor donor group with the strict criteria except for diabetes and NR as input. ZeroR: 87.0. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	98.1	74.0	8	15
0.75	1	98.1	74.0	8	15
0.5	1	100.0	74.0	7	13
0.25	1	87.0	75.9	1	1
0.999	2	98.1	77.8	7	13
0.75	2	98.1	77.8	7	13
0.5	2	98.1	79.6	7	13
0.25	2	87.0	79.6	1	1
0.999	3	94.4	81.5	5	9
0.75	3	94.4	81.5	5	9
0.5	3	94.4	83.3	5	9
0.25	3	87.0	83.3	1	1
0.999	5	92.6	79.6	4	7
0.75	5	92.6	79.6	4	7
0.5	5	92.6	81.5	4	7
0.25	5	87.0	85.2	1	1
0.999	10	87.0	87.0	1	1
0.75	10	87.0	87.0	1	1
0.5	10	87.0	87.0	1	1
0.25	10	87.0	87.0	1	1
0.999	20	87.0	87.0	1	1
0.75	20	87.0	87.0	1	1
0.5	20	87.0	87.0	1	1
0.25	20	87.0	87.0	1	1

Table 52. All risk factors except for diabetes – NRs & co-factor results. The table shows the results from the data processing with the gene expression data chosen for the all risk factor donor group with the strict criteria except for diabetes and a combination of NR and co-factors as input. ZeroR: 87.0. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	70.4	7	13
0.75	1	100.0	70.4	7	13
0.5	1	98.1	72.2	6	11
0.25	1	98.1	77.8	6	11
0.999	2	98.1	75.9	6	11
0.75	2	98.1	75.9	6	11
0.5	2	98.1	77.8	6	11
0.25	2	98.1	81.5	6	11
0.999	3	96.3	79.6	6	11
0.75	3	96.3	79.6	6	11
0.5	3	96.3	79.6	6	11
0.25	3	87.0	81.5	1	1
0.999	5	94.4	75.9	5	9
0.75	5	94.4	75.9	5	9
0.5	5	94.4	77.8	5	9
0.25	5	87.0	83.3	1	1
0.999	10	87.0	87.0	1	1
0.75	10	87.0	87.0	1	1
0.5	10	87.0	87.0	1	1
0.25	10	87.0	87.0	1	1
0.999	20	87.0	87.0	1	1
0.75	20	87.0	87.0	1	1
0.5	20	87.0	87.0	1	1
0.25	20	87.0	87.0	1	1

Table 53. All risk factors except for diabetes – co-factors results. The table shows the results from the data processing with the gene expression data chosen for the all risk factor donor group with the strict criteria except for diabetes and co-factors as input. ZeroR: 87.0. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

Two trees were chosen for the runs performed on NR data, co-factor data and the combination of NR and co-factors in order to transform the trees into rules. The selection of trees that would be transformed was performed with the same criteria as when choosing trees to transform for the individual risk factors, see chapter 4.10. The chosen trees were transformed into rules and the results from the transformation are collected in figures in Appendix H.

Tables 54-56 shows the results from the data processing performed on the gene expression data chosen for the donor groups generated from less stringent criteria. In table 54 the results from the classifications of the gene expression data chosen for the donor group created from less stringent criteria and NR-data are shown. The cross-validation of the ZeroR learning scheme is 95.5 percent correctly classified instances and no trees exist that have a higher percentage of cross-validation compared with the ZeroR cross-validation. However, seventeen trees exist that have an equal cross-validation as ZeroR. The corresponding percentage correctly classified instances on the training set were also the same as ZeroR. The tree with the highest percentage correctly classified instances (i.e. 100 percent) resulted in a value of cross-validation that was not higher compared with the ZeroR classifier (i.e. 89.2 percent). Table 55 shows the results from the data processing performed on the gene expression data chosen for the less stringent donor group and a combination of NRs and co-factors. The cross-validation of the ZeroR learning scheme is 95.5 percent and no trees exist that have a cross-validation that is higher than the ZeroR cross-validation. Table 56 shows the results from the processing of the gene expression data chosen from the less stringent donor group and co-factor data. The ZeroR cross-validation is 95.5 percent correctly classified instances and no tree exist that have a cross-validation that is higher compared with the ZeroR cross-validation.

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	89.2	14	27
0.75	1	100.0	89.2	14	27
0.5	1	95.5	91.7	1	1
0.25	1	95.5	95.5	1	1
0.999	2	96.8	91.1	8	15
0.75	2	95.5	92.4	1	1
0.5	2	95.5	95.5	1	1
0.25	2	95.5	95.5	1	1
0.999	3	96.8	94.3	6	11
0.75	3	95.5	95.5	1	1
0.5	3	95.5	95.5	1	1
0.25	3	95.5	95.5	1	1
0.999	5	95.5	94.3	1	1
0.75	5	95.5	95.5	1	1
0.5	5	95.5	95.5	1	1
0.25	5	95.5	95.5	1	1
0.999	10	95.5	95.5	1	1
0.75	10	95.5	95.5	1	1
0.5	10	95.5	95.5	1	1
0.25	10	95.5	95.5	1	1
0.999	20	95.5	95.5	1	1
0.75	20	95.5	95.5	1	1
0.5	20	95.5	95.5	1	1
0.25	20	95.5	95.5	1	1

Table 54. All risk factors except for diabetes – NR results. The table shows the results from the data processing with the gene expression data chosen for the all risk factor donor group with the loose criteria except for diabetes and a combination of NR as input. ZeroR: 95.5. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	86.6	14	27
0.75	1	100.0	86.6	14	27
0.5	1	95.5	89.2	1	1
0.25	1	95.5	95.5	1	1
0.999	2	99.4	87.9	9	17
0.75	2	99.4	87.9	9	17
0.5	2	99.4	89.8	8	15
0.25	2	95.5	95.5	1	1
0.999	3	98.1	91.7	5	9
0.75	3	98.1	92.4	5	9
0.5	3	98.1	94.3	5	9
0.25	3	95.5	95.5	1	1
0.999	5	95.5	94.9	1	1
0.75	5	95.5	95.5	1	1
0.5	5	95.5	95.5	1	1
0.25	5	95.5	95.5	1	1
0.999	10	95.5	95.5	1	1
0.75	10	95.5	95.5	1	1
0.5	10	95.5	95.5	1	1
0.25	10	95.5	95.5	1	1
0.999	20	95.5	95.5	1	1
0.75	20	95.5	95.5	1	1
0.5	20	95.5	95.5	1	1
0.25	20	95.5	95.5	1	1

Table 55. All risk factors except for diabetes – NRs & co-factor results. The table shows the results from the data processing with the gene expression data chosen for the all risk factor donor group with the strict criteria except for diabetes and a combination of NRs and co-factors as input. ZeroR: 95.5. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

CF	MNO	TS	CV	Leaves	Size
0.999	1	100.0	86.0	14	27
0.75	1	100.0	86.0	14	27
0.5	1	95.5	86.6	1	1
0.25	1	95.5	95.5	1	1
0.999	2	99.4	89.2	9	17
0.75	2	99.4	60.9	24	47
0.5	2	96.7	60.9	22	43
0.25	2	96.7	61.6	22	43
0.999	3	94.0	57.0	20	39
0.75	3	98.1	91.1	5	9
0.5	3	98.1	95.5	5	9
0.25	3	95.5	95.5	1	1
0.999	5	95.5	94.3	1	1
0.75	5	95.5	95.5	1	1
0.5	5	95.5	95.5	1	1
0.25	5	95.5	95.5	1	1
0.999	10	95.5	95.5	1	1
0.75	10	95.5	95.5	1	1
0.5	10	95.5	95.5	1	1
0.25	10	95.5	95.5	1	1
0.999	20	95.5	95.5	1	1
0.75	20	95.5	95.5	1	1
0.5	20	95.5	95.5	1	1
0.25	20	95.5	95.5	1	1

Table 56. All risk factors except for diabetes – co-factor results. The table shows the results from the data processing with the gene expression data chosen for the all risk factor donor group with the strict criteria except for diabetes and co-factors as input. ZeroR: 95.5. The CF column corresponds to the ConfidenceFactor parameter, the MNO column corresponds to the Minimum Number of objects parameter. TS is the column for the percentage correctly classified instances on the training set and the CV column is the cross-validation. Finally, in the leaves and size columns the number of leaves and the size of the tree are shown.

The results from the Weka-3-2 runs with the gene expression data chosen for the donor group representing all risk factor except for diabetes mellitus type II, showed that no tree existed that had a cross-validation that was higher compared with the corresponding ZeroR cross-validation. Two trees were chosen for the Weka-3-2 runs on NR data, Co-factor data and a combination of NR and Co-factor data. The trees were transformed into rules that are showed in figures in Appendix H.

6. Discussion and analysis

The aim of this project was to investigate whether nuclear receptors can be used as genetic markers for the metabolic syndrome by classifying each risk factor individually and then try to compare the results from the individual classifiers. The advantage with this approach as opposed to the classification of the metabolic syndrome as a whole is that data can be extracted for the specified risk factors separately. This data can then be stored in a database and analysed with different approaches. The analysis can be performed on individual risk factors, in groups of three risk factors or by using a collection of data for all of the specified risk factors. It was hypothesised that the method would result in that more information could be extracted about genes that possibly act as gene markers both for individual risk factors and for the metabolic syndrome.

6.1 Data analysis with Weka-3-2

The data analysis on the gene expression data for the four specified risk factors performed with Weka-3-2 did in many cases not generate classifiers with high cross-validation. When comparing the cross-validation with the ZeroR learning scheme, the C4.5 algorithm in most cases performed worse than or equal to ZeroR, which could indicate serious overfitting to the training set. The low values of cross-validation compared with the ZeroR algorithm could probably be explained by the fact that the quality of the data was low, meaning that the data included missing values in the measurements of the included risk factors, which probably have had negative impact of the data processing. Furthermore, the data includes errors, which could have been generated by several causes. The problems with bad quality data will be discussed later in chapter 6.2. However, it is also possible that the poor results could be explained by the fact that NRs are not suitable as markers for the specified risk factors. One of the main reasons for changing the gene sets from GPCRs to NRs was that GPCRs did not vary considerably between metabolic syndrome patients and non-metabolic syndrome patients or even between different tissues (Halinen and Norseng (2002). It is however difficult to say how NRs vary compared with GPCRs and by studying the NR gene expression data manually it was impossible to see such trends. It is however probable that NR expression is constant between patients belonging to the specified risk factor(s) and patients that are considered as normal in this study. This could in fact explain the poor results where a large fraction of the trees produced for the specified risk factors do not get a higher cross-validation compared with the ZeroR learning scheme.

One interesting fact that was noticed during the analysis of the data processing results was that no trees generated from the gene expression data chosen for the risk factors created by the less stringent criteria did have a cross-validation that was higher that the cross-validation for the ZeroR learning scheme. These results suggest that the less stringent donor groups cannot be used in searching for gene markers for the metabolic syndrome. In the gene expression data chosen for the risk factors created by

the stringent criteria and NR data, trees were found that had a better value of cross-validation compared with ZeroR only in the obesity group. The corresponding percentage correctly classified instances in the training set was above 80 percent and therefore higher than the cross-validation, which gives a clue about the quality of the tree by investigating how many errors that the classifier have made on the training set. These results could in fact imply that the genes included in the obesity trees that showed a better cross-validation compared with ZeroR, could be used as gene markers for obesity. The data processing on the gene expression data generated for stringent risk factor groups and co-factor data showed that there existed trees in the dyslipidemia and hypertension group that performed better compared with ZeroR. The corresponding percentage correctly classified instances on the training set were higher that the cross-validation. It is however important both to get a high cross-validation and a high percentage correctly classified instances on the training set in order to determine which trees to consider as "good". The cross-validation gives a measurement of the error rate of a learning technique given a fix set of data while the percentage of correctly classified instances gives a measurement of how many errors the classifier have made on the training set. The trees that have both high cross-validation and high percentage correctly classified instances could possibly be explained by the fact that the co-factors in these specific trees can be used as markers for dyslipidemia and hypertension. However, the results generated from the gene expression data chosen for risk factors that were created from stringent criteria and a combination of NRs and co-factors, resulted in trees being found for all risk factors that performed better compared with the corresponding ZeroR cross-validation and with the percentage correctly classified instances on the training set being higher than the corresponding cross-validation, which possibly could be explained by the fact that the combination of NRs and co-factors included genes that could be used as genetic markers for the specific risk factors.

These results are interesting because they support the theory behind co-factors, which are non-protein substances that are required by a protein for biological activity. Since a protein that exists in absence of corresponding co-factor(s) is considered not to function correctly, the results from the data processing on the gene expression data chosen for only NR data can be seen as reasonable, with the fact that no trees existed in the diabetes, dyslipidemia and hypertension groups that had a higher percentage of correctly classified instances compared with ZeroR. However, the poor results from the data processing performed on gene expression data chosen for only co-factors could probably also be explained by the fact that co-factors alone should not be able serve as markers for the individual risk factors. However, the co-factor data used in this study were not completely certain because of the fact that some of the NR genes were included in both the NR table and the co-factor table in the NR-database. The reason was that it is not yet known if these genes could serve as both co-factors and NRs or just as NRs. In addition, it is also important to remember that the NR database, including the NRs and co-factors that were used in the project, is not complete. The NR database used in this thesis consisted of 22 known NRs but according to Francis et al. (2002) a total of 49 distinct members of the nuclear receptor family exist. It is therefore likely that genes exist that not yet have been added to the database but that still should have been included in the data set used in this project. Furthermore, it is possible that these genes could have acted as gene markers and that the results from the data analysis could have shown a higher cross-validation and

percentage correctly classified instances on the training data compared with the results generated in this study.

It was difficult to determine which of the risk factors that seemed most promising in classifying the gene expression data, because the results did not vary extensively. Furthermore, for some risk factors trees that performed better than ZeroR in the data processing of for example only NRs and stringent criteria, the same risk factor did not perform better than ZeroR in the data processing of for example a combination of NR and co-factors and stringent criteria and so on.

By investigating the results from the data analysis with Weka-3-2 it was not obvious that some of the risk factors were more closely related to each other. It was however noticed that the data analysis for diabetes only resulted in that a total of two trees were generated that had a cross-validation that were higher than ZeroR. Both of these trees existed in the results from the data analysis of the gene expression data constructed from a combination of NRs and co-factors. For the three other risk factors, dyslipidemia, hypertension and obesity, a larger number of trees existed that had a cross-validation that was higher compared with ZeroR and that had been generated from both NRs, a combination of NRs and co-factors and only co-factors. This investigation could possibly imply that dyslipidemia, hypertension and obesity are more closely related to each other compared with diabetes. However, it is difficult to draw any certain conclusions about this relation because of the fact that donors were included in the negative group because of missing values and that it is probable that errors exist in the data that constitute the risk factor groups. Furthermore, there are no clear relations that show that some of the risk factor groups have results that are substantially better compared with the other risk factor groups. It is only possible to see a tendency that implies that the dyslipidemia, hypertension and obesity groups have results that are slightly better compared with the results from the diabetes group.

6.2 Effects of data quality issues

6.2.1 Errors in data

Several factors exist that may have had negative impact on the results. The first aspect to this problem involves the fact that the data used to generate the different donor groups was taken from the BioExpress™ database, which contains data that has been inserted into the database manually at a hospital. The possibility exists that the data includes errors in a way that the data that was inserted into the database is incorrect. One real example that was actually found in the obesity group before the extreme values were removed is a donor included in the database that has a BMI of 381.9-kg m^{-2}. This BMI value corresponds to an individual that has a body weight of 700 kilos and a height of 1.83 m, if he/she would exist in reality. One additional example is a donor that has a body weight of 10 kilos when his length is 1.78 meters. Even though these are extreme cases they thus show the problem of incorrect data in the BioExpress™ database.

The result from an additional investigation, which could be worth noticing, is that that the majority of patients in the risk factor groups and in the corresponding normal groups constructed both from stringent and less stringent criteria have some

type of cancer, where the most usual types are malign cancer in the kidney. This could possibly have affected the results from the study, but since both the positive groups and the corresponding normal groups included a majority of patients with this bias towards cancer, it is however difficult to speculate in the possible degree of influence that this bias could have caused.

The distribution of gender in the different risk factor groups showed that there was an over representation of males in all risk factor groups except for the hypertension group created from stringent criteria, where females were over represented. For the less stringent criteria there was also an over representation of males in all groups except for the dyslipidemia group. This over representations could have had effected the results. It was however not possible to perform any detailed investigation in the time frame of this project of how this distribution may have influenced the results.

6.2.2 Incomplete data

Another risk that exists with the data in the BioExpress™ database is that all measurement values and diagnosis for the donors are not included in the database. This could possibly be due to the fact that the donors have come to the hospital to get help for some completely other reason than their major problems or diseases and that the clinical experts at the hospital made a judgement that it was not necessary to take the specific test for that patient, which could help to explain the presence of missing values in the BioExpress™ database. In this study the missing values were a problem mostly for selecting the risk factor groups that would represent the healthy, normal patients that did not have any of the specified risk factors in the metabolic syndrome. The expectation before selecting patients to the different donor groups were likewise that it would be much more difficult to create the donor groups including donors that would be considered as normal, because the BioExpress™ database was hypothesised to contain data from a population that is more "sick" than the actual population in the United states.

The handling of missing values was a critical step in this project. The mostly poor results that were generated in the data processing phase of this study suggest that the method used for the handling of missing values has had negative effect in the construction of the risk factor groups. The method selected, to set the missing values as normal in the negative dyslipidemia and hypertension group, has the disadvantage that it introduces an insecurity factor to the negative groups. It is therefore probable that donors have been falsely included in the risk factor groups The investigation showed that all of the donors in the negative group constructed from stringent criteria were set to normal because of missing values in the measurement of dyslipidemia or hypertension. These results are indeed not very unexpected because the number of donors in the negative dyslipidemia and hypertension group was very low as shown earlier in table 5. In the negative groups for dyslipidemia and hypertension constructed from less stringent criteria a large fraction of the dataset was also represented by donors that had been included in the groups because of missing values. 174 donors from a total of 189 (i.e. 92%) were included in the negative dyslipidemia group because they had missing values in triglycerides or HDL. Similarly, 122 donors from a total of 141 (i.e. 86%) were included in the negative hypertension group. Because of

the fact that the extensive part of the normal groups were represented by donors that had been included because of missing values it is probable that these donors can have had negative impact on the results from the data processing. Since it was not possible to determine to what actual group, if any, these donors should belong to it is possible that the donors have been included in the wrong risk factor group.

6.2.3 Problems related to microarray techniques

The BioExpress™ database includes gene expression data, which is extracted by microarray techniques. Microarray techniques give the ability to group analysis of many hybridisations in order to reveal common patterns of the gene statement. Microarrays are suitable for research because it is a relatively cheap procedure and fast in converting information. The disadvantages of microarray experiments are though that human errors have a large influence on the results. According to Schuchhardt et al. (2000) manufacturing processes and labelling techniques will lead to different performances and detection ranges on microarrays. Schuchhardt et al listed the major sources of fluctuations for cDNA that are to be expected in microarray experiments and they can be summarized as fluctuations in probe, targets and array preparation, in the hybridisation process and overshining effects and effects resulting from image processing. It is hypothesised that some of the fluctuations also should exist on oligonucleotide arrays since cDNA and oligonucleotide microarrays share the same principles. Examples of such sources of fluctuation that should be shared between both cDNA and oligonucleotide arrays are random fluctuations in the geometry of the pin that transports the molecules to the slide, variability in the amount of molecules that are fixed on the slide and unequal distribution of the probes on the slide. Because of the fact that these sources of fluctuations exist, the gene expression data generated from the microarrays and that are used in this project probably includes this kind of fluctuations.

6.2.4 The importance of high quality data

On the basis of the above discussion regarding the major drawbacks and restrictions that were faced throughout this project it is clearly easy to realise the importance of good quality data. The BioExpress™ database consists of data that is both erroneous, incomplete and furthermore it consists of gene expression data, which quality could differ considerably because of the major fluctuations described earlier. Because of this and that the results from the data processing mostly did not get higher values of cross-validation compared with the ZeroR algorithm and that the percentage correctly classified instances on the training data in many cases was not very high, it is concluded that it is extremely important to have strict control of the data and to spend an adequate amount of time on the data cleaning phase of the study. In order to get the high level of control on the data, it would be interesting to perform a similar study on a closed patient group where it is known whether the donors should be included in the specified risk factor group or not. This would reduce the number of possible errors considerably and also completely remove the existing missing values.

6.3 The suitability of NRs as markers

The possibility exists that NRs cannot be used as genetic markers for the metabolic syndrome and that this is the explanation of the mostly poor results of the data analysis with Weka-3-2. However, nuclear receptors, which also have been named metabolic receptors, have recently been associated with metabolic syndrome risk factors in the literature (Francis et al., 2002). These receptors have been associated with for example adipocyte metabolism, insulin sensitivity, liver fat metabolism and reverse cholesterol transport. The fact that the receptors seem to be dysregulated in many common metabolic diseases and that they are designed to respond to small molecules supports the theory of nuclear receptors as being excellent therapeutic targets for the development of new therapeutics for metabolic disorders. However, the mostly poor results generated from the data analysis with Weka-3-2 could suggest that NRs and their corresponding co-factors cannot be used as genetic markers for metabolic syndrome risk factors. In order to investigate the causes of the poor data analysis results, one possible method could be to randomly choose a set of genes from the BioExpress™ database that does not have to be NRs and where everything is unknown about the genes and to perform exactly the same analysis on these genes. If the randomly chosen genes do get a better result compared with the NR genes, this could in fact give possible clues about nuclear receptors and their usefulness as genetic markers for the metabolic syndrome. Even if the data analysis with Weka-3-2 in most cases did not result in classifications of risk factors that had as high cross-validation and corresponding percentage correctly classified instances on the training set as would have been desired in order to consider NRs suitable as genetic markers for metabolic syndrome risk factors, it was still hypothesised that interesting information could be extracted if the results from the different risk factor classifiers were compared.

6.4 Comparison of rules

Even though many of the classifications with the C4.5 algorithm did not have as high cross-validation as desired it was however possible that the combination of the trees could give interesting results. There are several motivations for this claim, for example if any overlapping rules would be found between all risk factor groups, this could in fact mean that the specific rule includes genes that are important for presence of the metabolic syndrome. However, if the trees would have had a higher cross-validation the results from the combinations would have had a more strong ground to fall back on.

However, the results from the comparison of the different rule sets showed that no rule existed that were overlapping between the different risk factors. Not even when comparing only two risk factor rule sets against each other. These results point to that the NR genes included in the rules are not suitable as genetic markers for the metabolic syndrome. The first hypothesis on the comparisons of rules in this project was that if one rule was found to be overlapping in the results for diabetes, dyslipidemia, hypertension and obesity generated by different classifiers, it was hypothesised that the genes in that rule could be used as markers for the metabolic syndrome. The results from the combinations show that the hypothesis can be

falsified, on the basis of the data set used in this project. It is however probable that the quality of the data has had negative impacts on the data processing phase of the thesis, which was discussed earlier in chapter 6.2 and it is still possible that NRs can act as marker genes for the metabolic syndrome. It is also possible that rules exist that are overlapping between risk factor groups in the trees that were not chosen to be transformed into rules. For each risk factor input to Weka-3-2, two trees were chosen out of twenty-four, which imply that a great number of trees existed that possibly could include rules that would give results that could differ compared with the rules generated in this study. In addition, it is also possible that these rules include genes that actually could be used as markers for the metabolic syndrome. It would therefore be interesting to choose a larger number of rules and to perform the same comparison again. This task was not performed in this study because of the time limitations of the project.

The results from the search of overlapping genes that possibly existed in rules in the different rule sets also showed that no genes could be considered to act as marker genes for the metabolic syndrome or for a combination of metabolic syndrome risk factors. The results therefore show that the second hypothesis on the comparisons of rules that was stated could be falsified, on the basis of the current data set. The hypothesis was: if individual genes can be found to be overlapping in the rule sets representing the different risk factors, it is hypothesised that the genes can act as gene markers for the metabolic syndrome and the including risk factors. The results showed that the genes that were considered as overlapping between different risk factor rules classified as Yes, also existed in the corresponding rule set(s) classified as No with exactly the same abscall. It was also possible that the gene existed in several rules for one risk factor but with different abscalls. It is however possible that NRs in fact can act as biological marker genes for the metabolic syndrome and that the negative results could be explained by the poor quality of the data, as discussed earlier in chapter 6.2. It is also possible that the rules generated for each risk factor includes genes that could act as biological marker genes for individual metabolic syndrome risk factors. These genes do in this case not have to be overlapping between the different risk factors, and no genes should be overlapping between the rule sets for the specific risk factor classified as Yes and the rule sets for the specific risk factor that were classified as No. This possibility has unfortunately not been investigated closer in this thesis, because of the limited amount of time that was available for this project.

The rules that were assembled for all the risk factors that were classified as Yes did however result in the findings of a set of genes that possibly could act as marker genes for the metabolic syndrome. The genes that existed in the rules and that had the same abscall in every rule that the gene existed in could possibly be considered to act as a marker for the metabolic syndrome. The genes that existed in the rules but that did have both abscalls in different rules could be important for the metabolic syndrome but it is impossible to determine if the specific gene should be Present or Absent. In cases where the gene exists in a large number in one abscall and only in small numbers in the other abscall, it could possibly be hypothesised that the gene maybe should be in the abscall that occurs at greatest numbers. These results give supporting evidence that the third hypothesis on the assembly of rules could not be clearly falsified, given the data set used in the project. The hypothesis was: if an assembly of rules are created that contains the rules for each risk factor it is hypothesised that genes can be identified, which can act as genetic markers for the

metabolic syndrome. In this thesis three interesting genes were found when the assembly of rules were created, but it is however difficult to decide completely from the investigations performed in this study that the genes actually can act as marker genes for the metabolic syndrome. Additional studies are needed in order to further support or contradict to the third hypothesis stated in the project. The investigation regarding the assembly of rules did not include the analysis of similarities and dissimilarities between the rules in the risk factor rule sets and the rules in the negative groups because of the limited amount of time in this project. It is possible that this investigation may have shown that the same genes could be found also in the negative groups, which then would support the falsification of the hypothesis. Because of the fact that this investigation has not been performed in the thesis it is therefore difficult to clearly falsify or verify the hypothesis.

The main hypothesis in this study was by creating four different classifiers - one for each individual risk factor - and then comparing the results from the classifications, it is possible to identify biological marker genes involved in the metabolic syndrome. The results from the study show that, given this data set, it is difficult to find biological markers involved in the metabolic syndrome. Three interesting genes were found that possibly could act as marker genes, but supporting evidence show that additional experimental studies are needed in order to determine if NRs can act as marker genes for the metabolic syndrome.

However, it is likely that more complex relations exist between the expressions of genes that are actually involved in the metabolic syndrome. The metabolic syndrome is indeed a complicated disorder where many questions are still unsolved and it is not unlikely that a network of genes exists whose interactions are intensely complex.

6.5 Genes resulting from the comparison

In order to decide if any of the genes that were found when creating the assembled rules for all risk factors could be associated with the metabolic syndrome, it was hypothesised that if genes were found in this study that also exist in the Knowledge Bank created by Halinen and Norseng (2002) this could imply that the genes are in fact important for the specific risk factor(s). The Knowledge Bank includes genes that in the scientific literature are associated with metabolic syndrome risk factors and if some of these genes are also found in this project it is possible that they could be used as biological marker genes. The genes that were found in the assembly of rules that also were found in the Knowledge Bank were:

- The Peroxisome Proliferator – Activated Receptor – Alpha (PPARA)
- The Peroxisome Proliferator – Activated Receptor – Gamma (PPARG)
- Musculin (MSC)
- The Glucocorticoid Receptor (NR3C1)
- The Retinoic Acid Receptor Alpha (RARA)

Two additional genes were found in the assembly of rules that were also found in the Knowledge Bank but that existed with both abscalls in the assembly of rules. These genes were:

- The Mads Box Transcription Enhancer Factor 2, Polypeptide A (MEF2A)
- The Core – Binding Factor, Alpha subunit 2, translocated to, 1 (CBFA2T1)

Two of the found genes, PPARA and PPARG, were also found in the BioCarta pathway named "Basic mechanisms of action of PPARa, PPARb(d) and PPARg and effects on gene expression". The complete set of genes included in the pathway are PPARA, PPARD, PPARG, RXRA and RXRG. This pathway was considered as interesting because of its obvious relation to the metabolic syndrome.

Very much information regarding the Peroxisome proliferator – activated receptors (PPARs) is available in the literature and they are known to be key players in the lipid and glucose metabolism and are implicated in metabolic disorders such as the metabolic syndrome risk factors dyslipidemia and diabetes (Zhu & Reddy, 2000). Peroxisome proliferator –activated receptors (PPARs) are members of the nuclear receptor superfamily and they heterodimerize with the retinoic acid receptor (RXR) and bind to the target gene promoter to induce transcriptional activity. Three isotypes of this family of nuclear receptors, namely PPARA, PPARG and PPARD have been identified. The major known role of Peroxisome proliferator activated receptors so far in biological processes is in the storage and catabolism of fatty acids (Tenenbaum, Fisman & Motro, 2003). PPARA play a central role as the receptor for the fibrate drugs, which are widely used in the treatment of coronary artery disease by lowering triglycerides and raising high-density lipoprotein cholesterol levels. The PPARG subtype has been intensively studied because of the thiazoledinediones used in the treatment for diabetes mellitus type II that are ligands for PPARG (Kliewer et al, 2001). PPARG is therefore implicated in the pathogenesis of obesity, insulin resistance and diabetes. Both PPARA and PPARG are expressed in kidney.

Increasing interest has been focused on the Glucocorticoid Receptor (NR3C1) and its association to the metabolic syndrome (Rosmond, 2002). Studies in human have suggested a positive association between hypertension, insulin resistance and obesity with the Glucocorticoid Receptor gene. NR3C1 existed only as Absent in the assembly of rules that were constructed from both stringent and less stringent criteria and NRs. NR3C1 and PPARG also existed in the pathway named "Visceral Fat Deposits and the Metabolic Syndrome", which could further support the theory of NR3C1 being associated with the metabolic syndrome.

The fact that the Peroxisome proliferator – activated receptor alpha and the Peroxisome proliferator – activated receptor gamma were found in the rules generated by creating a assembly of rules classified as Yes and that they also were found in the Knowledge Bank further confer the suggested association between PPARA and PPARG and the metabolic syndrome. PPARA were found to be Present in the assembly of rules created from the stringent criteria and NR data and in the assembly of rules created from the less stringent criteria and a combination of NRs and co-factors.

However, PPARA were found to be both Present and Absent in the assembly of rules generated by the stringent criteria and a combination of NRs and co-factors. PPARA was however mostly Present, the gene existed 72 times as Present in the rules and 36 times as Absent. Furthermore, PPARG were only found to be Absent. PPARG were found in the assembly of rules created by less stringent criteria and a combination of NRs and co-factors.

Because of the time restriction in this study it was not possible to conduct further investigations about how the missing values had affected the results. One possible explanation to the ambiguous expression of PPARA in the assembly of rules is that it could be due to the negative effects of missing values. Further research must however be performed in order to support or contradict this hypothesis.

When comparing these results with the results from the data processing on all risk factors except for Diabetes mellitus type II, it became clear that NR3C1 was found in the rules created from stringent criteria and NRs while PPARG was found in the rules created from less stringent criteria and NRs that also existed in the assembly of rules. However, PPARG was Present in these rules, which could be thought as ambiguous compared with the results previously discussed for the assembly of rules. NR3C1 was Absent both in the rules generated from all risk factors except for diabetes and the assembly of rules. None of the other genes that were found in the assembly of rules and in the Knowledge Bank were also found in the rules generated from all risk factors except for diabetes mellitus type II. It would have been more interesting to compare the results if a data analysis could have been performed on all risk factors. In the method used in this study this was not possible because the samples in kidney for all risk factors were too few and because of this it is not as straightforward to interpret the comparison between the two results. Furthermore, it is also unfortunate that the number of samples in the other tissues except for kidney was so low, that it was not possible to go on with the studies on tissues that were most interesting such as adipose tissue or liver. The possibility exists that important information has been missed when these tissues were removed because of the fact that many of the interesting nuclear receptors are known to be expressed in these tissues. However, kidney is also a possible interesting tissue in context of the metabolic syndrome and it is important to remember that the molecular background on the syndrome is not known so far.

The additional genes that were found in the assembly of rules and in the Knowledge Bank were associated with any of the included risk factors in the metabolic syndrome. These genes were: Musculin (MSC), Retinoic Acid Receptor Alpha (RARA) and the Mads Box Transcription Enhancer Factor 2, Polypeptide A (MEF2A), which were associated with insulin resistance while the Core Binding Factor, Alpha Subunit 2, Translocated to, 1 (CBFA2T1) instead was associated with obesity according to the Knowledge Bank. No association between the previously mentioned genes and the metabolic syndrome were found when searching PubMed. The remaining genes that were also found in the assembly of rules but not in the Knowledge Bank were investigated by conducting a literature search to investigate if there was any clear connection between the given genes and metabolic syndrome risk factors. No such connection was however found. It was nonetheless not possible to conduct a comprehensive literature review within the time frame of this thesis as actually would have been required.

6.6 Selection of probe sets

One possible factor that could give information about the significance of the found genes comes from the fact that different probe sets were used for the same gene, which resulted in that one specific gene included in the rules generated from the trees could occur more than once in the specific rule, but the gene would be given different names. This made the interpretation of the results much more difficult, but since it was not possible to determine which probe set that should be considered as best, this approach was necessary. To select the best probe sets the ones with fragment warning were removed and the resulting probe sets were analysed manually with the tool E-lab. In addition to this investigation, it could have been possible to also examine the probe set names, which could have given clues about the quality of the probe sets. In this project, all probe sets with names extended with "_at", "_s_at" and "_x_at" were used and the only restriction was that the probe sets could not have any fragment warning. It has however upon closer examination become clear that the different probe set names could give clues about which probe set to consider as best. The "_s_at" extensions are given to probe sets when all the probes exactly match multiple transcripts (Affymetrix, 2003). Probe sets with probes among multiple transcripts are common and are due to alternative polyadenylation and alternative splicing. In most cases the "_s_at" probe sets represents transcripts from the same gene, but it can also represent transcripts from homologous genes. One transcript can be represented both by a "_s_at" probe set and a unique "_at" probe set (Affymetrix, 2003). The probe sets with the "_x_at" extension contain some probes that are identical, or highly similar, to unrelated sequences. Data generated from these probes should be interpreted with caution. An investigation of the genes that were found in the assembly of rules and also in the Knowledge Bank showed that PPARA were the only gene that had a probe set with the "_at" extension. PPARG, MSC, NR3C1, RARA_4, MEF2A and CBFA1T1 did all have probe sets with the "_s_at" extension. Since it is possible that the probe sets with the "_s_at " extension represents transcripts from homologous genes these results should be interpreted with care.

7 Conclusions

The aim of this project was to investigate whether nuclear receptors and their corresponding co-factors could be used as genetic markers for the metabolic syndrome, by using four different classifiers – one for each risk factor – and then to compare the results from the individual classifiers. The study shows that a set of NRs were found when comparing the rules generated from the individual risk factors. From the genes that were found in the assembly of rules including all of the metabolic syndrome risk factors three genes seemed most interesting:

- The Peroxisome Proliferator – Activated Receptor – Alpha (PPARA)
- The Peroxisome Proliferator – Activated Receptor – Gamma (PPARG)
- The Glucocorticoid Receptor (NR3C1)

These genes are interesting because they are found both in the assembly of rules and in the Knowledge Bank created by Halinen and Norseng (2002). Furthermore, the genes are also included in pathways that are associated with the metabolic syndrome and a previous association between these NRs and the metabolic syndrome has also been suggested in recent scientific literature (Francis et al., 2002; Rosmond, 2002). It is previously known that PPARs and glucocorticoid receptors are in fact involved in the genetic background of metabolic syndrome risk factors, but it is difficult to determine if PPARA, PPARG and NR3C1 are the hidden links between the included risk factors on the basis of this thesis, because of the poor results in the data analysis. The poor results from the data analysis can probably be explained by the low quality data, which involve both erroneous and incomplete data.

The first hypothesis on the comparisons of rules in this project was that if one rule was found to be overlapping in the results for diabetes, dyslipidemia, hypertension and obesity generated by different classifiers, it was hypothesised that the genes in that rule could be used as markers for the metabolic syndrome. The results from the combinations show that the hypothesis could be falsified, on the basis of the data set used in this project. The results from the investigation of overlapping genes showed that the second hypothesis on the comparisons of rules that was stated could be falsified, on the basis of the current data set. The hypothesis was: if individual genes can be found to be overlapping in the rule sets representing the different risk factors, it is hypothesised that the genes can act as gene markers for the metabolic syndrome and the including risk factors. The results from the assembly of rules showed supporting evidence that the third hypothesis on the assembly of rules could not be clearly falsified, given the data set used in the project. The hypothesis was: if an assembly of rules are created that contains the rules for each risk factor it is hypothesised that genes can be identified, which can act as genetic markers for the metabolic syndrome. In order to draw more certain conclusions on the three hypotheses is was concluded that additional investigations are needed in order to determine the real properties of NRs to act as marker genes for the metabolic syndrome. The main hypothesis in this study was by creating four different classifiers - one for each individual risk factor - and then comparing the results from the classifications, it is possible to identify biological marker genes involved in the

metabolic syndrome. The results from the study show that, given this data set, it is difficult to find biological markers involved in the metabolic syndrome. Three interesting genes were found that possibly could act as marker genes, but supporting evidence show that additional experimental studies are needed in order to determine if NRs can act as marker genes for the metabolic syndrome.

An additional conclusion drawn from this project is that the BioExpress™ database is not suitable for studies concerning metabolic diseases because the low quality of the data will have a great influence on the results and greatly complicate the interpretation of the results. The BioExpress™ database includes data that is both erroneous and incomplete and that has a strong bias towards cancer. Furthermore, the database also includes gene expression data, which quality could differ considerably because of major fluctuations in microarray techniques. Because of these drawbacks, the results generated in this study have probably been affected in a negative way. This could for example be reflected by the mostly poor results from the data analysis phase, where a great fraction of the generated trees performed even worse than the ZeroR learning scheme, which was used as a lower threshold for the cross-validation of the decision trees generated by the C4.5 algorithm.

Because of the limitations of data in the database, the occurrence of erroneous and missing data, the disadvantages of microarray techniques and the existence of several probe sets it is recommended that future studies are conducted that could further support or contradict the hypothesis of nuclear receptors acting as biological marker genes for the metabolic syndrome.

7.1 Future work

It is recommended that additional work is performed to further investigate the hypothetical link between nuclear receptors and the metabolic syndrome. Further research regarding the Peroxisome Proliferator Activated Receptors should be performed in order to generate additional results relevant to this study. Analysis similar to this study should be performed on additional tissues such as adipose tissue and liver, which are known to be highly interesting tissues for nuclear receptor expression. Furthermore, other methods could be used in the handling of the missing values in the BioExpress™ database in order to try to compare and evaluate the probable impact of the method used to handle missing values in this study. Larger tissue sample sets should be used for patients with similar clinical information and background. It would be interesting to perform a similar study on a closed patient group where it is known whether the donors should be included in the specified risk factor group or not. This would probably reduce the number of possible errors considerably and also completely remove the problem of missing values. Additional information that could be included in the analysis of metabolic syndrome risk factors are the age and race of the donors included in the different risk factor groups.

Further research regarding searching for metabolic syndrome risk factors should also be conduced by performing the analysis on all of the six risk factors included in the metabolic syndrome. In this study insulin resistance and microalbuminurea could not be taken into account since no diagnostic tests for these risk factors existed in the BioExpress™ database. Insulin resistance has in previous studies been proposed as the link between the metabolic syndrome risk factors. This hypothesis should be investigated further.

One suggestion of a further analysis that could be performed in order to investigate the link between metabolic syndrome risk factors and nuclear receptors is to use an unsupervised method like for example clustering. Clustering of examples from metabolic syndrome risk factors could give additional clues about which genes are associated and which could be used as genetic markers for the syndrome.

References

Affymetrix. (2003) Genechip® arrays. Affymetrix©. Available from: http://www.affymetrix.com/products/arrays/specific/hgu133.affx [Accessed13 May, 2003].

Alberti, K. G & Zimmet, P. Z. (1998) Definition, diagnosis and Classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus: Report of a WHO consultation. *Diabetic Medicine*, 15, 539-553.

Andersson, D. M. (1994) *Dorland's illustrated medical dictionary*. (Edition 28). Philadelphia: W. B. Saunders Company.

Bertone, P. & Gerstein, M. (2001) Integrative Data Mining: The New Direction in Bioinformatics. *IEEE engineering in medicine and biology magazine: the quarterly magazine of the Engineering in Medicine & Biology Society*. 20, 33-40.

BioCarta. (2003). Pathways. Available from http://www.biocarta.com/genes/index.asp [Accessed 22 May, 2003].

Björntorp, P. (1997) Body Fat Distribution, Insulin Resistance, and Metabolic Diseases. *Nutrition*. 13, 795-803.

Campbell N. A., Reece, J. B. & Mitchell, L. G. (1999) *Biology* (fifth edition). Addison Wesley Longman.

Chawla, A., Repa, J. J., Evans, R. M., Mangelsdorf, D., J. (2001) nuclear receptors and Lipid Physiology: Opening the X-files. *Science*. 294, 1866-1870.

Debouck, C. & Goodfellow, P. N. (1999) DNA microarrays in drug discovery and development. *Nature genetics supplement*, 21.

Deogun, J. S., Raghavan, V. V., Sarkar, A. & Sever, H. (1997) Data Mining: Research Trends, Challenges, and Applications. In: T. Y. Lin N. Cercone (eds.), *In Roughs Sets and Data Mining: Analysis of Imprecise data*. (9-45). Kluwer Academic Publishers.

Eliasson, B., Mero, N., Taskinen, M. & Smith, U. (1996) The insulin resistance syndrome and postprandial lipid intolerance in smokers. *Atherosclerosis*, 129, 79-88.

Elomaa, T. 1996. *Tools and Techniques for Decision Tree Learning*. Series of publications A, Report A-1996-2. Helsinki: Department of Computer Science, University of Helsinki.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to knowledge Discovery in Databases. *American Association for Artificial Intelligence*, 37-54.

Francis, G. A., Fayard, E., Picard, F. & Auwerx, J. (2002). nuclear receptors and the Control of Metabolism. *Annual Review of Physiology*, 65, 24.1-24.51.

Grundy, S. M. (1999). Hypertriglyceridemia, Insulin Resistance, and the Metabolic Syndrome. *The American Journal of Cardiology*. 83(9B), 25F-29F.

Grundy, S. M., Abate, N. & Chandalia, M. (2002) Diet Composition and the Metabolic Syndrome: What is the Optimal Fat Intake? *American Journal of Medicine*, 113(9B), 25S-29S.

Halinen, H. & Norseng, E. (2002). *Relating GPCR Expression Patterns to Metabolic Syndrome Risk Factors.* Unpublished master thesis from the University of Chalmers. Göteborg.

Han, J. & Kamber, M. (2001). Data Mining – Concepts and Techniques. Academic Press.

Hansen, B. C. (1999) The Metabolic Syndrome X: A Work in Progress. In: B, C, Hansen, J, Saye & L, P, Wennogle (eds.), *The Metabolic Syndrome. Convergence of Insulin Resistance, Glucose Intolerance, Hypertension, Obesity, and Dyslipidemias – Searching for the Underlying Defects.* (p. 1-24). 19-22 February, 1999, Jacksonville, Florida, USA.

Hellenius, M., Johansson, J., Karlberg, B. E., Landin-Wilhelmssen, K. & Walldius, G. (1991) *Det metabola syndromet.* Södertälje: Lindfors & Andersson.
Informatics. Association of Intelligent Machinery. (to appear)
International Conference on Computational Biology and Genome.

Kliewer, S. A, Xu, H. E.,Lambert. M. H., Willson, T. M. (2001). Peroxisome Proliferator – Activated Receptors: From Genes to Physiology. *Recent progress in Hormone Research.* 56. 239-265.

Laudet, V. & Gronemeyer, H. (2002) *The nuclear receptor factsbook.* London: Academic Press.

Meigs, J. B. (2000). Invited Commentary: Insulin Resistance Syndrome? Syndrome X? Multiple Metabolic Syndrome? A Syndrome at all? Factor Analysis Reveals Patterns in the Fabric of Correlated Metabolic Risk Factors. *American Journal of Epidemiology.* 152, 908-911.

Mitchell, T. M. (1997) *Machine Learning.* McGraw-Hill Science/Engineering/Math.

National Cholesterol Education Program (NCEP). (2001) Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) *JAMA*, 285, 2486-2497.

Okosun, I. E., Liao, Y., Rotimi, C. N., Prewitt, T. E. & Cooper, R. S. (2000) Abdominal Adiposy and Clustering of Multiple Metabolic Syndrome in White, Black and Hispanic Americans. *AEP*, 10, 263-270.

Petersen, K. F. & Shulman, G. I. (2002). Cellular mechanisms of insulin resistance in skeletal muscle. *Journal of the Royal Society of Medicine*, 95, 8-13.

Pruitt, K. D. & Maglott, D. R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research.* 29(1). 137-140.

Qi, C., Zhu. Y., Reddy. J. K. (2000). Peroxisome proliferator – activated receptors, coactivators and downstream targets. *Cell Biochemistry and Biophysics.* 32. 187-204.

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

Rahpeymai, N. (2002). Data Mining with Decision Trees in the Gene Logic Database – A Breast Cancer Study. the University of Skövde, Sweden. HS-IDA-MD-02-207

Rahpeymai, N., Olsson, B. and Andersson, M.L. (2003) Microarray-based diagnosis of breast cancer using decision trees. *Proceedings of the 5th International Conference on Computational Biology and Genome Informatics*

Reaven, G. (2002). Metabolic Syndrome – Pathophysiology and Implications for Management of Cardiovascular Disease. *Circulation,* 106, 286-288.

Rosmond, R. (2002). The Glucocorticoid Receptor Gene and Its Association to Metabolic Syndrome. *Obesity Recearch.*10(10), 1078-1086.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., Herzel, H. (2000). Normalisation strategies for cDNA microarrays. *Oxford University Press.* 28(10), i-v.

Tamames, J., Clark, D., Herrero, J., Dopazo, J., Blaschke, C., Fernandez, J. M., Oliveros, J. C. & Valencia, A. (2002). *Journal of Biotechnology,* 98, 296-283.

Tenenbaum, A., Fisman, E. Z. (2003). *Metabolic Syndrome and type 2 diabetes mellitus: focus on peroxisome proliferator activated receptors (PPAR).* Cardiovascular Diabetology. 2(4). 1-7.

Thurasingham, B. (1998) *Data Mining: Technologies, Techniques, Tools and Trends.* CRC Press.

Wamala, S, P., Lynch, J.,Horsten, M., Mittleman, M. A., Schenk-Gustafsson, K.& Orth-Gomer, K. (1999). Education and the Metabolic Syndrome in Women. *Diabetes Care,* 22(12), 1999-2003.

Witten, I. H. & Frank, E. (1999) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann publishers.

Yip, J. & Trevisan, R. (1999) Microalbuminuria and Insulin Resistance. In: G. Reaven and A. Laws (eds.), *Contemporary Endocrinology: Insulin Resistance* (309-316). Human Press Inc.

Appendix A. The Nuclear Receptor database – NRs

The nuclear receptor database available at AstraZeneca in Mölndal. The table includes gene symbols and probeset_ids for known nuclear receptors.

GENE SYMBOL	PROBESET_ID
AR	205239_at
AR	211110_s_at
ESR1	211234_x_at
ESR1	211233_x_at
ESR1	205225_at
ESR1	211235_s_at
ESR1	215552_s_at
ESR1	215551_at
ESR2	211120_x_at
ESR2	211117_x_at
ESR2	210780_at
ESR2	211118_x_at
ESRRA	1487_at
ESRRA	203193_at
ESRRB	223858_at
ESRRG	207981_s_at
ESRRG	209966_x_at
HNF4A	214851_at
HNF4A	216889_s_at
HNF4A	208429_x_at
HNF4A	230914_at
HNF4A	214832_at
HNF4G	207456_at
NR0B1	206645_s_at
NR0B1	206644_at
NR0B2	206410_at
NR1D1	31637_s_at
NR1D1	204760_s_at
NR1H2	218215_s_at
NR1H3	203920_at
NR1H4	206340_at
NR1I2	207202_s_at
NR1I2	207203_s_at
NR1I3	207007_at
NR2C1	204791_at
NR2C2	206038_s_at

GENE SYMBOL	PROBESET_ID
NR2E1	207443_at
NR2E3	208388_at
NR2E3	208385_at
NR2F1	209506_s_at
NR2F2	209121_x_at
NR2F2	209119_x_at
NR2F6	209262_s_at
NR2F6	213354_s_at
NR2F6	209261_s_at
NR3C1	201866_s_at
NR3C1	211671_s_at
NR3C1	201865_x_at
NR3C2	205259_at
NR4A1	202340_x_at
NR4A2	204622_x_at
NR4A2	216248_s_at
NR4A2	204621_s_at
NR4A3	207978_s_at
NR5A1	210333_at
NR5A2	208343_s_at
NR5A2	208337_s_at
NR6A1	210392_x_at
NR6A1	210391_at
NR6A1	211402_x_at
NR6A1	207742_s_at
PGR	208305_at
PPARA	206870_at
PPARD	208044_s_at
PPARD	37152_at
PPARG	208510_s_at
RARA	211605_s_at
RARA	216300_x_at
RARA	203750_s_at
RARA	203749_s_at
RARB	205080_at
RARB	208530_s_at

RARB	208412_s_at	RXRB	209148_at
RARG	204189_at	RXRG	205954_at
RARG	204188_s_at	THRA	35846_at
RORA	210479_s_at	THRA	204100_at
RORA	210426_x_at	THRB	207044_at
RORB	206443_at	VDR	204253_s_at
RORC	206419_at	VDR	213692_s_at
RXRA	202426_s_at	VDR	204255_s_at
RXRA	202449_s_at	VDR	204254_s_at
RXRB	215099_s_at		

Appendix B. The Nuclear Receptor database – Co-factors

The nuclear receptor database available at AstraZeneca in Mölndal. The table includes gene symbols and probeset_ids for genes known as co-factors to the nuclear receptors.

GENE SYMBOL	PROBESET_ID
ADPRT	208644_at
AES	217729_s_at
AIP	201782_s_at
AIP	201781_s_at
AR	205239_at
AR	211110_s_at
ARNT	218222_x_at
ARNT	210828_s_at
ARNT	231016_s_at
ARNT	218221_at
ARNT2	202986_at
ATF2	205446_s_at
ATF3	202672_s_at
ATF4	200779_at
ATF5	217389_s_at
ATF5	204999_s_at
ATF5	204998_s_at
ATF6	203952_at
ATF7	232732_at
ATF7	206684_s_at
ATF7	228830_s_at
BAG1	211475_s_at
BAG1	202387_at
BATF	205965_at
BCL3	204908_s_at
BCL3	204907_s_at
BCL6	215990_s_at
BCL6	203140_at
BLZF1	203840_at
BLZF1	32088_at
BRCA1	211851_x_at
BRCA1	204531_s_at
BRCA2	208368_s_at
BRDT	206787_at
BRPF1	204481_at

GENE SYMBOL	PROBESET_ID
BTAF1	209430_at
BTG1	200921_s_at
BTG1	200920_s_at
BTG2	201236_s_at
BTG2	201235_s_at
C19orf2	214173_x_at
C19orf2	211563_s_at
C4orf1	202614_at
CALR	214315_x_at
CALR	212952_at
CALR	212953_x_at
CALR	200935_at
CALR	214316_x_at
CART1	206837_at
CBFA2T1	205529_s_at
CBFA2T1	216831_s_at
CBFA2T1	205528_s_at
CBFA2T2	209145_s_at
CBFA2T2	207625_s_at
CBFA2T2	209144_s_at
CBFB	206788_s_at
CBFB	202370_s_at
CBX4	206724_at
CBX5	209715_at
CITED1	207144_s_at
CITED2	227287_at
CITED2	209357_at
CITED2	207980_s_at
CKN1	205162_at
CNOT7	225053_at
CNOT7	218250_s_at
COPS5	201652_at
CREB1	204314_s_at
CREB1	214513_s_at
CREB1	204312_x_at

CREB1	204313_s_at	DYRK1B	217270_s_at
CREBBP	211808_s_at	DYRK1B	204954_s_at
CREBBP	216362_at	E2F1	2028_s_at
CREBBP	202160_at	E2F1	204947_at
CREG	201200_at	E2F6	203957_at
CRKL	206184_at	E4F1	218524_at
CRSP2	202611_s_at	EDF1	209059_s_at
CRSP2	217120_s_at	EDF1	209058_at
CRSP2	202612_s_at	EED	209572_s_at
CRSP2	202610_s_at	EED	210656_at
CRSP3	218846_at	EGR1	201694_s_at
CRSP3	223947_s_at	EGR1	227404_s_at
CRSP3	242706_s_at	EGR1	201693_s_at
CRSP6	221517_s_at	EGR2	205249_at
CRSP6	223115_at	EIF3S2	208756_at
CRSP7	231724_at	ELF2	203822_s_at
CRSP8	51176_at	ELF2	210361_s_at
CRSP8	221598_s_at	ELF3	210827_s_at
CRSP9	204349_at	ELF3	201510_at
CRSP9	204350_s_at	ELF3	208270_s_at
CRY1	209674_at	ELF4	31845_at
CRY2	212695_at	ELF4	203490_at
CSEN	228269_x_at	ELK3	206127_at
CSEN	231774_at	ELK4	205994_at
CTBP1	213980_s_at	ELK4	206919_at
CTBP1	212863_x_at	ENO1	201231_s_at
CTBP1	203392_s_at	ENO1	216554_s_at
CTBP1	213979_s_at	ENO1	217294_s_at
CTCF	202521_at	EP300	213579_s_at
CTNNB1	201533_at	EPAS1	200879_s_at
DAP3	208822_s_at	EPAS1	200878_at
DAXX	201763_s_at	ERF	203643_at
DDIT3	209383_at	ESR1	211234_x_at
DDX17	208151_x_at	ESR1	211233_x_at
DDX17	208719_s_at	ESR1	205225_at
DDX17	213998_s_at	ESR1	211235_s_at
DEDD	202480_s_at	ESR1	215552_s_at
DEDD	215158_s_at	ESR1	215551_at
DEDD	211255_x_at	ESR2	211120_x_at
DR1	209188_x_at	ESR2	211117_x_at
DR1	207654_x_at	ESR2	210780_at
DR1	216652_s_at	ESR2	211118_x_at
DR1	230073_at	EWSR1	210011_s_at
DR1	209187_at	EWSR1	209214_s_at
DRAP1	203258_at	EWSR1	210012_s_at
DTX1	227336_at	EYA1	214608_s_at

EYA2	209692_at	HDAC7A	217937_s_at
FHL2	202949_s_at	HEY1	218839_at
FMR2	216364_s_at	HEY1	44783_s_at
FMR2	206105_at	HEYL	226828_s_at
FMR2	210957_s_at	HEYL	220662_s_at
FOS	209189_at	HIPK2	224066_s_at
FOSL1	204420_at	HIPK2	219028_at
FOXF1	205935_at	HIRA	217427_s_at
FOXF2	206377_at	HMGB1	200679_x_at
FOXH1	207644_at	HMGB1	200680_x_at
FOXO1A	202723_s_at	HMGCS2	204607_at
FOXO1A	228484_s_at	HMGN3	209377_s_at
FOXO1A	202724_s_at	HNF4A	214851_at
FUS	217370_x_at	HNF4A	216889_s_at
FUS	200959_at	HNF4A	208429_x_at
GABPA	210188_at	HNF4A	230914_at
GABPB1	206173_x_at	HNF4A	214832_at
GABPB1	204618_s_at	HOXC6	206858_s_at
GABPB2	206173_x_at	HRMT1L2	206445_s_at
GABPB2	204618_s_at	HSBP1	200942_s_at
GADD45B	209305_s_at	HSF2	209657_s_at
GADD45B	207574_s_at	HSF2	211220_s_at
GADD45B	213560_at	HSPA8	210338_s_at
GADD45B	209304_x_at	HSPA8	208687_x_at
GCN5L2	202182_at	HSPA8	221891_x_at
GLI2	207034_s_at	HSPA8	224187_x_at
GLI2	208057_s_at	HTATIP	206689_x_at
GLI4	243076_x_at	HTATIP	209192_x_at
GLI4	238364_x_at	HTATIP	214258_x_at
GLI4	227023_at	HTATIP2	209448_at
GMEB1	220938_s_at	HTATIP2	229102_at
GPS2	209350_s_at	HTATIP2	207180_s_at
GPS2	226703_at	ID1	208937_s_at
GTF2B	208066_s_at	ID3	207826_s_at
GTF2F1	202355_s_at	ID4	209293_x_at
GTF2F1	202354_s_at	ID4	209291_at
GTF2F1	202356_s_at	ID4	226933_s_at
HCFC1	202473_x_at	IFI35	209417_s_at
HCFC1	231177_at	ING1	209808_x_at
HCFC1	202474_s_at	ING1	210350_x_at
HD	202390_s_at	ING4	218234_at
HD	202389_s_at	ING4	48825_at
HDAC1	201209_at	IRF1	202531_at
HDAC4	204225_at	IRF2	203275_at
HDAC5	202455_at	IRF3	202621_at
HDAC5	229408_at	ITGB3BP	205176_s_at

JUNB	201473_at	MEIS2	216526_x_at
JUP	201015_s_at	MEIS2	208729_x_at
KLF12	214276_at	MEIS2	209140_x_at
KLF12	239019_at	MEN1	202645_s_at
KLF12	229881_at	MHC2TA	205101_at
KLF12	206966_s_at	MHC2TA	211884_s_at
KLF12	227261_at	MN1	205330_at
KLF12	208467_at	MNT	204206_at
KLF12	238940_at	MSC	209928_s_at
KLF12	206965_at	MTF1	205323_s_at
LDB1	203451_at	MTF1	205322_s_at
LDB1	35160_at	MXI1	202364_at
LDB2	206481_s_at	MYCBP	203359_s_at
LHX3	221670_s_at	MYCBP	203360_s_at
LMO1	206718_at	MYCBP	203361_s_at
MAD	206877_at	MYOD1	206657_s_at
MADH2	203076_s_at	NAB1	211139_s_at
MADH2	203077_s_at	NAB1	209272_at
MADH2	203075_at	NAB1	208047_s_at
MADH3	205397_x_at	NAB2	216017_s_at
MADH3	205398_s_at	NAB2	212803_at
MADH4	202527_s_at	NCOA1	209107_x_at
MADH4	202526_at	NCOA1	210249_s_at
MAFK	206750_at	NCOA2	205731_s_at
MAGED2	208682_s_at	NCOA2	205732_s_at
MAGED2	213627_at	NCOA3	211352_s_at
MAML1	202360_at	NCOA3	209060_x_at
MAZ	207824_s_at	NCOA3	207700_s_at
MAZ	229807_s_at	NCOA3	209062_x_at
MBD1	208595_s_at	NCOA4	210774_s_at
MBD1	203353_s_at	NCOA5	234471_s_at
MBD1	226862_at	NCOA5	225145_at
MCM5	216237_s_at	NCOA6	208979_at
MCM5	201755_at	NCOA6IP	238346_s_at
MECP2	202616_s_at	NCOA6IP	236371_s_at
MECP2	202617_s_at	NCOA6IP	219231_at
MECP2	202618_s_at	NCOR1	200857_s_at
MED6	207079_s_at	NCOR2	208888_s_at
MED6	207078_at	NCOR2	215205_x_at
MEF2A	208328_s_at	NCOR2	207760_s_at
MEF2B	205124_at	NCOR2	208889_s_at
MEF2C	207968_s_at	NCOR2	211388_s_at
MEF2C	209199_s_at	NEUROD6	220045_at
MEF2D	203004_s_at	NFATC3	207416_s_at
MEF2D	203003_at	NFATC3	210555_s_at
MEIS2	207480_s_at	NFATC3	225137_at

NFATC4	205897_at
NFE2	209930_s_at
NFE2L1	200759_x_at
NFE2L1	214179_s_at
NFE2L1	200758_s_at
NFIL3	203574_at
NFKB2	209636_at
NFKB2	207535_s_at
NFKBIB	214062_x_at
NFKBIB	214448_x_at
NFKBIB	228388_at
NFYA	204109_s_at
NFYA	215720_s_at
NFYC	211251_x_at
NFYC	211797_s_at
NFYC	202215_s_at
NFYC	202216_x_at
NMI	203964_at
NOTCH1	218902_at
NOTCH2	202445_s_at
NOTCH2	210756_s_at
NOTCH2	212377_s_at
NOTCH2	202443_x_at
NOTCH2	227067_x_at
NR0B2	206410_at
NR1D1	31637_s_at
NR1D1	204760_s_at
NR1H3	203920_at
NR1H4	206340_at
NR1I2	207202_s_at
NR1I2	207203_s_at
NR1I3	207007_at
NR2C2	206038_s_at
NR2F1	209506_s_at
NR2F2	209121_x_at
NR2F2	209119_x_at
NR2F6	209262_s_at
NR2F6	213354_s_at
NR2F6	209261_s_at
NR3C1	201866_s_at
NR3C1	211671_s_at
NR3C1	201865_x_at
NR4A2	204622_x_at
NR4A2	216248_s_at
NR4A2	204621_s_at
NR5A1	210333_at
NRIP1	202599_s_at
NRIP1	202600_s_at
NT5C	219214_s_at
OLIG2	213825_at
OLIG2	213824_at
OPTN	202074_s_at
PAK6	219461_at
PAWR	214090_at
PAWR	204005_s_at
PAWR	214237_x_at
PAX8	207924_x_at
PAX8	214528_s_at
PAX8	207923_x_at
PAX8	207921_x_at
PAX8	209552_at
PAX8	121_at
PCBD	203557_s_at
PCQAP	222175_s_at
PFDN5	210908_s_at
PFDN5	207132_x_at
PIAS1	217864_s_at
PLRG1	225194_at
PMF1	202337_at
PML	206503_x_at
PML	210362_x_at
PML	211013_x_at
PML	211588_s_at
PML	211014_s_at
PML	211012_s_at
PMX1	205991_s_at
PMX2B	207009_at
POU2AF1	205267_at
PPARBP	203497_at
PPARBP	203496_s_at
PPARG	208510_s_at
PPARGC1	219195_at
PQBP1	214527_s_at
PQBP1	207769_s_at
PQBP1	207440_at
PRKACA	216234_s_at
PRKACA	202801_at
PSIP1	210758_at
PSIP1	205961_s_at
PSIP1	209337_at
PSIP2	205961_s_at
PSIP2	209337_at

PSMC3	201267_s_at	SAP18	208740_at
PSMC5	209503_s_at	SAP30	204899_s_at
PTMA	200772_x_at	SAP30	213963_s_at
PTMA	211921_x_at	SAP30	204900_x_at
PTMA	200773_x_at	SF1	208313_s_at
PURA	204021_s_at	SFPQ	226898_s_at
PURA	213806_at	SFPQ	201586_s_at
RAI17	212124_at	SFPQ	201585_s_at
RAI17	233060_at	SIAH2	209339_at
RAN	200750_s_at	SIN3A	225135_at
RAN	200749_at	SKI	213755_s_at
RARA	211605_s_at	SKIL	232379_at
RARA	216300_x_at	SKIL	206675_s_at
RARA	203750_s_at	SLC26A3	206143_at
RARA	203749_s_at	SMARCA2	206544_x_at
RB1	203132_at	SMARCA2	212257_s_at
RB1	211540_s_at	SMARCA2	206542_s_at
RCOR	212612_at	SMARCA2	228926_s_at
RCOR	237041_x_at	SMARCA2	212258_s_at
RDBP	209219_at	SMARCA4	213720_s_at
RELA	209878_s_at	SMARCA4	212520_s_at
RELA	230202_at	SMARCA4	208794_s_at
RELA	201783_s_at	SMARCA4	215714_s_at
REST	204535_s_at	SMARCA4	208793_x_at
RFX1	206321_at	SMARCA4	213719_s_at
RFX1	226786_at	SMARCA4	214728_x_at
RFX5	202963_at	SMARCB1	206532_at
RFX5	202964_s_at	SMARCB1	212167_s_at
RFXANK	202758_s_at	SMARCC1	201074_at
RFXAP	208492_at	SMARCC1	201072_s_at
RFXAP	229431_at	SMARCC1	201073_s_at
RING1	208371_s_at	SMARCC1	201075_s_at
RING1	35685_at	SMARCC2	201321_s_at
RIPK3	228139_at	SMARCD1	209518_at
RNF14	201824_at	SMARCD1	203183_s_at
RNF14	201823_s_at	SMARCD2	201827_at
RNF4	212696_s_at	SMARCD3	204099_at
RXRA	202426_s_at	SMARCD3	231144_at
RXRA	202449_s_at	SMARCE1	211989_at
RXRB	215099_s_at	SMARCF1	212152_x_at
RXRB	209148_at	SMARCF1	210649_s_at
RYBP	201846_s_at	SMARCF1	218917_s_at
RYBP	201844_s_at	SMARCF1	207591_s_at
RYBP	201845_s_at	SMN1	203852_s_at
SAP18	208742_s_at	SOX10	209842_at
SAP18	208741_at	SOX10	209843_s_at

SP100	202864_s_at	TCF4	203753_at
SP100	210985_s_at	TCF7L2	212759_s_at
SP100	210218_s_at	TCF7L2	212762_s_at
SP140	207777_s_at	TCF8	212758_s_at
SP3	227537_s_at	TCF8	208078_s_at
SP3	213168_at	TCFL4	213951_s_at
SP3	229217_at	TCFL4	217910_x_at
SP4	206663_at	TCFL4	231401_s_at
SPANXC	220922_s_at	TCFL4	213708_s_at
SPANXC	220217_x_at	TCFL4	217909_s_at
SPANXC	224032_x_at	TCFL4	210752_s_at
SSX1	206627_s_at	TFAP2A	204654_s_at
SSX1	206626_x_at	TFAP4	205688_at
SUFU	222749_at	TFDP1	204147_s_at
SUFU	224201_s_at	TFDP1	212330_at
SUPT3H	206506_s_at	TFDP2	203589_s_at
SURB7	209363_s_at	TFDP2	203588_s_at
SURB7	209362_at	TFEC	206715_at
SUV39H1	218619_s_at	TGFB1I1	209651_at
SVIL	202566_s_at	TGIF	203313_s_at
SVIL	202565_s_at	TGIF2	218724_s_at
TADA2L	209938_at	TGIF2	216262_s_at
TADA2L	210537_s_at	TGIF2LX	233178_at
TAF1	216711_s_at	THRA	35846_at
TAF11	209358_at	THRA	204100_at
TAF4	208545_x_at	THRB	207044_at
TAF4	213090_s_at	TIEG2	218486_at
TAF4B	216226_at	TIF1	204391_x_at
TAF6	203572_s_at	TIF1	213301_x_at
TAF7	201023_at	TLE1	203220_s_at
TAF9	202168_at	TLE1	227130_s_at
TAF9	203893_at	TLE1	203222_s_at
TBL1X	201867_s_at	TLE1	203221_at
TBL1X	213401_s_at	TLE2	204431_at
TBL1X	201868_s_at	TLE2	40837_at
TBL1X	201869_s_at	TNRC11	214275_at
TBL1X	213400_s_at	TNRC11	211342_x_at
TBL1X	222634_s_at	TNRC11	203506_s_at
TBL2	212685_s_at	TNRC11	216071_x_at
TBP	203135_at	TP53	211300_s_at
TBPL1	208398_s_at	TP53	201746_at
TCERG1	202396_at	TRAF2	204413_at
TCF15	207306_at	TRIM28	200990_at
TCF2	205313_at	TRIM32	203846_at
TCF2	208135_at	TRIM33	210266_s_at
TCF4	222146_s_at	TRIP11	209778_at

TRIP13	204033_at	ZFP95	203731_s_at
TRIP3	212544_at	ZFP95	203730_s_at
TRIP4	203732_at	ZFX	207247_s_at
TRIP6	209129_at	ZFX	207920_x_at
TRIP8	228793_at	ZIC2	223642_at
TRIP8	224933_s_at	ZNF136	206240_s_at
TRIP8	221763_at	ZNF145	205883_at
TRRAP	202642_s_at	ZNF161	202173_s_at
TRRAP	214908_s_at	ZNF214	243457_s_at
TRRAP	220687_at	ZNF214	220497_at
TSG101	201758_at	ZNF226	219603_s_at
TTF2	204407_at	ZNF226	233461_x_at
UBP1	218082_s_at	ZNF226	231717_s_at
USF1	231768_at	ZNF226	224004_at
UTF1	208275_x_at	ZNF230	205791_x_at
WHSC2	203112_s_at	ZNF274	204937_s_at
WHSC2	34225_at	ZNF286	220250_at
WNT6	221608_at	ZNF302	218490_s_at
WNT6	221609_s_at	ZNF302	228393_s_at
WNT6	222086_s_at	ZNF304	207753_at
YAP1	224894_at	ZNF317	225296_at
YAP1	213342_at	ZNF337	214760_at
YAP1	224895_at	ZNF337	37860_at
YWHAH	201020_at	ZNF397	229133_s_at
YY1	201901_s_at	ZNF397	235271_s_at
YY1	200047_s_at	ZNF85	206572_x_at
YY1	213494_s_at		

Appendix C. The distribution of samples in different tissues

The resulting number of samples in the different tissues. The tissues marked in light grey refers to the tissues that according to the NR group at AstraZeneca in Mölndal was interesting to use in this project.

TISSUE	NUMBER OF SAMPLES
ADIPOSE TISSUE	1
ADRENAL GLAND, NOS	5
AMPULLA OF VATER	1
ARTERY, NOS	2
BLADDER, NOS	4
BONES, NOS	4
BRAIN, NOS	14
BREAST, NOS	4
BRONCHUS, NOS	2
CERVIX, NOS	2
CINGULATE GYRUS	2
COLON, NOS	45
DUODENUM, NOS	12
ENDOMETRIUM, NOS	4
ESOPHAGUS, NOS	10
FALLOPIAN TUBE, NOS	3
GALLBLADDER, NOS	9
HEART, NOS	1
INFERIOR FRONTAL GYRUS	1
INFERIOR TEMPORAL GYRUS	1
KIDNEY, NOS	21
LACRIMAL GLAND, NOS	1
LEFT VENTRICLE, NOS	1
LIVER, NOS	17
LUNG, NOS	44
LYMPH NODE, NOS	20
MEDIASTINUM, NOS	2
MESENTERY, NOS	1
MIDDLE FRONTAL GYRUS	1
MUSCLES, NOS	5
MYOMETRIUM, NOS	8
NERVE, NOS	1
OMENTUM, NOS	6
OVARY, NOS	10
PANCREAS, NOS	18
PARATHYROID GLAND, NOS	1
PAROTID GLAND, NOS	3

PROSTATE, NOS	67
RECTUM, NOS	16
RHINAL SULCUS	2
RIGHT VENTRICLE, NOS	1
SALIVARY GLAND, NOS	1
SKIN, NOS	3
SMALL INTESTINE, NOS	8
SOFT TISSUES, NOS	19
SPLEEN, NOS	4
STOMACH, NOS	21
SUPERIOR FRONTAL GYRUS	2
SUPERIOR PARIETAL LOBULE	2
SUPERIOR TEMPORAL GYRUS	2
TENDON AND TENDON SHEATH, NOS	1
TESTIS, NOS	2
THYMUS, NOS	2
THYROID GLAND, NOS	2
TONGUE, NOS	1
TONSIL, NOS	1
WHITE BLOOD CELL, NOS	4
VISUAL CORTEX	2
VULVA, NOS	1

Table C1. The number of samples in the different tissue types in the diabetes group constructed from stringent criteria.

TISSUE	NUMBER OF SAMPLES
ADIPOSE TISSUE	1
ADRENAL GLAND, NOS	2
AMYGDALOID NUCLEUS	1
AORTA, NOS	1
ATRIUM, NOS	1
BLADDER, NOS	2
BREAST, NOS	1
BRONCHUS, NOS	1
CAUDATE NUCLEUS, NOS	1
CERVIX, NOS	2
COLON, NOS	5
CORONARY ARTERY, NOS	1
CORTEX OF FRONTAL LOBE	1
CORTEX OF OCCIPITAL LOBE	1
CORTEX OF PARIETAL LOBE	1
CORTEX OF TEMPORAL LOBE	1
HIPPOCAMPUS	1
KIDNEY, NOS	8
LIVER, NOS	6
LUNG, NOS	7

LYMPH NODE, NOS	1
MEDULLA OBLONGATA, NOS	1
MYOMETRIUM, NOS	1
OMENTUM, NOS	1
OVARY, NOS	5
PANCREAS, NOS	2
PITUITARY GLAND, NOS	1
SALIVARY GLAND, NOS	1
SMALL INTESTINE, NOS	4
SPINAL CORD, NOS	1
SUBSTANTIA NIGRA	1
THALAMUS, NOS	1
THYROID GLAND, NOS	1
TONSIL, NOS	1
TRACHEA, NOS	1
URETHRA, NOS	1
UTERUS, NOS	2
VEIN, NOS	1

Table C2. The number of samples in the different tissue types in the dyslipidemia group constructed from stringent criteria.

TISSUE	NUMBER OF SAMPLES
ADIPOSE TISSUE	5
ADRENAL GLAND, NOS	8
AMPULLA OF VATER	2
AMYGDALOID NUCLEUS	2
APPENDIX, NOS	1
BLADDER, NOS	1
BREAST, NOS	24
CARDIAC MYOCYTE, NOS	1
CARTILAGE, NOS	1
CAUDATE NUCLEUS, NOS	3
CEREBELLAR HEMISPHERE, NOS	3
CERVIX, NOS	8
CINGULATE GYRUS	38
COLON, NOS	49
CORPUS CALLOSUM, NOS	1
CORTEX OF FRONTAL LOBE	2
CORTEX OF OCCIPITAL LOBE	1
CORTEX OF PARIETAL LOBE	2
CORTEX OF TEMPORAL LOBE	1
DUODENUM, NOS	18
ENDOMETRIUM, NOS	3
ENTORHINAL CORTEX	1
ESOPHAGUS, NOS	4
FALLOPIAN TUBE, NOS	6

FRONTAL POLE	20
GALLBLADDER, NOS	4
GLOBUS PALLIDUS, NOS	1
HEPATOCYTE	15
HIPPOCAMPUS	2
HYPOTHALAMUS, NOS	2
INFERIOR FRONTAL GYRUS	18
INFERIOR TEMPORAL GYRUS	19
INSULA, NOS	2
KIDNEY, NOS	31
LEFT ATRIUM, NOS	42
LEFT VENTRICLE, NOS	74
LIVER, NOS	11
LUNG, NOS	41
LYMPH NODE, NOS	4
MIDDLE FRONTAL GYRUS	21
MIDDLE TEMPORAL GYRUS	18
MOTOR CORTEX	1
MUSCLES, NOS	10
MYOMETRIUM, NOS	20
NASAL SINUS, NOS	1
NERVE, NOS	1
NUCLEUS BASALIS OF MEYNERT	2
OMENTUM, NOS	14
ORBITAL GYRI	2
OVARY, NOS	15
PANCREAS, NOS	19
PAROTID GLAND, NOS	1
PERITONEUM, NOS	2
PRECENTRAL GYRUS	2
PROSTATE, NOS	10
PULVINAR, NOS	1
PUTAMEN	1
RAPHE OF MEDULLA OBLONGATA	2
RECTUM, NOS	8
RHINAL SULCUS	23
RIGHT ATRIUM, NOS	48
RIGHT VENTRICLE, NOS	63
SALIVARY GLAND, NOS	1
SEMINAL VESICLE, NOS	1
SEPTAL AREA OF PARATERMINAL BODY OF RHINENCEPHALON	1
SKIN, NOS	4
SMALL INTESTINE, NOS	15
SOFT TISSUES, NOS	2
SPINAL CORD, NOS	1
SPLEEN, NOS	6
STOMACH, NOS	18

SUPERIOR FRONTAL GYRUS	21
SUPERIOR PARIETAL LOBULE	17
SUPERIOR TEMPORAL GYRUS	21
SUPRAMARGINAL GYRUS	1
SYNOVIUM OF JOINT, NOS	1
THYROID GLAND, NOS	25
URETER, NOS	1
UTERUS, NOS	4
VAS DEFERENS, NOS	1
VEIN, NOS	1
WHITE BLOOD CELL, NOS	19
VISUAL CORTEX	23

Table C3. The number of samples in the different tissue types in the hypertension group constructed from stringent criteria.

TISSUE	NUMBER OF SAMPLES
ADIPOSE TISSUE	6
ADRENAL GLAND, NOS	1
AMYGDALOID NUCLEUS	1
APPENDIX, NOS	1
BREAST, NOS	40
CAUDATE NUCLEUS, NOS	3
CEREBELLAR HEMISPHERE, NOS	3
CEREBELLAR VERMIS, NOS	2
CEREBELLUM, NOS	1
CERVIX, NOS	17
CINGULATE GYRUS	5
COLON, NOS	23
CORPUS CALLOSUM, NOS	1
ENDOMETRIUM, NOS	19
ENTORHINAL CORTEX	2
FALLOPIAN TUBE, NOS	11
FRONTAL POLE	4
GLOBUS PALLADIUS, NOS	3
HEPATOCYTE	10
HIPPOCAMPUS	2
HYPOTHALAMUS, NOS	2
INFERIOR TEMPORAL GYRUS	4
INSULA, NOS	1
KIDNEY, NOS	17
LEFT ATRIUM, NOS	17
LEFT VENTRICLE, NOS	24
LOCUS CERULEUS	3
LUNG, NOS	9
LYMPH NODE, NOS	1

LYMPHOCYTE, NOS	17
MIDDLE TEMPORAL GYRUS	3
MONOCYTE, NOS	6
MOTOR CORTEX	4
MUSCLES, NOS	2
MYOMETRIUM, NOS	39
NUCLEUS BASALIS OF MEYNERT	2
OMENTUM, NOS	5
ORBITAL GYRI	4
OVARY, NOS	48
PENIS, NOS	1
PRECENTRAL GYRUS	2
PROSTATE, NOS	11
PULVINAR, NOS	2
PUTAMEN	2
RAPHE OF MEDULLA OBLONGATA	3
RECTUM, NOS	4
RIGHT ATRIUM, NOS	17
RIGHT VENTRICLE, NOS	22
SEPTAL AREA OF PARATERMINAL BODY OF RHINENCEPHALON	3
SKIN, NOS	17
SMALL INTESTINE, NOS	8
SOFT TISSUES, NOS	1
SPLEEN, NOS	6
SUBSTANTIA NIGRA	2
SUBTHALAMIC NUCLEUS	1
SUPERIOR PARIETAL LOBULE	2
SUPERIOR TEMPORAL GYRUS	3
SUPRAMARGINAL GYRUS	2
SYNOVIUM OF JOINT, NOS	2
TEMPORAL GYRUS, NOS	3
TEMPORAL POLE	1
THALAMUS, NOS	3
URETER, NOS	6
UTERUS, NOS	1
VEIN, NOS	1
WHITE BLOOD CELL, NOS	21
WHITE MATTER OF OCCIPITAL LOBE	2
VISUAL CORTEX	2

Table C4. The number of samples in the different tissue types in the obesity group constructed from stringent criteria.

TISSUE	NUMBER OF SAMPLES
ADIPOSE TISSUE	5
ADRENAL GLAND, NOS	3
APPENDIX, NOS	2
BREAST, NOS	54
BRONCHUS, NOS	2
CEREBELLUM, NOS	1
CERVIX, NOS	32
COLON, NOS	84
CORTEX OF FRONTAL LOBE	1
CORTEX OF PARIETAL LOBE	1
CORTEX OF TEMPORAL LOBE	1
DUODENUM, NOS	3
ENDOMETRIUM, NOS	12
ESOPHAGUS, NOS	7
FALLOPIAN TUBE, NOS	11
GALLBLADDER, NOS	3
HIPPOCAMPUS	1
KIDNEY, NOS	37
LEFT ATRIUM, NOS	56
LEFT VENTRICLE, NOS	70
LIVER, NOS	40
LUNG, NOS	97
LYMPH NODE, NOS	30
MUSCLES, NOS	15
MYOMETRIUM, NOS	54
OMENTUM, NOS	11
OVARY, NOS	54
PANCREAS, NOS	8
PAROTID GLAND, NOS	6
PERITONEUM, NOS	1
PROSTATE, NOS	4
RECTUM, NOS	26
RIGHT ATRIUM, NOS	69
RIGHT VENTRICLE, NOS	58
SALIVARY GLAND, NOS	5
SEMINAL VESICLE, NOS	2
SKIN, NOS	14
SMALL INTESTINE, NOS	25
SMOOTH MUSCLE, NOS	1
SOFT TISSUES, NOS	4
SPLEEN, NOS	14
STOMACH, NOS	26
THYMUS, NOS	6
THYROID GLAND, NOS	37
UTERUS, NOS	29

VAGINA, NOS	1
WHITE BLOOD CELL, NOS	14
VULVA, NOS	2

Table C5. The number of samples in the different tissues in the negative group constructed from stringent criteria.

TISSUE	NUMBER OF SAMPLES
ADIPOSE TISSUE	9
ADRENAL GLAND, NOS	11
AMPULLA OF VATER	2
AMYGDALOID NUCLEUS	3
APPENDIX, NOS	1
ARTERY, NOS	7
BLADDER, NOS	12
BLOOD VESSEL, NOS	1
BONE MARROW, NOS	1
BONES, NOS	9
BRAIN, NOS	21
BREAST, NOS	17
BRONCHUS, NOS	3
CAUDATE NUCLEUS, NOS	4
CEREBELLAR HEMISPHERE, NOS	2
CEREBELLAR VERMIS, NOS	3
CEREBELLUM, NOS	2
CERVIX, NOS	17
CINGULATE GYRUS	15
COLON, NOS	117
CORPUS CALLOSUM, NOS	3
CORTEX OF FRONTAL LOBE	1
DUODENUM, NOS	35
ENDOMETRIUM, NOS	15
ENTORHINAL CORTEX	3
ESOPHAGUS, NOS	18
FALLOPIAN TUBE, NOS	12
FRONTAL POLE	7
GALLBLADDER, NOS	14
GLOBUS PALLIDUS, NOS	2
HEART, NOS	1
HIPPOCAMPUS	4
HYPOTHALAMUS, NOS	2
INFERIOR FRONTAL GYRUS	5
INFERIOR TEMPORAL GYRUS	7
INSULA, NOS	3
KIDNEY, NOS	57

LACRIMAL GLAND, NOS	1
LEFT VENTRICLE, NOS	2
LIVER, NOS	52
LOCUS CERULEUS	1
LUNG, NOS	107
LYMPH NODE, NOS	34
LYMPHATIC SYSTEM, NOS	1
MEDIASTINUM, NOS	2
MEDULLA OBLONGATA, NOS	1
MESENTERY, NOS	1
MIDDLE FRONTAL GYRUS	5
MIDDLE TEMPORAL GYRUS	9
MOTOR CORTEX	4
MUSCLES, NOS	18
MYOMETRIUM, NOS	26
NERVE, NOS	1
NUCLEUS BASALIS OF MEYNERT	2
OMENTUM, NOS	22
ORBITAL GYRI	4
OVARY, NOS	56
PANCREAS, NOS	53
PARATHYROID GLAND, NOS	1
PAROTID GLAND, NOS	4
PLEURA, NOS	1
PRECENTRAL GYRUS	1
PROSTATE, NOS	99
PULVINAR, NOS	1
PUTAMEN	2
RAPHE OF MEDULLA OBLONGATA	2
RECTUM, NOS	27
RHINAL SULCUS	7
RIGHT VENTRICLE, NOS	1
SALIVARY GLAND, NOS	1
SEPTAL AREA OF PARATERMINAL BODY OF RHINENCEPHALON	4
SKIN, NOS	19
SMALL INTESTINE, NOS	34
SOFT TISSUES, NOS	31
SPINAL CORD, NOS	2
SPLEEN, NOS	21
STOMACH, NOS	51
SUBSTANTIA NIGRA	3
SUBTHALAMIC NUCLEUS	1
SUPERIOR FRONTAL GYRUS	8
SUPERIOR PARIETAL LOBULE	6
SUPERIOR TEMPORAL GYRUS	10
SUPRAMARGINAL GYRUS	2
TEMPORAL GYRUS, NOS	3

TEMPORAL POLE	1
TENDON AND TENDON SHEATH, NOS	1
TESTIS, NOS	3
THALAMUS, NOS	3
THYMUS, NOS	7
THYROID GLAND, NOS	14
TONGUE, NOS	2
TONSIL, NOS	2
UTERUS, NOS	9
VEIN, NOS	1
WHITE BLOOD CELL, NOS	5
WHITE MATTER OF OCCIPITAL LOBE	1
VISUAL CORTEX	6
VULVA, NOS	3

Table C6. The number of samples in the different tissue types in the diabetes group constructed from less stringent criteria.

TISSUE	NUMBER OF SAMPLES
ADIPOSE TISSUE	7
ADRENAL GLAND, NOS	4
AMYGDALOID NUCLEUS	1
AORTA, NOS	1
ATRIUM, NOS	1
BLADDER, NOS	2
BREAST, NOS	1
BRONCHUS, NOS	1
CAUDATE NUCLEUS, NOS	1
CERVIX, NOS	7
COLON, NOS	36
CORONARY ARTERY, NOS	1
CORTEX OF FRONTAL LOBE	1
CORTEX OF OCCIPITAL LOBE	1
CORTEX OF PARIETAL LOBE	1
CORTEX OF TEMPORAL LOBE	1
DUODENUM, NOS	10
ENDOMETRIUM, NOS	1
FALLOPIAN TUBE, NOS	6
GALLBLADDER, NOS	3
HIPPOCAMPUS	1
KIDNEY, NOS	32
LIVER, NOS	16
LUNG, NOS	43
LYMPH NODE, NOS	2
MEDULLA OBLONGATA, NOS	1
MYOMETRIUM, NOS	12
OMENTUM, NOS	4
OVARY, NOS	23
PANCREAS, NOS	12
PITUITARY GLAND, NOS	1

PLEURA, NOS	1
SALIVARY GLAND, NOS	1
SKIN, NOS	1
SMALL INTESTINE, NOS	11
SOFT TISSUES, NOS	1
SPINAL CORD, NOS	1
SPLEEN, NOS	5
STOMACH, NOS	3
SUBSTANTIA NIGRA	1
THALAMUS, NOS	1
THYROID GLAND, NOS	1
TONSIL, NOS	1
TRACHEA, NOS	1
URETHRA, NOS	1
UTERUS, NOS	5
VEIN, NOS	5

Table C7. The number of samples in the different tissue types in the dyslipidemia group constructed from less stringent criteria.

TISSUE	NUMBER OF SAMPLES
ADIPOSE TISSUE	19
ADRENAL GLAND, NOS	16
AMPULLA OF VATER	3
AMYGDALOID NUCLEUS	5
APPENDIX, NOS	2
ARTERY, NOS	3
BLADDER, NOS	4
BLOOD VESSEL, NOS	1
BONE MARROW, NOS	1
BREAST, NOS	52
CARDIAC MYOCYTE, NOS	1
CARTILAGE, NOS	1
CAUDATE NUCLEUS, NOS	8
CEREBELLAR HEMISPHERE, NOS	7
CEREBELLAR VERMIS, NOS	5
CEREBELLUM, NOS	1
CERVIX, NOS	39
CINGULATE GYRUS	53
COLON, NOS	141
CORPUS CALLOSUM, NOS	6
CORTEX OF FRONTAL LOBE	1
DUODENUM, NOS	41
ENDOMETRIUM, NOS	11
ENTORHINAL CORTEX	5
ESOPHAGUS, NOS	12
FALLOPIAN TUBE, NOS	21
FRONTAL POLE	30
GALLBLADDER, NOS	6
GLOBUS PALLIDUS, NOS	5
HEPATOCYTE	15
HIPPOCAMPUS	6

HYPOTHALAMUS, NOS	6
INFERIOR FRONTAL GYRUS	22
INFERIOR TEMPORAL GYRUS	27
INSULA, NOS	7
KIDNEY, NOS	100
LEFT ATRIUM, NOS	64
LEFT VENTRICLE, NOS	106
LIVER, NOS	43
LOCUS CERULEUS	1
LUNG, NOS	120
LYMPH NODE, NOS	9
MEDULLA OBLONGATA, NOS	1
MIDDLE FRONTAL GYRUS	25
MIDDLE TEMPORAL GYRUS	30
MOTOR CORTEX	7
MUSCLES, NOS	23
MYOMETRIUM, NOS	62
NASAL SINUS, NOS	1
NERVE, NOS	1
NUCLEUS BASALIS OF MEYNERT	4
OMENTUM, NOS	29
ORBITAL GYRI	7
OVARY, NOS	83
PANCREAS, NOS	50
PAROTID GLAND, NOS	1
PERITONEUM, NOS	2
PLEURA, NOS	1
PRECENTRAL GYRUS	3
PROSTATE, NOS	26
PULVINAR, NOS	3
PUTAMEN	5
RAPHE OF MEDULLA OBLONGATA	6
RECTUM, NOS	16
RHINAL SULCUS	27
RIGHT ATRIUM, NOS	72
RIGHT VENTRICLE, NOS	87
SALIVARY GLAND, NOS	1
SEMINAL VESICLE, NOS	1
SEPTAL AREA OF PARATERMINAL BODY OF RHINENCEPHALON	6
SKIN, NOS	29
SMALL INTESTINE, NOS	43
SOFT TISSUES, NOS	8
SPINAL CORD, NOS	4
SPLEEN, NOS	22
STOMACH, NOS	57
SUBSTANTIA NIGRA	4
SUBTHALAMIC NUCLEUS	1
SUPERIOR FRONTAL GYRUS	25
SUPERIOR PARIETAL LOBULE	23
SUPERIOR TEMPORAL GYRUS	29
SUPRAMARGINAL GYRUS	4
SYNOVIUM OF JOINT, NOS	1
TEMPORAL GYRUS, NOS	5

THALAMUS, NOS	1
THYROID GLAND, NOS	50
URETER, NOS	1
UTERUS, NOS	20
VAS DEFERENS, NOS	1
VEIN, NOS	6
WHITE BLOOD CELL, NOS	25
WHITE MATTER OF OCCIPITAL LOBE	3
VISUAL CORTEX	29

Table C8. The number of samples in the different tissue types in the hypertension group constructed from less stringent criteria.

TISSUE	NUMBER OF SAMPLES
ADIPOSE TISSUE	20
ADRENAL GLAND, NOS	1
AMYGDALOID NUCLEUS	1
APPENDIX, NOS	1
ARTERY, NOS	1
BLADDER, NOS	3
BREAST, NOS	61
CAUDATE NUCLEUS, NOS	5
CEREBELLAR HEMISPHERE, NOS	4
CEREBELLAR VERMIS, NOS	4
CEREBELLUM, NOS	2
CERVIX, NOS	43
CINGULATE GYRUS	7
COLON, NOS	79
CORPUS CALLOSUM, NOS	3
CORTEX OF FRONTAL LOBE	1
ENDOMETRIUM, NOS	34
ENTORHINAL CORTEX	4
FALLOPIAN TUBE, NOS	27
FRONTAL POLE	6
GLOBUS PALLIDUS, NOS	4
HEPATOCYTE	10
HIPPOCAMPUS	3
HYPOTHALAMUS, NOS	4
INFERIOR TEMPORAL GYRUS	6
INSULA, NOS	3
KIDNEY, NOS	48
LEFT ATRIUM, NOS	33
LEFT VENTRICLE, NOS	46
LOCUS CERULEUS	3
LUNG, NOS	62
LYMPH NODE, NOS	3
LYMPHOCYTE, NOS	17
MEDULLA OBLONGATA, NOS	1
MIDDLE TEMPORAL GYRUS	5
MONOCYTE, NOS	6
MOTOR CORTEX	7
MUSCLES, NOS	12
MYOMETRIUM, NOS	80
NUCLEUS BASALIS OF MEYNERT	2

OMENTUM, NOS	13
ORBITAL GYRI	6
OVARY, NOS	116
PANCREAS, NOS	3
PENIS, NOS	1
PRECENTRAL GYRUS	3
PROSTATE, NOS	36
PULVINAR, NOS	3
PUTAMEN	3
RAPHE OF MEDULLA OBLONGATA	4
RECTUM, NOS	11
RIGHT ATRIUM, NOS	36
RIGHT VENTRICLE, NOS	37
SEPTAL AREA OF PARATERMINAL BODY OF RHINENCEPHALON	5
SKIN, NOS	38
SMALL INTESTINE, NOS	21
SOFT TISSUES, NOS	7
SPLEEN, NOS	17
STOMACH, NOS	2
SUBSTANTIA NIGRA	4
SUBTHALAMIC NUCLEUS	1
SUPERIOR PARIETAL LOBULE	3
SUPERIOR TEMPORAL GYRUS	5
SUPRAMARGINAL GYRUS	2
SYNOVIUM OF JOINT, NOS	2
TEMPORAL GYRUS, NOS	5
TEMPORAL POLE	1
THALAMUS, NOS	4
URETER, NOS	1
UTERUS, NOS	21
VEIN, NOS	6
WHITE BLOOD CELL, NOS	25
WHITE MATTER OF OCCIPITAL LOBE	3
VISUAL CORTEX	3

Table C9. The number of samples in the different tissue types in the obesity group constructed from less stringent criteria.

TISSUE	NUMBER OF SAMPLES
ADIPOSE TISSUE	2
CERVIX, NOS	1
COLON, NOS	4
FALLOPIAN TUBE, NOS	1
KIDNEY, NOS	2
LUNG, NOS	3
MYOMETRIUM, NOS	4
OVARY, NOS	6
PANCREAS, NOS	1
SKIN, NOS	1
SMALL INTESTINE, NOS	2
UTERUS, NOS	2

Table C10. The number of samples in the different tissue types in the donor groups constructed from all risk factor groups.

TISSUE	NUMBER OF SAMPLES
ADIPOSE TISSUE	4
CERVIX, NOS	3
COLON, NOS	9
ENDOMETRIUM, NOS	1
FALLOPIAN TUBE, NOS	3
KIDNEY, NOS	7
LUNG, NOS	13
MYOMETRIUM, NOS	7
OMENTUM, NOS	1
OVARY, NOS	8
PANCREAS, NOS	1
SKIN, NOS	1
SMALL INTESTINE, NOS	3
UTERUS, NOS	2
VEIN, NOS	4

Table C11. The number of samples in the different tissue types in the donor group constructed from all risk factors except for diabetes.

Appendix D. Donors with missing values

The 174 donors set to normal because they have missing values in the measurements for dyslipidemia. The donors are included in the negative dyslipidemia group constructed from less stringent criteria in kidney.

DONOR_ID	DONOR_ID	DONOR_ID	DONOR_ID
100076	107857	114296	125622
100077	107859	114302	125624
100078	107935	114308	125861
100086	107943	114343	125863
100096	108009	115748	127234
100109	108393	115750	129754
100194	108563	115751	129884
100195	108565	115752	130805
100253	108571	115753	130821
100254	108573	115755	131377
100279	108575	115756	131379
100797	108576	115757	131555
104864	108632	115758	133765
104865	108635	115759	133777
105089	109045	115761	135621
105172	109050	115762	135675
105323	109063	115763	136488
105578	109093	115764	136489
105579	109105	115766	137124
105613	110110	115767	138327
105615	110132	120372	138436
105617	110151	120397	139816
105618	110825	120699	139819
105620	110831	120825	140903
105623	110833	120826	141003
105625	110849	120828	141981
105706	110850	121608	142015
106209	110863	121666	142018
106454	111454	121670	142625
106455	111455	121702	142912
106456	112082	121730	143987
106457	113351	122903	145483
106599	113357	122906	146819
106613	113379	124446	151553
106664	113409	124465	151566
106665	113986	124875	152255
106737	113990	125525	152385
107614	114024	125527	152434

152727	161936	173783	184180
156383	161976	174077	184199
156426	162010	177904	184257
156445	162711	178627	187442
156875	167904	178639	
159917	173776	181943	

Appendix E. Donors with missing values

The 122 donors set to normal because of missing values in the measurements for hypertension. The donors are included in the negative hypertension group constructed from less stringent criteria in kidney

DONOR_ID	108571	115753	138436
100077	108573	115756	139816
100078	108575	115758	139818
100109	108576	115759	139819
100195	108632	115761	140903
100253	109045	115763	142625
100279	109050	115766	142912
100797	109063	120825	143987
104864	109093	120826	145483
104865	109105	120828	151553
105172	110151	121608	152385
105323	110825	121666	152434
105578	110833	121670	152463
105579	110849	121702	152727
105613	110850	124293	156426
105615	111455	124446	156445
105617	112082	125525	156875
105620	113351	125527	159917
105621	113357	125622	161936
105625	113379	125863	161976
106454	113409	129754	162711
106455	113986	129884	167904
106457	113990	130819	173776
106599	114024	130821	173783
106664	114296	131379	174077
106737	114302	131490	178627
107857	114308	133765	181943
107859	114343	133777	184180
107943	115748	135675	187442
108009	115751	136488	
108563	115752	136489	

Appendix F. Donors with missing values

The 28 donors set to normal because of missing values in measurements in dyslipidemia or hypertension. The donors are included in the negative group constructed from stringent criteria.

DONOR_ID
100279
100797
105323
105613
105615
106599
108563
108575
108576
108632
110825
110850
111455
112082
115756
115759
115761
115763
115766
121666
121702
125525
130821
133777
135675
136488
156426
156445

Appendix G. Decision tree rules

Tables G1-G47 in this appendix shows the resulting rules transformed from the trees chosen for each risk factor.

```
Diabetes & NRs
Criteria: stringent
CF: 0,5 MNO:1
Class: Yes

!NR1D1_2,!RARA_4,NR3C2,!ESR1_2,!PPARG,VDR_1,!AREG,ESRRG
!NR1D1_2,!RARA_4,NR3C2,!ESR1_2,!PPARG,VDR_1,AREG,!HNF4A_4
!NR1D1_2,!RARA_4,NR3C2,!ESR1_2,!PPARG,!VDR_1
!NR1D1_2,!RARA_4,NR3C2,ESR1_2
!NR1D1_2,!RARA_4,!NR3C2,NR2F6
!NR1D1_2,RARA_4,!NR4A3
NR1D1_2,!NR2F1,VDR_2

Diabetes & NRs
Criteria: stringent
CF: 0,5 MNO:3
Class: Yes

!NR1D1_2,!RARA_4,NR3C2,!NR0B1_1,!NR2F1,NR4A3,!THRA_2,!VDR_2
!NR1D1_2,!RARA_4,NR3C2,!NR0B1_1,NR2F1
!NR1D1_2,!RARA_4,NR3C2,NR0B1_1
```

Figure G1. Rules generated from the diabetes group constructed from stringent criteria and NRs. The rules classify the samples to the Yes class.

```
Dyslipidemia & NRs
 Criteria: stringent
CF: 0,999 MNO:2
Class: Yes

!RORA_1,!NR2F1,!PPARG,NR1I3
!RORA_1,!NR2F1,!PPARG,!NR1I3,!NR3C1,!ESR1_2,!NR4A2_3,ESRRG
!RORA_1,!NR2F1,!PPARG,!NR1I3,!NR3C1,ESR1_2

Dyslipidemia & NRs
Criteria: stringent
CF: 0,999 MNO:5
Class: Yes

!RORA_1,!NR2F1,!PPARG,ESRRG,ESRRA_2
```

Figure G2. Rules generated from the dyslipidemia group constructed from stringent criteria and NRs. The rules classify the samples to the Yes class.

```
Hypertension & NRs
Criteria: stringent
CF: 0,999 MNO:2
Class: Yes

PPARA,PPARG
PPARA,!PPARG,AREG,VDR_2
PPARA,!PPARG,!AREG
!PPARA,!NR1I3,!NR4A2_1,!PPARG,RORA_2,RORA_1,VDR_2
!PPARA,!NR1I3,!NR4A2_1,!PPARG,!RORA_2,RORA_1
!PPARA,!NR1I3,!NR4A2_1,!PPARG,!RORA_2,!RORA_1,VDR_1,NR4A3
!PPARA,!NR1I3,!NR4A2_1,!PPARG,!RORA_2,!RORA_1,!VDR_1,!AREG
!PPARA,!NR1I3,NR4A2_1,NR4A3,!NR1H4
!PPARA,!NR1I3,NR4A2_1,!NR4A3

Hypertension & NRs
Criteria: stringent
CF: 0,999 MNO:5
Class: Yes

PPARA
!PPARA,!NR4A2_1,!PPARG,!RORA_2,NR3C2
!PPARA,NR4A2_1,!NR4A3
```

Figure G3. Rules generated from the hypertension group constructed from stringent criteria and NRs. The rules classify the samples to the Yes class.

```
Obesity & NRs
Criteria: stringent
CF: 0,75 MNO:3
Class: Yes

PPARA,AREG,NR1I3
PPARA,!AREG

Obesity & NRs
Criteria: stringent
CF: 0,999 MNO:5
Class: Yes

!PPARA
```

Figure G4. Rules generated from the obesity group constructed from stringent criteria and NRs. The rules classify the samples to the Yes class.

```
Diabetes & NRs
Criteria: stringent
CF: 0,5 MNO:1
Class: No

!NR1D1_2,!RARA_4,NR3C2,!ESR1_2,!PPARG,VDR_1,!AREG,!ESRRG
!NR1D1_2,!RARA_4,NR3C2,!ESR1_2,!PPARG,VDR_1,AREG,HNF4A_4
!NR1D1_2,!RARA_4,NR3C2,!ESR1_2,PPARG
!NR1D1_2,!RARA_4,!NR3C2,!NR2F6
!NR1D1_2,RARA_4,NR4A3
NR1D1_2,!NR2F1,!VDR_2
NR1D1_2,NR2F1


Diabetes & NRs
Criteria: stringent
CF: 0,5 MNO:3
Class: No

!NR1D1_2,!RARA_4,NR3C2,!NR0B1_1,!NR2F1,NR4A3,THRA_2
!NR1D1_2,!RARA_4,NR3C2,!NR0B1_1,!NR2F1,NR4A3,!THRA_2,VDR_2
!NR1D1_2,!RARA_4,NR3C2,!NR0B1_1,!NR2F1,!NR4A3
!NR1D1_2,!RARA_4,!NR3C2
!NR1D1_2,RARA_4
NR1D1_2
```

Figure G5. Rules generated from the diabetes group constructed from stringent criteria and NRs. The rules classify the samples to the No class.

```
Dyslipidemia & NRs
Criteria: stringent
CF: 0,999 MNO:2
Class: No

!RORA_1,!NR2F1,!PPARG,!NR1I3,!NR3C1,!ESR1_2,!NR4A2_3,!ESRRG
!RORA_1,!NR2F1,!PPARG,!NR1I3,!NR3C1,!ESR1_2,NR4A2_3
!RORA_1,!NR2F1,PPARG
!RORA_1,NR2F1
RORA_1

Dyslipidemia & NRs
Criteria: stringent
CF: 0,999 MNO:5
Class: No

!RORA_1,!NR2F1,!PPARG,ESRRG,!ESRRA_2
!RORA_1,!NR2F1,!PPARG,!ESRRG
!RORA_1,!NR2F1,PPARG
!RORA_1,NR2F1
RORA_1
```

Figure G6. Rules generated from the dyslipidemia group constructed from stringent criteria and NRs. The rules classify the samples to the No class.

```
Hypertension & NRs
Criteria: stringent
CF: 0,999 MNO:2
Class: No

PPARA,!PPARG,AREG,!VDR_2
!PPARA,NR1I3
!PPARA,!NR1I3,!NR4A2_1,PPARG
!PPARA,!NR1I3,!NR4A2_1,!PPARG,RORA_2,RORA_1,!VDR_2
!PPARA,!NR1I3,!NR4A2_1,!PPARG,RORA_2,!RORA_1
!PPARA,!NR1I3,!NR4A2_1,!PPARG,!RORA_2,!RORA_1,VDR_1,!NR4A3
!PPARA,!NR1I3,!NR4A2_1,!PPARG,!RORA_2,!RORA_1,!VDR_1,AREG
!PPARA,!NR1I3,NR4A2_1,NR4A3,NR1H4

Hypertension & NRs
Criteria: stringent
CF: 0,999 MNO:5
Class: No

!PPARA,!NR4A2_1,PPARG
!PPARA,!NR4A2_1,!PPARG,RORA_2
!PPARA,!NR4A2_1,!PPARG,!RORA_2,!NR3C2
!PPARA,NR4A2_1,NR4A3
```

Figure G7. Rules generated from the hypertension group constructed from stringent criteria and NRs. The rules classify the samples to the No class.

```
Obesity & NRs
Criteria: stringent
CF: 0,75 MNO:3
Class: No

PPARA,AREG,!NR1I3
!PPARA

Obesity & NRs
Criteria: stringent
CF: 0,999 MNO:5
Class: No

PPARA
```

Figure G8. Rules generated from the obesity group constructed from stringent criteria and NRs. The rules classify the samples to the No class.

```
Diabetes & NRs
Criteria: less stringent
CF: 0,5 MNO:5
Class: Yes

VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,RARB,!ESRRG,THRA_2
VDR_1,NR4A1,!NR1I3,!AR,HNF4A_1,!NR1D1_2
VDR_1,!NR4A1,NR2F2

Diabetes & NRs
Criteria: less stringent
CF: 0,999 MNO:5
Class: Yes

VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,RARB,!ESRRG,THRA_2
VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,RARB,ESRRG,!RARA_4,!PPARA,!RXRG
,AREG,ESR1_1,!VDR_2
VDR_1,NR4A1,!NR1I3,!AR,HNF4A_1,!NR1D1_2
VDR_1,!NR4A1,NR2F2
```

Figure G9. Rules generated from the diabetes group constructed from less stringent criteria and NRs. The rules classify the samples to the Yes class.

```
Dyslipidemia & NRs
Criteria: less stringent
CF: 0,75 MNO:2
Class: Yes

!NR2F1,!NR3C1,!RORA_1,!NR0B2,!ESR2_1,RXRG,NR4A2_3
!NR2F1,!NR3C1,!RORA_1,!NR0B2,ESR2_1,RORA_2,!VDR_2

Dyslipidemia & NRs
Criteria: less stringent
CF: 0,75 MNO:3
Class: Yes

!NR2F1,!NR3C1,!RORA_1,!NR0B2,!ESR2_1,!RXRG,!VDR_2,!RORA_2,NR2F6
,!NR1D1_2,VDR_1
!NR2F1,!NR3C1,!RORA_1,!NR0B2,!ESR2_1,RXRG
!NR2F1,!NR3C1,!RORA_1,!NR0B2,ESR2_1,!VDR_2
```

Figure G10. Rules generated from the dyslipidemia group constructed from less stringent criteria and NRs. The rules classify the samples to the Yes class.

```
Hypertension & NRs
Criteria: less stringent
CF: 0,25 MNO:10
Class: Yes

!RARA_4,!VDR_2,!NR4A2_1,!NR3C1,NR1H4,!NR4A1
!RARA_4,!VDR_2,NR4A2_1,!RORA_1
!RARA_4,VDR_2,THRA_2

Hypertension & NRs
Criteria: less stringent
CF: 0,5 MNO:20
Class: Yes

!RARA_4,VDR_2,THRA_2
```

Figure G11. Rules generated from the hypertension group constructed from less stringent criteria and NRs. The rules classify the samples to the Yes class.

```
Obesity & NRs
Criteria: less stringent
CF:0,25 MNO:1
Class:Yes

PPARA,VDR_1,NR1H3,NR1H4,!NR2E3_2,ESRRA_2,!HNF4A_1,RORA_1
PPARA,VDR_1,NR1H3,NR1H4,!NR2E3_2,ESRRA_2,HNF4A_1
PPARA,VDR_1,NR1H3,!NR1H4
PPARA,VDR_1,!NR1H3

Obesity & NRs
Criteria: less stringent
CF: 0,25 MNO:2
Class: Yes

PPARA,NR4A2_3,ESRRA_2,VDR_2,!AREG
PPARA,!NR4A2_3
```

Figure G12. Rules generated from the obesity group constructed from less stringent criteria and NRs. The rules classify the samples to the Yes class.

```
Diabetes & NRs
Criteria: less stringent
CF: 0,5 MNO:5
Class: No

VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,RARB,!ESRRG,!THRA_2
VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,RARB,ESRRG
VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,!RARB
VDR_1,NR4A1,!NR1I3,!AR,HNF4A_1,NR1D1_2
VDR_1,NR4A1,!NR1I3,AR
VDR_1,NR4A1,NR1I3
VDR_1,!NR4A1,!NR2F2
!VDR_1

Diabetes & NRs
Criteria: less stringent
CF: 0,999 MNO:5
Class: No

VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,RARB,!ESRRG,!THRA_2
VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,RARB,ESRRG,!RARA_4,!PPARA,!RXRG,
AREG,ESR1_1,VDR_2
VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,RARB,ESRRG,!RARA_4,!PPARA,!RXRG,
AREG,!ESR1_1
VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,RARB,ESRRG,!RARA_4,!PPARA,!RXRG,
!AREG
VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,RARB,ESRRG,!RARA_4,!PPARA,RXRG
VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,RARB,ESRRG,!RARA_4,PPARA
VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,RARB,ESRRG,RARA_4
VDR_1,NR4A1,!NR1I3,!AR,!HNF4A_1,!RARB
VDR_1,NR4A1,!NR1I3,!AR,HNF4A_1,NR1D1_2
VDR_1,NR4A1,!NR1I3,AR
VDR_1,NR4A1,NR1I3
VDR_1,!NR4A1,!NR2F2
!VDR_1
```

Figure G13. Rules generated from the diabetes group constructed from less stringent criteria and NRs. The rules classify the samples to the No class.

Dyslipidemia & NRs
Criteria: less stringent
CF: 0,75 MNO:2
Class: No

```
!NR2F1,!NR3C1,!RORA_1,!NR0B2,!ESR2_1,!RXRG
!NR2F1,!NR3C1,!RORA_1,!NR0B2,!ESR2_1,RXRG,!NR4A2_3
!NR2F1,!NR3C1,!RORA_1,!NR0B2,ESR2_1,RORA_2,VDR_2
!NR2F1,!NR3C1,!RORA_1,!NR0B2,ESR2_1,!RORA_2
!NR2F1,!NR3C1,!RORA_1,NR0B2
!NR2F1,!NR3C1,RORA_1
!NR2F1,NR3C1
NR2F1
```

Dyslipidemia & NRs
Criteria: less stringent
CF: 0,75 MNO:3
Class: No

```
!NR2F1,!NR3C1,!RORA_1,!NR0B2,!ESR2_1,!RXRG,!VDR_2,RORA_2
!NR2F1,!NR3C1,!RORA_1,!NR0B2,!ESR2_1,!RXRG,!VDR_2,!RORA_2,NR2F6,
!NR1D1_2,!VDR_1
!NR2F1,!NR3C1,!RORA_1,!NR0B2,!ESR2_1,!RXRG,!VDR_2,!RORA_2,NR2F6,
NR1D1_2
!NR2F1,!NR3C1,!RORA_1,!NR0B2,!ESR2_1,!RXRG,!VDR_2,!RORA_2,!NR2F6
!NR2F1,!NR3C1,!RORA_1,!NR0B2,!ESR2_1,!RXRG,VDR_2
!NR2F1,!NR3C1,!RORA_1,!NR0B2,ESR2_1,VDR_2
!NR2F1,!NR3C1,!RORA_1,NR0B2
!NR2F1,!NR3C1,RORA_1
!NR2F1,NR3C1
NR2F1
```

Figure G14. Rules generated from the dyslipidemia group constructed from less stringent criteria and NRs. The rules classify the samples to the No class.

Hypertension & NRs
Criteria: less stringent
CF: 0,25 MNO:10
Class: No

```
!RARA_4,!VDR_2,!NR4A2_1,!NR3C1,!NR1H4
!RARA_4,!VDR_2,!NR4A2_1,!NR3C1,NR1H4,NR4A1
!RARA_4,!VDR_2,!NR4A2_1,NR3C1
!RARA_4,!VDR_2,NR4A2_1,RORA_1
!RARA_4,VDR_2,!THRA_2
RARA_4
```

Hypertension & NRs
Criteria: less stringent
CF: 0,5 MNO:20
Class: No

```
!RARA_4,!VDR_2
!RARA_4,VDR_2,!THRA_2
RARA_4
```

Figure G15. Rules generated from the hypertension group constructed from less stringent criteria and NRs. The rules classify the samples to the No class.

Obesity & NRs
Criteria: less stringent
CF: 0,25 MNO:1
Class: No

```
PPARA,VDR_1,NR1H3,NR1H4,!NR2E3_2,ESRRA_2,!HNF4A_1,!RORA_1
PPARA,VDR_1,NR1H3,NR1H4,!NR2E3_2,!ESRRA_2
PPARA,VDR_1,NR1H3,NR1H4,NR2E3_2
PPARA,!VDR_1
!PPARA
```

Obesity & NRs
Criteria: less stringent
CF: 0,25 MNO:2
Class: No

```
PPARA,NR4A2_3,ESRRA_2,VDR_2,AREG
PPARA,NR4A2_3,ESRRA_2,!VDR_2
PPARA,NR4A2_3,!ESRRA_2
!PPARA
```

Figure G16. Rules generated from the obesity group constructed from less stringent criteria and NRs. The rules classify the samples to the No class.

```
Diabetes and a combination of NRs & co-factors
Criteria: stringent
CF: 0,5 MNO:5
Class: Yes

!HTATIP_2,CRY1,!MEF2D_1,!BLZF1_1
!HTATIP_2,CRY1,MEF2D_1,ELF3_1

Diabetes and a combination of NRs & co-factors
Criteria: stringent
CF: 0,999 MNO:2
Class: Yes

!HTATIP_2,CRY1,!MEF2D_1,BLZF1_1,!ELF3_1
!HTATIP_2,CRY1,!MEF2D_1,!BLZF1_1
!HTATIP_2,CRY1,MEF2D_1,ELF3_1,!HDAC7A
```

Figure G17. Rules generated from the diabetes group constructed from stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the Yes class.

```
Dyslipidemia and a combination of NRs & co-factors
Criteria: stringent
CF: 0,5 MNO:3
Class: Yes

!RFX5,NFYA_1
!RFX5,!NFYA_1,!EPAS1_2

Dyslipidemia and a combination of NRs & co-factors
Criteria: stringent
CF: 0,999 MNO:5
Class: Yes

!RFX5,TCFL4_1
```

Figure G18. Rules generated from the dyslipidemia group constructed from stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the Yes class.

```
Hypertension and a combination of NRs & co-factors
Criteria: stringent
CF: 0,25 MNO:10
Class: Yes

!CBFA2T1_1,!RFXAP
CBFA2T1_1

Hypertension and a combination of NRs & co-factors
Criteria: stringent
CF: 0,999 MNO:10
Class: Yes

!CBFA2T1_1,RFXAP,MED6_2
!CBFA2T1_1,!RFXAP
CBFA2T1_1
```

Figure G19. Rules generated from the hypertension group constructed from stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the Yes class.

```
Obesity and a combination of NRs & co-factors
Criteria: stringent
CF: 0,25 MNO:3
Class: Yes

PPARA,!ZFP95_1
!PPARA,!SMARCA4_5,!SP140,MSC

Obesity and a combination of NRs & co-factors
Criteria: stringent
CF: 0,25 MNO:5
Class: Yes

PPARA
```

Figure G19. Rules generated from the obesity group constructed from stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the Yes class.

```
Diabetes and a combination of NRs & co-factors
Criteria: stringent
CF: 0,5 MNO:5
Class: No

!HTATIP_2,CRY1,!MEF2D_1,BLZF1_1
!HTATIP_2,CRY1,MEF2D_1,!ELF3_1
!HTATIP_2,!CRY1
HTATIP_2

Diabetes and a combination of NRs & co-factors
Criteria: stringent
CF: 0,999 MNO:2
Class: No

!HTATIP_2,CRY1,!MEF2D_1,BLZF1_1,ELF3_1
!HTATIP_2,CRY1,MEF2D_1,!ELF3_1
!HTATIP_2,CRY1,MEF2D_1,ELF3_1,HDAC7A
!HTATIP_2,!CRY1
HTATIP_2
```

Figure G20. Rules generated from the diabetes group constructed from stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the No class.

```
Dyslipidemia and a combination of NRs & co-factors
Criteria: stringent
CF: 0,5 MNO:3
Class: No

!RFX5,!NFYA_1,EPAS1_2
RFX5

Dyslipidemia and a combination of NRs & co-factors
Criteria: stringent
CF: 0,999 MNO:5
Class: No

!RFX5,!TCFL4_1
RFX5
```

Figure G21. Rules generated from the dyslipidemia group constructed from stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the No class.

```
Hypertension and a combination of NRs & co-factors
Criteria: stringent
CF: 0,25 MNO:10
Class: No

!CBFA2T1_1,RFXAP

Hypertension and a combination of NRs & co-factors
Criteria: stringent
CF: 0,999 MNO:10
Class: No

!CBFA2T1_1,RFXAP,!MED6_2
```

Figure G22. Rules generated from the hypertension group constructed from stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the No class.

```
Obesity and a combination of NRs & co-factors
Criteria: stringent
CF: 0,25 MNO:3
Class: No

PPARA,ZFP95_1
!PPARA,SMARCA4_5
!PPARA,!SMARCA4_5,!SP140,!MSC
!PPARA,!SMARCA4_5,SP140

Obesity and a combination of NRs & co-factors
Criteria: stringent
CF: 0,25 MNO:5
Class: No

!PPARA
```

Figure G23. Rules generated from the obesity group constructed from stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the No class.

```
Diabetes and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,999 MNO:10
Class: Yes

!SUPT3H,PMF1,VDR_1,HTATIP2_1,RIPK3
!SUPT3H,PMF1,VDR_1,HTATIP2_1,!RIPK3,!BCL3_2
!SUPT3H,!PMF1,!HNF4A_1,!ARNT2,HMGCS2
!SUPT3H,!PMF1,HNF4A_1

Diabetes and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,5 MNO:10
Class: Yes

!SUPT3H,PMF1,VDR_1,HTATIP2_1,RIPK3
!SUPT3H,PMF1,VDR_1,HTATIP2_1,!RIPK3,!BCL3_2
!SUPT3H,!PMF1,!ARNT2,HMGCS2
```

Figure G24. Rules generated from the diabetes group constructed from less stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the Yes class.

```
Dyslipidemia and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,25 MNO:3
Class: Yes

!DDX17_3,!SMARCE1,!SAP30_2,!ELF4_1,TTF2
!DDX17_3,!SMARCE1,!SAP30_2,!ELF4_1,!TTF2,ZNF145
!DDX17_3,SMARCE1,!RIPK3,!PAX8_6,NR5A2_1,EGR2
!DDX17_3,SMARCE1,!RIPK3,PAX8_6,!TLE1_2,!RING1_2
!DDX17_3,SMARCE1,!RIPK3,PAX8_6,!TLE1_2,RING1_2,EGR2

Dyslipidemia and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,5 MNO:10
Class: Yes

!DDX17_3,!SMARCE1,SMARCA4_3
```

Figure G25. Rules generated from the dyslipidemia group constructed from less stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the Yes class.

```
Hypertension and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,25 MNO:5
Class: Yes

SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,ELF3_1,NR3C2,!TCF2_2,NFYC_4,
SUV39H1,ZFP95_2
SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,ELF3_1,NR3C2,!TCF2_2,NFYC_4,
!SUV39H1,MSC,TFDP2_1
SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,ELF3_1,NR3C2,!TCF2_2,NFYC_4,
!SUV39H1,!MSC
SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,ELF3_1,!NR3C2
SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,!ELF3_1,BTG2,!GLI2,TRIP11,ME
F2A,!PPARG
SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,!ELF3_1,BTG2,!GLI2,TRIP11,!M
EF2A
SMARCA2_2,ZNF226_1,!MED6_1,!CREB1_1
SMARCA2_2,ZNF226_1,MED6_1,GADD45B_3,NR2F1,NOTCH2_2
SMARCA2_2,ZNF226_1,MED6_1,GADD45B_3,!NR2F1,RYBP_3
SMARCA2_2,!ZNF226_1,!NT5C,NFYA_1

Hypertension and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,999 MNO:20
Class: Yes

SMARCA2_2,ZNF226_1,!MED6_1,ELF3_1,!TBL1X_2
SMARCA2_2,ZNF226_1,!MED6_1,ELF3_1,TBL1X_2,!E4F1
SMARCA2_2,ZNF226_1,!MED6_1,!TNRC11_2
```

Figure G26. Rules generated from the hypertension group constructed from less stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the Yes class.

```
Obesity and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,25 MNO:2
Class: Yes

PPARA,PAX8_4,CALR_1,SMARCD2,!MTF1_2
PPARA,PAX8_4,CALR_1,SMARCD2,MTF1_2,VDR_2,DEDD_1,!TTF2
PPARA,PAX8_4,CALR_1,SMARCD2,MTF1_2,VDR_2,!DEDD_1
PPARA,PAX8_4,CALR_1,!SMARCD2

Obesity and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,25 MNO:5
Class: Yes

PPARA,PAX8_4,!ZFP95_1,GPS2_2
```

Figure G27. Rules generated from the obesity group constructed from less stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the Yes class.

```
Diabetes and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,5 MNO:10
Class: No

!SUPT3H,PMF1,VDR_1,HTATIP2_1,!RIPK3,BCL3_2
!SUPT3H,PMF1,VDR_1,!HTATIP2_1
!SUPT3H,PMF1,!VDR_1
!SUPT3H,!PMF1,!ARNT2,!HMGCS2
!SUPT3H,!PMF1,ARNT2
SUPT3H

Diabetes and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,999 MNO:10
Class: No

!SUPT3H,PMF1,VDR_1,HTATIP2_1,!RIPK3,BCL3_2
!SUPT3H,PMF1,VDR_1,!HTATIP2_1
!SUPT3H,PMF1,!VDR_1
!SUPT3H,!PMF1,!HNF4A_1,!ARNT2,!HMGCS2
!SUPT3H,!PMF1,!HNF4A_1,ARNT2
SUPT3H
```

Figure G28. Rules generated from the diabetes group constructed from less stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the No class.

```
Dyslipidemia and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,25 MNO:3
Class: No

!DDX17_3,!SMARCE1,!SAP30_2,!ELF4_1,!TTF2,!ZNF145
!DDX17_3,!SMARCE1,!SAP30_2,ELF4_1
!DDX17_3,!SMARCE1,SAP30_2
!DDX17_3,SMARCE1,!RIPK3,!PAX8_6,!NR5A2_1
!DDX17_3,SMARCE1,!RIPK3,!PAX8_6,NR5A2_1,!EGR2
!DDX17_3,SMARCE1,!RIPK3,PAX8_6,!TLE1_2,RING1_2,!EGR2
!DDX17_3,SMARCE1,!RIPK3,PAX8_6,TLE1_2
!DDX17_3,SMARCE1,RIPK3
DDX17_3

Dyslipidemia and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,5 MNO:10
Class: No

!DDX17_3,!SMARCE1,!SMARCA4_3
!DDX17_3,SMARCE1
DDX17_3
```

Figure G29. Rules generated from the dyslipidemia group constructed from less stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the No class.

```
Hypertension and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,25 MNO:5
Class: No

SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,ELF3_1,NR3C2,!TCF2_2,NFYC_4,
SUV39H1,!ZFP95_2
SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,ELF3_1,NR3C2,!TCF2_2,NFYC_4,
!SUV39H1,MSC,!TFDP2_1
SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,ELF3_1,NR3C2,!TCF2_2,!NFYC_4
SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,ELF3_1,NR3C2,TCF2_2
SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,!ELF3_1,BTG2,GLI2
SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,!ELF3_1,BTG2,!GLI2,TRIP11,ME
F2A,PPARG
SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,!ELF3_1,BTG2,!GLI2,!TRIP11
SMARCA2_2,ZNF226_1,!MED6_1,CREB1_1,!ELF3_1,!BTG2
SMARCA2_2,ZNF226_1,MED6_1,GADD45B_3,NR2F1,!NOTCH2_2
SMARCA2_2,ZNF226_1,MED6_1,GADD45B_3,!NR2F1,!RYBP_3
SMARCA2_2,ZNF226_1,MED6_1,!GADD45B_3
SMARCA2_2,!ZNF226_1,NT5C
SMARCA2_2,!ZNF226_1,!NT5C,!NFYA_1
!SMARCA2_2

Hypertension and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,999 MNO:20
Class: No

SMARCA2_2,ZNF226_1,!MED6_1,ELF3_1,TBL1X_2,E4F1
SMARCA2_2,ZNF226_1,!MED6_1,TNRC11_2
SMARCA2_2,ZNF226_1,MED6_1
SMARCA2_2,!ZNF226_1
!SMARCA2_2
```

Figure G30. Rules generated from the hypertension group constructed from less stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the No class.

```
Obesity and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,25 MNO:2
Class: No

PPARA,PAX8_4,CALR_1,SMARCD2,MTF1_2,VDR_2,DEDD_1,TTF2
PPARA,PAX8_4,CALR_1,SMARCD2,MTF1_2,!VDR_2
PPARA,PAX8_4,!CALR_1
PPARA,!PAX8_4
!PPARA

Obesity and a combination of NRs & co-factors
Criteria: less stringent
CF: 0,25 MNO:5
Class: No

PPARA,PAX8_4,!ZFP95_1,!GPS2_2
PPARA,PAX8_4,ZFP95_1
PPARA,!PAX8_4
!PPARA
```

Figure G31. Rules generated from the obesity group constructed from less stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the No class.

```
Diabetes & co-factors
Criteria: stringent
CF: 0,5 MNO:5
Class: Yes

!HTATIP_2,CRY1,!MEF2D_1,!BLZF1_1
!HTATIP_2,CRY1,MEF2D_1,ELF3_1

Diabetes & co-factors
Criteria: stringent
CF: 0,999 MNO:2
Class: Yes

!HTATIP_2,CRY1,!MEF2D_1,BLZF1_1,!ELF3_1
!HTATIP_2,CRY1,!MEF2D_1,!BLZF1_1
!HTATIP_2,CRY1,MEF2D_1,ELF3_1,!HDAC7A
```

Figure G32. Rules generated from the diabetes group constructed from stringent criteria and co-factors. The rules classify the samples to the Yes class.

```
Dyslipidemia & co-factors
Criteria: stringent
CF: 0,5 MNO:3
Class: Yes

!RFX5,NFYA_1
!RFX5,!NFYA_1,!EPAS1_2

Dyslipidemia & co-factors
Criteria: stringent
CF: 0,999 MNO:5
Class: Yes

!RFX5,TCFL4_1
```

Figure G33. Rules generated from the dyslipidemia group constructed from stringent criteria and co-factors. The rules classify the samples to the Yes class.

```
Hypertension & co-factors
Criteria: stringent
CF: 0,25 MNO:10
Class: Yes

!CBFA2T1_1,!RFXAP
CBFA2T1_1

Hypertension & co-factors
Criteria: stringent
CF: 0,999 MNO:10
Class: Yes

!CBFA2T1_1,RFXAP,MED6_2
!CBFA2T1_1,!RFXAP
CBFA2T1_1
```

Figure G34. Rules generated from the hypertension group constructed from stringent criteria and co-factors. The rules classify the samples to the Yes class.

```
Obesity & co-factors
Criteria: stringent
CF: 0,999 MNO:20
Class: Yes

NCOA2_2

Obesity & co-factors
Criteria: stringent
CF: 0,999 MNO:3
Class: Yes

!HNF4A_1,NT5C,!NFKB2,LHX3
!HNF4A_1,NT5C,!NFKB2,!LHX3,DEDD_2,!BCL3_2
HNF4A_1,MEF2D_1
```

Figure G35. Rules generated from the obesity group constructed from stringent criteria and co-factors. The rules classify the samples to the Yes class.

```
Diabetes & co-factors
Criteria: stringent
CF: 0,5 MNO:5
Class: No

!HTATIP_2,CRY1,!MEF2D_1,BLZF1_1
!HTATIP_2,CRY1,MEF2D_1,!ELF3_1
!HTATIP_2,!CRY1
HTATIP_2

Diabetes & co-factors
Criteria: stringent
CF: 0,999 MNO:2
Class: No

!HTATIP_2,CRY1,!MEF2D_1,BLZF1_1,ELF3_1
!HTATIP_2,CRY1,MEF2D_1,!ELF3_1
!HTATIP_2,CRY1,MEF2D_1,ELF3_1,HDAC7A
!HTATIP_2,!CRY1
HTATIP_2
```

Figure G36. Rules generated from the diabetes group constructed from stringent criteria and co-factors. The rules classify the samples to the No class.

```
Dyslipidemia & co-factors
Criteria: stringent
CF: 0,5 MNO:3
Class: No

!RFX5,!NFYA_1,EPAS1_2
RFX5

Dyslipidemia & co-factors
Criteria: stringent
CF: 0,999 MNO:5
Class: No

!RFX5,!TCFL4_1
RFX5
```

Figure G37. Rules generated from the dyslipidemia group constructed from stringent criteria and co-factors. The rules classify the samples to the No class.

```
Hypertension & co-factors
Criteria: stringent
CF: 0,25 MNO:10
Class: No

!CBFA2T1_1,RFXAP

Hypertension & co-factors
Criteria: stringent
CF: 0,999 MNO:10
Class: No

!CBFA2T1_1,RFXAP,!MED6_2
```

Figure G38. Rules generated from the hypertension group constructed from stringent criteria and co-factors. The rules classify the samples to the No class.

```
Obesity & co-factors
Criteria: stringent
CF: 0,999 MNO:20
Class: No

!NCOA2_2

Obesity & co-factors
Criteria: stringent
CF: 0,999 MNO:3
Class: No

!HNF4A_1,NT5C,NFKB2
!HNF4A_1,NT5C,!NFKB2,!LHX3,DEDD_2,BCL3_2
!HNF4A_1,NT5C,!NFKB2,!LHX3,!DEDD_2
!HNF4A_1,!NT5C
HNF4A_1,!MEF2D_1
```

Figure G39. Rules generated from the obesity group constructed from stringent criteria and co-factors. The rules classify the samples to the No class.

```
Diabetes & co-factors
Criteria: less stringent
CF: 0,75 MNO:10
Class: Yes


!PMF1,!ARNT2,HMGCS2

Diabetes & co-factors
Criteria: less stringent
CF: 0,999 MNO:10
Class: Yes

!PMF1,!HNF4A_1,!ARNT2,HMGCS2
!PMF1,HNF4A_1
```

Figure G40. Rules generated from the diabetes group constructed from less stringent criteria and co-factors. The rules classify the samples to the Yes class.

```
Dyslipidemia & co-factors
Criteria: less stringent
CF: 0,25 MNO:3
Class: Yes

!DDX17_3,!SMARCE1,!SAP30_2,!ELF4_1,TTF2
!DDX17_3,!SMARCE1,!SAP30_2,!ELF4_1,!TTF2,ZNF145
!DDX17_3,SMARCE1,!RIPK3,PAX8_6,!TLE1_2,!RING1_2
!DDX17_3,SMARCE1,!RIPK3,PAX8_6,!TLE1_2,RING1_2,EGR2

Dyslipidemia & co-factors
Criteria: less stringent
CF: 0,5 MNO:10
Class: Yes

!DDX17_3,!SMARCE1,SMARCA4_3
```

Figure G41. Rules generated from the dyslipidemia group constructed from less stringent criteria and co-factors. The rules classify the samples to the Yes class.

```
Hypertension & co-factors
Criteria: less stringent
CF: 0,25 MNO:10
Class: Yes

SMARCA2_2,ZNF226_1,!MED6_1,ELF3_1,MBD1_3,TCFL4_1,NFATC3_2,!ESR2
_1
SMARCA2_2,ZNF226_1,!MED6_1,ELF3_1,!MBD1_3
SMARCA2_2,ZNF226_1,!MED6_1,!ELF3_1,!GLI2,TNRC11_2,SIAH2
SMARCA2_2,ZNF226_1,!MED6_1,!ELF3_1,!GLI2,!TNRC11_2,!FOXF2

Hypertension & co-factors
Criteria: less stringent
CF: 0,999 MNO:20
Class: Yes

SMARCA2_2,ZNF226_1,!MED6_1,ELF3_1,!TBL1X_2
SMARCA2_2,ZNF226_1,!MED6_1,ELF3_1,TBL1X_2,!E4F1
SMARCA2_2,ZNF226_1,!MED6_1,!ELF3_1,!TNRC11_2
```

Figure G42. Rules generated from the hypertension group constructed from less stringent criteria and co-factors. The rules classify the samples to the Yes class.

```
Obesity & co-factors
Criteria: less stringent
CF: 0,25 MNO:5
Class:Yes

HNF4A_1,IRF1,!ZFP95_1,!DKFZp761F0118_2,ELF3_1
HNF4A_1,IRF1,!ZFP95_1,DKFZp761F0118_2

Obesity & co-factors
Criteria: less stringent
CF: 0,5 MNO:10
Class:Yes

HNF4A_1,DDX17_3
```

Figure G43. Rules generated from the obesity group constructed from less stringent criteria and co-factors. The rules classify the samples to the Yes class.

Diabetes & co-factors
Criteria: less stringent
CF: 0,75 MNO:10
Class: No

PMF1
!PMF1,!ARNT2,!HMGCS2
!PMF1,ARNT2

Diabetes & co-factors
Criteria: less stringent
CF: 0,999 MNO:10
Class: No

PMF1
!PMF1,!HNF4A_1,!ARNT2,!HMGCS2
!PMF1,!HNF4A_1,ARNT2

Figure G44. Rules generated from the diabetes group constructed from less stringent criteria and co-factors. The rules classify the samples to the No class.

Dyslipidemia & co-factors
Criteria: less stringent
CF: 0,25 MNO:3
Class: No

!DDX17_3,!SMARCE1,!SAP30_2,!ELF4_1,!TTF2,!ZNF145
!DDX17_3,!SMARCE1,!SAP30_2,ELF4_1
!DDX17_3,!SMARCE1,SAP30_2
!DDX17_3,SMARCE1,!RIPK3,!PAX8_6
!DDX17_3,SMARCE1,!RIPK3,PAX8_6,!TLE1_2,RING1_2,!EGR2
!DDX17_3,SMARCE1,!RIPK3,PAX8_6,TLE1_2
!DDX17_3,SMARCE1,RIPK3
DDX17_3

Dyslipidemia & co-factors
Criteria: less stringent
CF: 0,5 MNO:10
Class: No

!DDX17_3,!SMARCE1,!SMARCA4_3
!DDX17_3,SMARCE1
DDX17_3

Figure G45. Rules generated from the dyslipidemia group constructed from less stringent criteria and co-factors. The rules classify the samples to the No class.

```
Hypertension & co-factors
Criteria: less stringent
CF: 0,25 MNO:10
Class: No

SMARCA2_2,ZNF226_1,!MED6_1,ELF3_1,MBD1_3,TCFL4_1,NFATC3_2,ESR2_
1
SMARCA2_2,ZNF226_1,!MED6_1,ELF3_1,MBD1_3,TCFL4_1,!NFATC3_2
SMARCA2_2,ZNF226_1,!MED6_1,ELF3_1,MBD1_3,!TCFL4_1
SMARCA2_2,ZNF226_1,!MED6_1,!ELF3_1,GLI2
SMARCA2_2,ZNF226_1,!MED6_1,!ELF3_1,!GLI2,TNRC11_2,!SIAH2
SMARCA2_2,ZNF226_1,!MED6_1,!ELF3_1,!GLI2,!TNRC11_2,FOXF2
SMARCA2_2,ZNF226_1,MED6_1
SMARCA2_2,!ZNF226_1
!SMARCA2_2

Hypertension & co-factors
Criteria: less stringent
CF: 0,999 MNO:20
Class: No

SMARCA2_2,ZNF226_1,!MED6_1,ELF3_1,TBL1X_2,E4F1
SMARCA2_2,ZNF226_1,!MED6_1,!ELF3_1,TNRC11_2
SMARCA2_2,ZNF226_1,MED6_1
SMARCA2_2,!ZNF226_1
!SMARCA2_2
```

Figure G46. Rules generated from the hypertension group constructed from less stringent criteria and co-factors. The rules classify the samples to the No class.

```
Obesity & co-factors
Criteria: less stringent
CF: 0,25 MNO:5
Class: No

!HNF4A_1
HNF4A_1,IRF1,!ZFP95_1,!DKFZp761F0118_2,!ELF3_1
HNF4A_1,IRF1,ZFP95_1
HNF4A_1,!IRF1

Obesity & co-factors
Criteria: less stringent
CF: 0,5 MNO:10
Class: No

!HNF4A_1
HNF4A_1,!DDX17_3
```

Figure G47. Rules generated from the obesity constructed from less stringent criteria and co-factors. The rules classify the samples to the No class.

Appendix H. Rules for all risk factors except for diabetes

Tables H1-H12 in this appendix shows the rules generated from the trees constructed from all risk factors except for diabetes.

```
All risk factors except for diabetes & NRs
Criteria: stringent
CF: 0,999 MNO:1
Class: Yes

!NR1D1_2,RORA_2,!VDR_2,!NR4A1,!ESRRG,!ESR1_1,NR2F6
!NR1D1_2,RORA_2,!VDR_2,!NR4A1,!ESRRG,!ESR1_1,!NR2F6,!NR3C1
!NR1D1_2,RORA_2,VDR_2,!ESR1_1
!NR1D1_2,RORA_2,VDR_2,ESR1_1,ESRRA_2,NR2F2

All risk factors except for diabetes & NRs
Criteria: stringent
CF: 0,999 MNO:3
Class: Yes

!NR1D1_2,RORA_2,!VDR_2,!NR4A1,NR1H4
!NR1D1_2,RORA_2,VDR_2,NR2F2
```

Figure H1. Rules generated from the all risk factors except for diabetes constructed from stringent criteria and NRs. The rules classify the samples to the Yes class.

```
All risk factors except for diabetes & NRs
Criteria: stringent
CF: 0,999 MNO:1
Class: No

!NR1D1_2,RORA_2,!VDR_2,!NR4A1,!ESRRG,!ESR1_1,!NR2F6,NR3C1
!NR1D1_2,RORA_2,!VDR_2,!NR4A1,!ESRRG,ESR1_1
!NR1D1_2,RORA_2,!VDR_2,NR4A1
!NR1D1_2,RORA_2,VDR_2,ESR1_1,ESRRA_2,!NR2F2
!NR1D1_2,RORA_2,VDR_2,ESR1_1,!ESRRA_2
!NR1D1_2,!RORA_2
NR1D1_2

All risk factors except for diabetes & NRs
Criteria: stringent
CF: 0,999 MNO:3
Class: No

!NR1D1_2,RORA_2,!VDR_2,!NR4A1,!NR1H4
!NR1D1_2,RORA_2,!VDR_2,NR4A1
!NR1D1_2,RORA_2,VDR_2,!NR2F2
!NR1D1_2,!RORA_2
NR1D1_2
```

Figure H2. Rules generated from the all risk factors except for diabetes constructed from stringent criteria and NRs. The rules classify the samples to the No class.

All risk factors except for diabetes and a combination of NRs & co-factors
Criteria: stringent
CF: 0,5 MNO:3
Class: Yes

!TBL1X_2,RORA_2,NFYC_4,!MEF2C

All risk factors except for diabetes and a combination of NRs & co-factors
Criteria: stringent
CF: 0,5 MNO:5
Class: Yes

!TBL1X_2,RORA_2,!SAP30_2

Figure H3. Rules generated from the all risk factors except for diabetes constructed from stringent criteria and a combination of NRs & co-factors. The rules classify the samples to the Yes class.

All risk factors except for diabetes and a combination of NRs & co-factors
Criteria: stringent
CF: 0,5 MNO:3
Class: No

!TBL1X_2,RORA_2,NFYC_4,MEF2C
!TBL1X_2,RORA_2,!NFYC_4
!TBL1X_2,!RORA_2
TBL1X_2

All risk factors except for diabetes and a combination of NRs & co-factors
Criteria: stringent
CF: 0,5 MNO:5
Class: No

!TBL1X_2,RORA_2,SAP30_2
!TBL1X_2,!RORA_2
TBL1X_2

Figure H4. Rules generated from the all risk factors except for diabetes constructed from stringent criteria and a combination of NRs & co-factors. The rules classify the samples to the No class.

```
All risk factors except for diabetes & co-factors
Criteria: stringent
CF: 0,999 MNO:1
Class: Yes

!TBL1X_2,ZNF226_1,GADD45B_2,TFDP2_1,E2F1_1
!TBL1X_2,ZNF226_1,GADD45B_2,!TFDP2_1,!TNRC11_3
!TBL1X_2,ZNF226_1,!GADD45B_2

All risk factors except for diabetes & co-factors
Criteria: stringent
CF: 0,999 MNO:3
Class: Yes

!TBL1X_2,ZNF226_1,!NR1D1_1,!PAX8_4,!ESR1_1
!TBL1X_2,ZNF226_1,!NR1D1_1,PAX8_4
```

Figure H5. Rules generated from the all risk factors except for diabetes constructed from stringent criteria & co-factors. The rules classify the samples to the Yes class.

```
All risk factors except for diabetes & co-factors
Criteria: stringent
CF: 0,999 MNO:1
Class: No

!TBL1X_2,ZNF226_1,GADD45B_2,TFDP2_1,!E2F1_1
!TBL1X_2,ZNF226_1,GADD45B_2,!TFDP2_1,TNRC11_3
!TBL1X_2,!ZNF226_1
TBL1X_2

All risk factors except for diabetes & co-factors
Criteria: stringent
CF: 0,999 MNO:3
Class: No

!TBL1X_2,ZNF226_1,!NR1D1_1,!PAX8_4,ESR1_1
!TBL1X_2,ZNF226_1,NR1D1_1
!TBL1X_2,!ZNF226_1
TBL1X_2
```

Figure H6. Rules generated from the all risk factors except for diabetes constructed from stringent criteria & co-factors. The rules classify the samples to the No class.

All risk factors except for diabetes & NRs
Criteria: less stringent
CF: 0,999 MNO:2
Class: Yes

!NR1D1_2,!NR0B2,NR5A2_1,ESRRA_2
!NR1D1_2,!NR0B2,PPARG,!NR4A1
!NR1D1_2,NR0B2,AREG

All risk factors except for diabetes & NRs
Criteria: less stringent
CF: 0,999 MNO:3
Class: Yes

!NR1D1_2,NR5A2_1,ESRRA_2
!NR1D1_2,!NR5A2_1,PPARG,!NR4A1

Figure H7. Rules generated from the all risk factors except for diabetes constructed from less stringent criteria & NRs. The rules classify the samples to the Yes class.

All risk factors except for diabetes & NRs
Criteria: less stringent
CF: 0,999 MNO:2
Class: No

!NR1D1_2,!NR0B2,NR5A2_1,!ESRRA_2
!NR1D1_2,!NR0B2,!PPARG
!NR1D1_2,!NR0B2,PPARG,NR4A1
!NR1D1_2,NR0B2,!AREG
NR1D1_2

All risk factors except for diabetes & NRs
Criteria: less stringent
CF: 0,999 MNO:3
Class: No

!NR1D1_2,NR5A2_1,!ESRRA_2
!NR1D1_2,!NR5A2_1,!PPARG
!NR1D1_2,!NR5A2_1,PPARG,NR4A1
NR1D1_2

Figure H8. Rules generated from the all risk factors except for diabetes constructed from less stringent criteria & NRs. The rules classify the samples to the No class.

```
All risk factors except for diabetes and a combination of NRs
and co-factors
Criteria: less stringent
CF: 0,5 MNO:2
Class: Yes

SAP30_2,!SAP18_2,NCOA2_2,NR2F2
!SAP30_2,RYBP_3,!NFATC3_1,ZNF226_1

All risk factors except for diabetes and a combination of NRs
and co-factors
Criteria: less stringent
CF: 0,5 MNO:3
Class: Yes

!SAP30_2,RYBP_3,!NFATC3_1,ZNF226_1
```

Figure H9. Rules generated from the all risk factors except for diabetes constructed from less stringent criteria and a combination of NRs and co-factors. The rules classify the samples to the Yes class.

```
All risk factors except for diabetes and a combination of NRs
and co-factors
Criteria: less stringent
CF: 0,5 MNO:2
Class: No

SAP30_2,!SAP18_2,NCOA2_2,!NR2F2
SAP30_2,!SAP18_2,!NCOA2_2
SAP30_2,SAP18_2
!SAP30_2,!RYBP_3
!SAP30_2,RYBP_3,NFATC3_1
!SAP30_2,RYBP_3,!NFATC3_1,!ZNF226_1

All risk factors except for diabetes and a combination of NRs
and co-factors
Criteria: less stringent
CF: 0,5 MNO:3
Class: No


SAP30_2
!SAP30_2,!RYBP_3
!SAP30_2,RYBP_3,NFATC3_1
!SAP30_2,RYBP_3,!NFATC3_1,!ZNF226_1
```

Figure H10. Rules generated from the all risk factors except for diabetes constructed from less stringent criteria and co-factors. The rules classify the samples to the No class.

```
All risk factors except for diabetes & co-factors
Criteria: less stringent
CF: 0,999 MNO:2
Class: Yes

SAP30_2,!SAP18_2,NCOA2_2,NR2F2
!SAP30_2,RYBP_3,!NFATC3_1,ZNF226_1,CALR_1

All risk factors except for diabetes & co-factors
Criteria: less stringent
CF: 0,5 MNO:3
Class: Yes

!SAP30_2,RYBP_3,!NFATC3_1,ZNF226_1
```

Figure H11. Rules generated from the all risk factors except for diabetes constructed from less stringent criteria and co-factors. The rules classify the samples to the Yes class.

```
All risk factors except for diabetes & co-factors
Criteria: less stringent
CF: 0,999 MNO:2
Class: No

SAP30_2,!SAP18_2,NCOA2_2,!NR2F2
SAP30_2,!SAP18_2,!NCOA2_2
SAP30_2,SAP18_2
!SAP30_2,!RYBP_3
!SAP30_2,RYBP_3,NFATC3_1
!SAP30_2,RYBP_3,!NFATC3_1,ZNF226_1,!CALR_1
!SAP30_2,RYBP_3,!NFATC3_1,!ZNF226_1

All risk factors except for diabetes & co-factors
Criteria: less stringent
CF: 0,5 MNO:3
Class: No


SAP30_2
!SAP30_2,!RYBP_3
!SAP30_2,RYBP_3,NFATC3_1
!SAP30_2,RYBP_3,!NFATC3_1,!ZNF226_1
```

Figure H12. Rules generated from the all risk factors except for diabetes constructed from less stringent criteria and co-factors. The rules classify the samples to the No class.